UNIVERSITY OF SOUTHAMPTON

A Study of Early Indication Citation Metrics

by

David C. Tarrant

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy

in the
Faculty of Physical and Applied Sciences
Electronics and Computer Science

October 2011

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF PHYSICAL AND APPLIED SCIENCES
ELECTRONICS AND COMPUTER SCIENCE

Doctor of Philosophy

A STUDY OF EARLY INDICATION CITATION METRICS

by David C. Tarrant

Research outputs are growing in number and frequency, assisted by a greater number of publication mediums and platforms via which material can be disseminated. At the same time, the requirement to find acceptable, timely, objective measurements of research "quality" has become more important. Historically, citations have been used as an independent indication of the significance of scholarly material. However, citations are very slow to accrue since they can only be made by subsequently published material. This enforces a delay of a number of years before the citation impact of a publication can be accurately judged. By contrast, each new citation establishes a large number of co-citation relationships between that publication and older material whose citation impact is already well established. By taking advantage of this co-citation property, this thesis investigates the possibility of developing a metric that can provide an earlier indicator of a publication's citation impact. This thesis proposes a new family of co-citation based impact measures, describes a system to evaluate their effectiveness against a large citation database, and justifies the results of this evaluation against an analysis of a diverse range of research metrics.

Contents

List of Figures

List of Tables

DECLARATION OF AUTHORSHIP

I, **David Carl Tarrant**

declare that the thesis entitled

A Study of Early Indication Citation Metrics

and the work presented in are my own. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- parts of this work have been published as:

Tarrant, D., Carr, L. and Payne, T. (2008) Releasing the Power of Digital Metadata: Examining Large Networks of Co-Related Publications. *Research and Advanced Technology for Digital Libraries. In Press*

Signed: .....

Date: .....

Acknowledgements

Thanks first must go to my primary supervisor, Dr. Leslie Carr who has inspired, encouraged and funded me through the many years this work has taken. Thanks also go to Dr. Terry Payne, my second supervisor for the first year, whose dedication to good research, precise methods and thorough checking have had a profound effect on who I am today. Thanks also goes to Professor Nigel Shadbolt and Dr Nicholas Gibbins for their invaluable time put forward to be my internal examiners. My doctorate was funded by the Engineering and Physical Sciences Research Council (EPSRC).

With the hard work involved in a doctorate comes the highs and lows of when things do and don't go to plan and when you delete all the test results accidentally. In addition, I would like to thank my colleagues, Tim Brody for providing a sounding board and critical feedback at times when I have needed it most, and Steve Hitchcock, with whom I have had the pleasure of publishing many papers and gaining valuable conference experience as a result.

I've had the opportunity to present my work at many international conferences and workshops, allowing me to meet follow researchers from a wide range of countries. My thanks go to the many conference committees who have seen my work as being valuable in their community.

For my parents - Brenda and Paddy.

Chapter 1

Introduction

Bibliometrics, the study of written works, provides many techniques to identify the significance of articles and publications. Current bibliometric techniques can take three or more years to identify a publication as high impact (Moed 2005). This is too long to be useful for early career researchers wishing to quantify their value (Harnad 2006*b*). Thus there is a need to examine new techniques for predicting impact earlier in the life of a publication.

One existing bibliometric technique is to count the number of citations that a publication receives. This 'Citation Count' is an example metric that allows a reader to determine a publication's contribution to a research discipline, a technique first applied widely by Garfield (Garfield 1955). However, delays inherent in the publication process may mean the first citations only appear a year after publication. Impact is established over the following years, as the number of citations builds. Although a single factor, such as Citation Count, should not be used as the sole measurement for impact of a researcher (Moed 2009), it is one of the key factors and for early career researchers, previous indicators (such as h-index) may not be present.

Reducing the delay between publication and accruing citations can be achieved in a number of ways. Many have suggested that Open Access (OA) not only facilitates earlier citation but also holds a citation advantage over non-OA articles (Lawrence 2001*a*, Eysenbach 2006, Antelman 2004). This is achieved by encouraging authors to make their work freely available in many forms including pre-prints. This results in the work being available sooner to a greater audience, enabling the immediate citation upon official publication. The internet has provided a key enabler to Open Access upon which services such as arXiv (Ginsparg 1994*b*) have been created, as well as push forward the case for online Open Access to publishers, via the release of the final version by the publisher or a pre-print by an author (Harnad et al. 2008).

More recently, a series of studies have shown very little or no citation advantage to be gained from OA publishing (Davis 2010, Moed 2007, Gaule & Maystre 2008). In the

majority of cases (both in support and against OA citation advantage) these studies introduced a bias at some point, limiting their study to certain areas of research and time scales. Davis (2010) feels that finding no citation advantage for OA articles "calls into question the wisdom of looking only at citation counts to measure the impact of a journal article, particularly given the ease of tracking article downloads online".

Download count is one example of a potential early indicator, which has been shown to correlate with subsequent impact (Brody 2006), however such measures are often limited in scope when resources are widely distributed. Additionally, such metrics, while found to be correlated, often discard the importance of peer review indicated by a citation between one author's work and another. The process of peer review involves an author carefully considering another's work for purposes of citing it themselves. A citation represents an author exhibiting their intellectual honesty in their own work; it says nothing about the quality of the cited works.

In order to study early indication metrics, while maintaining focus on peer reviewed literature, the dependence on citations has to be maintained. This thesis investigates a number of novel techniques for making early impact judgement based upon a particular perspective of a publications' citations: its co-citation network. Each new citation establishes a number of co-citation relationships between a single publication and older, more established material. Due to the large number of co-citations established by a single citation, the co-citation network soon becomes sizeable when compared to the citation network. By taking advantage of these properties, this thesis investigates the possibility of developing a metric which can provide an earlier indicator of a publication's citation impact.

Each new publication will cite a number of existing publications, increasing the citation count of each of these by one. By citing a number of publications, a link is also formed between these, which relates them as co-cited publications, forming a network of links bigger than the single citation network. Previous work has found that the co-citation network can be used effectively to judge the main research areas of the publication (Small 1973). This thesis looks at the potential of using the larger co-citation network to form an early indication metric of subsequent impact. This higher impact score may potentially lead to greater visibility of this article and thus a greater number of citations at an early stage. If this can be shown to be true, then there is the potential for an articles' peak citation rate to be reached sooner, thus decreasing the timespan of a publication life cycle and potentially speeding up scientific progress, although this is a bold claim.

In order to process the co-citation network, a number of new and existing metrics are applied to a large corpus of publication data. Each metric is compared to the baseline Citation Count algorithm in order to find if any new metric is able to indicate later

citation impact earlier in the publication life cycle. In order to test the effectiveness of each new metrics the following hypotheses are going to be tested:

- Co-Citation relations can be used to create an early indication metric for publication impact which correlates well with existing metrics.
- A metric based on co-citations will identify high impact publications sooner in their lifecycle.
- Applying co-citation metrics in search ranking will promote more recent publications.

The first hypothesis can be tested by examining the correlation between the rank order (of the same set of publications) produced by each new metric, against that of Citation Count. Any new early indication metric should show good positive correlation to Citation Count sooner in the publication life cycle. Correlations at the same point in time are also necessary, in order to ensure each metric is maintaining some relation when looking at established publications, but not essential when identifying the best early indication metric. Each alternative metric will be used to generate a ranked list of publications at various points in their life cycle. In order to judge impact, the rank position of known high impact publications (by Citation Count) will be analysed. An effective new algorithm should be able to place these publications higher in the overall ranking than Citation Count at the same point in time, thus testing the second hypothesis. Finally, as a side effect of the previous points, any new algorithm should be revealing a number of more recent publications in higher search positions earlier in the publication life cycle.

In order to prove these hypotheses a system is required, which can study the life cycle of a number of publications and examine at each stage the relationship between early impact metrics, and subsequent impact of the same publications by Citation Count. This thesis presents such a system; built to be generic and capable of applying a diverse number of metrics to a large set of publications and their related citation data. Using this system, a new family of CoRank metrics are applied and analysed to discover if any can satisfy the criteria required to fulfil the hypotheses and reveal a new early indicator for Citation Count.

The results of this thesis show that total count of citations towards articles with which a publication is co-cited, is capable of indicating subsequent impact over the first 12 months after publication of an article. Additionally, after this period the same articles are still ranked highly within the entire dataset, implying that this metric (called CoRank-LinkCount) still favours high impact established articles, while revealing a number of newer, subsequent high impact articles sooner in the publication life cycle.

This thesis analyses a number of new and existing metrics and uses Principal Component Analysis (PCA) to demonstrate the differences between families of metrics based

upon their significant properties. PCA is a technique that is designed to clearly identify dominant factors in metrics and other mathematical techniques, making it particularly appropriate for use here. PCA is used to examine significant correlation differences between algorithms and identify potential reasons and benefits of each group of metrics discovered when using this and other techniques. Co-Citation metrics are just one example of an alternative metric, which could be applied to the area of scholarly communications. The system designed as part of this thesis provides a mechanism to evaluate further alternative metrics against a variety of datasets in an environment where this is becoming ever more important.

1.1 Approach

In order to judge the impact of publications and articles earlier in their life cycle, it is first necessary to define what is meant by impact and address how it is measured. Currently, one of the most highly regarded and widely used[1,2] indicators of an articles' impact is deduced from that item's citation count; a simple metric which represents the total number of references obtained from other publications.

The strongest characteristic of a citation is the element of peer review and consideration which has gone into the process of citing another authors' work. Once a publication or author becomes cited, many will assume that Citation Count is an indicator of the quality of the work, much to the frustration of others (Garfield 1973).

Early indication metrics are defined as those which can, to some degree of accuracy, predict impact obtained later in the publication life cycle. With Citation Count being used as a key factor when deducing impact, early indication metrics endeavour to predict this through use of other data such as download count, a technique available to those publishing on the Web. While studies have shown that download count is correlated to some degree with subsequent citation impact (Watson 2009, Brody 2006), download count is subject to high levels of inaccuracy due to the distribution of resources of the Web. As well as providing opportunities to obtain download statistics, the Web has also become a platform, from which detailed information pertaining to publications can be retrieved and processed in new ways. These changes, among others, have led to new openings in the field of bibliometrics and webometrics to look at alternative metrics (altmetrics) to rate the impact of scholarly communications (Priem et al. 2010).

Commercial secondary publishers have typically taken on the role of re-keying, mining and analysing publication metadata, but as the primary literature has moved from print to the more open world of the Web, this task can now be undertaken by new services.

[1] Google Scholar - http://scholar.google.com - Displays citation count on every search result
[2] ISI Highly Cited - http://isihighlycited.com

Citeseer[3], working on the Computer Science literature found on websites, Citebase[4], working on publications in Open Access repositories, and Google Scholar[5], drawing from published journal collections as well as the open Web, all provide some kind of alternative to commercial citation analyses.

Working from early impact metrics, this thesis presents the train of thought which led to looking at other publication metadata and analysing its use as a possible technique to judge impact. Specifically of interest is the co-citation network. Current techniques for rating publication using citations only look at direct relations with other publications, while co-citation data is already in widespread use when placing a publication in a research area; it is the co-citations which indicate in which communities a publication is being cited (Small 1973).

A co-citation network exists between a publication and the publications alongside which it is cited. By deduction this means that a single direct citation can provide a good number of co-citations. Additionally, the set of co-cited publications is likely to consist of established publications which have already obtained some indication of impact. This is due to the very nature of the publication life cycle, reflecting the ways scholars will research and cite established and respected works. So the fact that the co-citation network will always be bigger than the citation network, and contain more established publications, makes it an ideal candidate for early impact metric examination.

The main body of this thesis focuses on a number of novel metrics designed to evaluate the potential of using the co-citation network as a source of impact information. In order to carry out this evaluation a system was designed to process co-relationship data, apply any number of metrics and compare the different sets of results.

Judging the effectiveness of any new metric requires comparison with currently available metrics, including Citation Count. It is this metric, and its performance on a network of publications that defines the set of test criteria, against which all other algorithms are evaluated.

With the changing nature of scholarly communications and the shift to online publishing and dissemination, additional metrics that are traditionally applied on the Web, including PageRank (used by Google), are also applied to the corpus of publications. When evaluated against the criteria set out by Citation Count, a number of interesting conclusions are discovered suggesting that a number of families of metrics exist, all designed for varying purposes.

Towards the end of this thesis, it is these families of metrics which are focused upon in more detail, looking at how different metrics can be used to measure variations in behaviour across a number of subject areas. The evolution of scholarly communication and

[3]http://citeseer.ist.psu.edu/
[4]http://www.citebase.org
[5]http://scholar.google.com

assessment techniques leads to new questions on how these behaviours can be accurately modelled and evaluated. The realisation of co-relation networks in this environment and the techniques used to model such metrics could provide important mechanisms needed for the future of scholarly impact studies.

1.2 Structure

The main body of this thesis focuses on analysing the use of co-citations for ranking of scholarly publications. Here is presented CoRank, a new metric designed to be an earlier indicator of subsequent impact measured by Citation Count. The history and respectability of Citation Count within the scholarly communications environment means that this accepted standard metric is used throughout to define the test criteria and evaluation methods. In addition to CoRank, a number of other metrics are investigated which have been designed to handle other factors present as a result of the now dynamic environment surrounding scholarly communications. Finally, an overall comparison of all of the metrics is drawn, showing how families of metrics can be classified and how these compare with one another.

Chapter 2 looks at the changing nature of scholarly communications. The amount of available journals and publications has been growing exponentially for a number of years due to a variety of social and economic factors. This growth means that the importance of quantitative measures to help organise this deluge of information is now even more apparent. At the same time, open access to research and online publishing has been suggested to speed up the rate of citation (Eysenbach 2006), meaning that any impact factor designed to handle the deluge of information, must also attempt to do it quicker.

The Open Access Initiative has played a key role in establishing easy to understand policies regarding open access publishing and commercial journals. Chapter 2 looks in more detail at the work of this initiative and the effect mandates have had on Open Access publishing. Finally, this chapter examines the overall impact of open access journals compared to those with a subscription based model. This study makes use of the data published online by Thomson Reuters in their Web of Science (WoS) index[6], who apply a number of metrics to the journal data in order to provide a series of citation reports each year.

Bibliometrics, the statistical study of written documents, although established for print media, has become an important area of study in other communities as well, especially on the Web. Chapter 3 introduces some of the basic laws surrounding bibliometrics, including those which rely on the mathematics behind the power law.

[6]Web of Science - http://thomsonreuters.com/products_services/science/science_products/a-z/web_of_science/

After looking at a number of different metrics and introducing the basic principles of the citation network, Chapter 3 applies a number of these metrics to the WoS journal data to compare the properties of each. This study reveals how the principal component of all of the WoS metrics is the journal citation score, including the metric named "Article Influence Score", which is not entirely what its name suggests.

When considering the impact of individual articles rather than issues of journals, the WoS metrics are found to be inappropriate for use. Chapter 3 looks a broader range of metrics to discover how areas which are not publication-based rate individual articles and items. Chapter 3 introduces two of the Web's metrics, Hubs & Authorities (Kleinberg 1999) and PageRank (Brin et al. 1998), discussing the merits of each algorithm and why each was designed to focus on the free form structure of the Web. Finally, in this chapter the idea of bibliographic coupling and co-citation networks is introduced, along with a brief explanation of how these networks can be used with existing citation metrics.

Continuing this theme, Chapter 4 looks at the various available early indication metrics applied both on the Web and to publication data. This chapter begins by looking at the division of metrics according to their type and data source and introduces a simple diagrammatic way of classifying the various metric types. Taking a live institutional scholarly repository containing over 15,000 publications, allows the application of many of the metrics which are encompassed by this classification. These include download and pathway metrics which track usage patterns in Web based systems.

Chapter 5 then starts to build towards the idea of using the co-citation network to rank publications. This chapter begins by summarising the various techniques outlined to this point and makes observations about what represents a "better" metric. It is these observations which form the basis of measurable criteria that can be used to examine if any new metric successfully fulfils the hypotheses. With these conditions and evaluation criteria established, Chapter 5 introduces the first iteration of the CoRank algorithm. A small artificial network of publications is used to demonstrate practically how a network of co-citations builds from the citation network, showing how a network of 32 citations (from 18 publications) produces a total of 82 co-citations.

While applying each of the chosen metrics to the artificial network, including CoRank, the characteristics of each algorithm are looked at in more detail outlining how the input data affects the performance of each. Most significantly, this involves examining the iterative metrics to show how the number of citations affects the number of iterations which need to be performed. Finally in Chapter 5, the results of applying the different metrics to the test network are analysed in order to demonstrate how CoRank performs within acceptable boundaries on the artificial network.

In Chapter 6, CoRank is applied to a real citation network sourced from Citebase. The Citebase dataset provides a set of over 300,000 publications with more than 3.37 million citations indexed relating to these publications. In the same manner as in the artificial

network, analysis of Citebase ensures that each algorithm being applied produces an accurate result prior to live application.

Following the complete life cycle of a publication requires the tracking of a series of publications over a number of years, from the point when they are added into Citebase. In order to obtain such data, instead of waiting a number of years for it to accumulate, a number of snapshots were created representing a number of historic views of Citebase. Having many algorithms to apply to these datasets, each containing several million citation links, implies that the results of each algorithm could take some time to compute, especially when the algorithm has to perform a number of iterations. Chapter 6 introduces the Co-Ordinator system for handling this complete operation, showing how it was designed to scale for processing results quickly and in a distributed fashion.

Finally in Chapter 6, the initial set of algorithms is applied to the Citebase data and results presented. Each set of results is presented individually before being collated and verified against a wider sample. A thorough evaluation of further factors is then undertaken in order to evaluate the difference between the artificial network results and those obtained from the Citebase data, before concluding that performance of the CoRank algorithm is not as positive as hypothesised.

Chapter 7 builds on the CoRank algorithm and looks at other variations and possibilities for using the co-citation network for ranking. The target of each algorithm remains the same and the idea behind each variation is explained before the results are presented. A total of six further algorithms are presented in this chapter, these results are then collated with the data from Chapter 6 and analysed together. With all the data collected and basic conclusions drawn, statistical significance calculations ensure that results are sound and reliable. Although it may seem odd to perform this process after some conclusions have been drawn, there are several arguments for and against statistical significance testing which influenced this decision; these are also explained here.

Chapter 7 contains the main result of this thesis, which is split into a three part summary covering the performance of all ten metrics against the hypotheses. Section 7.2 looks at the correlations between each metric in an attempt to identify a potential early indicator of Citation Count rank order other than itself. Section 7.3 looks at the mean rank positions of known high impact publications, while Section 7.4 examines the ages of publications in the top 5% of all publications as ranked by each metric. While many of the metrics show a good positive correlation in Section 7.2, the benefit of CoRank-LinkCount (the best of the ten metrics) is much clearer to see in Section 7.3. Here it is able to reveal subsequent high impact publications within the first 12 months after publication. Combined with positive performance in all other tests, CoRank-LinkCount is the best of the new algorithms introduced in this thesis.

With all the metrics introduced, applied, and basic analysis performed, Chapter 7 then takes a broader look at how the algorithms relate to each other. The characteristics of

each algorithm map to different usage concepts and behaviours, and this can be seen in the results and the subsequent "family" groups which the algorithms form. Taking a closer look at the characteristics of each of these groups, allows the evaluation and discovery of a whole landscape of different metrics.

Chapter 8 addresses how bibliometrics studies human behaviour and how it changes over time. By looking at the application of bibliometrics to rate institutions and academics, it is possible to show how the "publish or perish" paradigm has been applied, and how it can fail to accurately track the methodologies applied by different subject areas.

In a related study, Bollen et al. (2008) looks more generally at the behaviour of 47 bibliometric algorithms on usage data, and maps them according to characteristics into a visualised space showing the clustering and separation between groups of algorithms. With similar groupings already observed in Chapter 7, a similar study can be undertaken which looks at how the algorithms presented in this thesis relate to each other, as well as those analysed by Bollen. Performing this mapping is not only of interest to the work in this thesis, but a good way of proving the methodology used and results found using a completely different dataset. In order to map the algorithms the principal eigenvectors (or components) of each algorithm have to be obtained, these are then transformed by Principal Component Analysis (a technique demonstrated on a much simpler dataset in Appendix 8) into values which can be mapped onto a 2D visualisation. Chapter 8 concludes by presenting the results of this analysis over the metrics applied and compares these to the findings of Bollen.

Chapter 9 then provides a summary of findings from this thesis, including stumbling points and future directions. The work on CoRank and related metrics reinforces the opportunities for different application depending on the behaviour to be modelled. The original CoRank metric is a quite complex algorithm based upon PageRank, both of which perform badly against the test criteria. The characteristics of these metrics compared to Citation Count, the target metric, dictate that similar performance is never possible. By relaxing some of the set out criteria, the greater benefits of analysing co-relations may have been revealed more clearly. The Co-Ordinator system provides a mechanism by which many novel techniques and refinements can be easily investigated and tested against a variety of criteria. It is this system, and the realisation of the value of co-citations, which can support the development of an evidence base to satisfy the requirements of the academic and further research community.

Chapter 2

The Movement Towards Online Scholarly Communications

Scholarly communications are the means through which researchers, academics and scholars share their work with the wider academic community. They play a key role in the dissemination of knowledge gathered through continuous research and experimentation.

Scholarly communications has a long history dating back to the 17th century and the inception of the Royal Society in London who are credited as the publishers of the worlds first scientific peer reviewed journal in 1665. The society was established as a place of research and discussion. The publishing of the first journal enabled such practices to continue outside of the society itself (Syfret 1948). The Royal Society continues to play an important role in the scholarly community and has many significant figures associated with it, who have been recognised for their contributions to science.

Contributions begin with research; a practice where an academic attempts to build upon existing work by gathering together results which constitute their own original research. The academic may then submit this work to be published, at which point it will undergo review by peers who are also experts in this field. If found to be of high enough quality and significance, this article may be accepted and published in a scholarly communications journal. At this point another researcher may pick up on this work and the scholarly communications cycle, depicted by Figure 2.1, starts again.

At the start of this cycle, background study requires access to the current publications. There is the potential that a researcher could miss a publication on the exact principal which they are intending to investigate, thus potentially wasting their time with repeated results. To ensure that this is not the case, a researcher should be able to obtain access to all journals relevant to their area of research; a problem in itself if the institution or individual does not subscribe to all of the available journals.

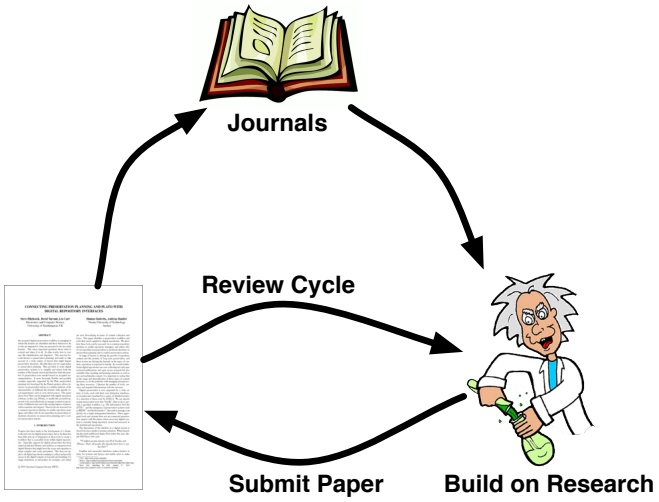

FIGURE 2.1: The basic scholarly communications life cycle (adapted from the EPrints Handbook[1])

Over the last decade the number of academic journals has been growing exponentially (Figure 2.2); likewise, so has the subscription cost related to the top rated journals. This has led to problems in institutions maintaining subscriptions to all of the relevant journals their academic population requires access to. This problem became known as the "Serials Crisis" (see Section 2.1), referring to the number of serials (another word for journals) and the decreasing budgets with which they can be purchased (Panitch & Michalak 2005).

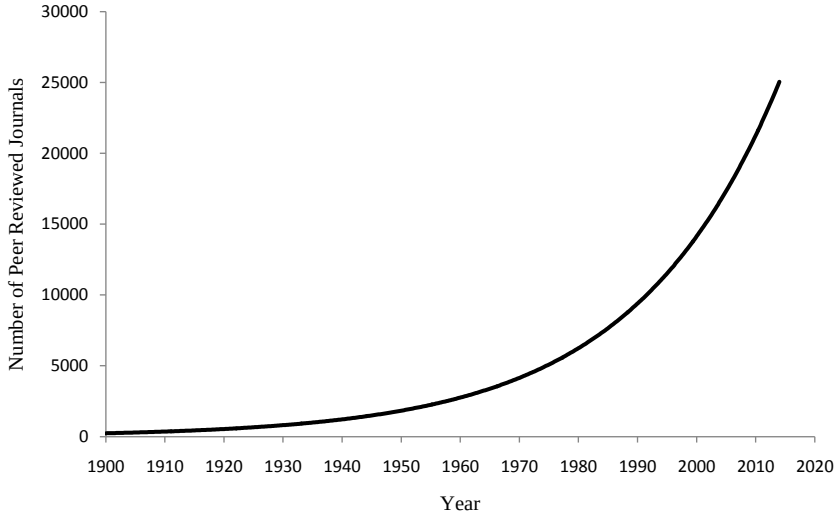

FIGURE 2.2: Academic Journal Growth Rate 1900 - 2004 (source: http://www.ulrichsweb.com/)

Similarly, when publishing their research a scholar may wish to publish in a journal which is not only going to be accessible to their peers, but also going to reflect well on themselves. Due to the competitive nature of modern institutions, publishing of high impact articles is now used as a measure of success of both the institution and

the individual (Day 2004). At the individual level this pressure is known as "publish or perish". The publish or perish imperative refers to the culture that a good academic or researcher, is one who has achieved many high impact journal publications.

The exponential growth in number of available journals has been driven by a number of factors including, a distinct increase in the number of scientific specialisations, access to technologies such as the internet, and partly by the "publish or perish" imperative (Clapham 2005, Harnad 2006*b*). Fortunately this increase in the number of available journals has been shown not to decrease the standard of scholarly publishing (Goel & Faria 2007).

According to Harnad (1995), the Web provides a potential solution to the serials crisis, reducing both the cost in publishing and more significantly, the cost of dissemination. The Web opens up a whole new environment for both printed and, more recently, digital only journals to publicise themselves at a much reduced cost. A publisher only needs to host an item online once (at the cost of a small amount of storage), in order for everyone connected to the Web to gain access to this resource. When comparing this to the dissemination requirements for printed material, the reduction in costs is significant. The size of the audience on the Web is also much greater; particularly as the Web is now regarded as such a critical source of information and means of communication that people have a "fundamental right" to have access to it (Hick et al. 2000 and BBC News[2]). Lawrence (2001*b*) also realised the Web as a critical method of dissemination for authors. His study looked at the different citation rates between articles available online and those which are not and finds a significant benefit to being online. Lawrence concludes with the question "If you are not online, are you are invisible?".

There is no doubt that the Web has had a profound effect on modern day society (Castells 2010), especially in the virtual lowering of distances between people. It is now very easy to stay in communication with fellow peers in a completely different time zone via email and other social technologies. Long distance collaborations are now fuelled by instant conversation technologies and the ability to be able to share large datasets and findings over the Internet.

While there are a great number of new journals being born, both printed and digitally, the Open Access (OA) movement (see Section 2.2) is spurring the online publication of the same peer reviewed items in full or pre-print version. OA publishing on the Web provides an opportunity to mitigate the power of traditional publishers, relax locational constraints and change the scale of operations (Goel 2003, Bergstrom & Bergstrom 2004), each helping to further reduce the cost of publishing. The other main benefit comes from the speed at which material can be made available online compared to via a journal or other carefully collated publication. This enables easier discovery and access to the articles themselves, a solution which has been shown to have a direct positive

[2]Internet access is 'a fundamental right' - http://news.bbc.co.uk/1/hi/technology/8548190.stm

effect on achieved impact (Harnad 2006*a*). In the same study, Harnad defends the simple principal that the more people who have access to the research, the more your potential impact. Similarly, the quicker this access is given the more relevant the work is likely to be to current topics of research. Harnad also separates the serials crisis from the access/impact problem, stating that making everything available will not necessarily make it discoverable and thus citable (Harnad et al. 2008).

Figure 2.3 shows a more modern take on the scholarly communications life cycle, adding global publishing via the Web as a possible dissemination mechanism. This cycle shows the two possible routes to OA publishing discussed later in section 2.2; pre-print publishing by the author (with permission) and post-print publishing by either the author or the journal. While the post-print represents the fully reviewed and edited publication, a pre-print could be an earlier version which is just as valid in results and content. Being available on the Web then leads to new impact cycles being created, with access being easier than to the subsequently printed subscription journals. Figure 2.3 has been adapted from a version in the EPrints Handbook[3], which also states that due to online publishing "research impact is greater (and faster) because access is maximized (and accelerated)".

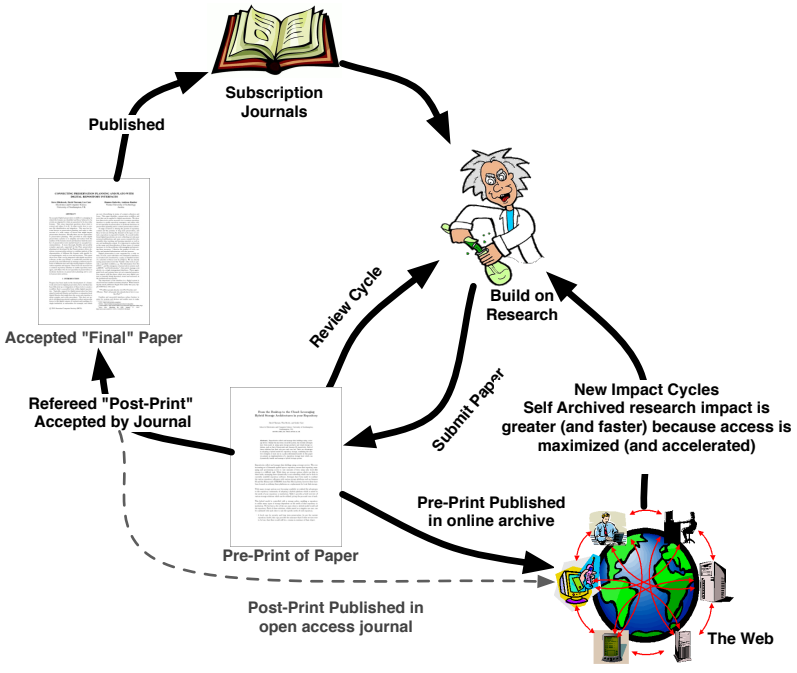

FIGURE 2.3: The scholarly communications life cycle, improved by online publishing (adapted from the EPrints Handbook[4])

With 25,000 journals (in 2004) responsible for around 2.5 million articles a year (plus those articles now getting published via OA methods), even when split by subject or research area, there is still an enormous number of articles to consider when doing background research. Odlyzko (2002) gives a good analogy of finding something relevant

[3]EPrints Handbook - http://www.eprints.org/documentation/handbook/golden.php (Jan 2011)

in amongst these 2.5 million articles being similar to swimming across a raging "river of knowledge".

Even if a researcher has access to all the relevant journals and publications, through a variety of means, there is still a problem in knowing which of these are the most authoritative and trusted sources of information. What is needed is some mechanism to be able to rank the located articles in order of perceived "importance". This weighting became known as the "impact" with a series of metrics devised from which the impact of a journal or article can be judged.

The initial discovery of impact metrics can be attributed to Garfield (1955) who, while looking for information in a subject area about which he knew very little, realised that he had no idea if the information he was finding was authoritative. Garfield's realisation was that a publication can be viewed as authoritative if there exists a number of citations towards it; the higher this number the more "impact" this publication can be said to have. Since Garfield's initial discovery of impact metrics (discussed further in Chapter 3), these have become a method through which journals can be assessed and judged. With the increased number of available journals, there is a lot of competition to be high impact in order to appeal to a larger number of subscribers.

A citation, often referred to as a reference, exhibits an authors intellectual honesty. It provides readers an indication of the authors background knowledge and source for information. During the background research stage, an academic will gather together this evidence base of information in order to help guide their research. These background publications may then be cited in subsequent works in order to communicate the depth of background research covered to the reader. Citing publications in this way can also assist an academic to prove statements without having to carry out the work themselves as exemplified in the following sentence: Men and the elderly could be said to be better drivers than women and young people as they have less accidents, however these are more likely to be fatal (Massie et al. 1995). At the time of writing (Jan 2011) this article was known to be cited by 155 other articles implying that it is a high impact and potentially trustworthy source. More about the history of impact metrics and bibliometrics is introduced in Chapter 3.

During the study carried out by Lawrence (2001*b*), it was found that online dissemination of publications leads to a 336% increase in citations. Thus the access/impact problem is reduced to a simple concept; the more people that can access an article, the more can cite it. Since this initial study this figure has been found to be generous with variations observed across different disciplines (Eysenbach 2006, Antelman 2004, Brody & Harnad 2004). Conversely, a series of other studies have shown very little or no citation advantage to be gained from OA publishing (Davis 2010, Moed 2007, Gaule & Maystre 2008). In the majority of cases (both in support and against OA citation advantage) these studies have introduced a bias at some point, limiting their study to

certain areas of research and time scales, implying that the question of whether OA publishing does give a citation advantage is still open.

As people turn to online methods of discovery (such as Google and Google Scholar), it is generally accepted that online dissemination, not necessarily linked to Open Access, is essential for discovery. As the number of publication records online becomes larger, the more critical metrics are to enable the locating of not just relevant articles, but also those which are authoritative.

Online dissemination speeds up the whole dissemination process, but with the number of materials available, the need for metrics to judge impact does not disappear. Garfield was an early pioneer in the application of quantitative impact metrics to scholarly publications and since his initial work (Garfield 1955) the statistical study of publications has become its own research area titled bibliometrics (see Chapter 3). It is the theory and the study of impact metrics which forms the basis of this thesis. The move towards online dissemination and publication brings a number of questions. Are existing metrics suitable for discovery of authoritative sources of information online? Are there any better metrics available which can make use of the plethora of information and raw data which is now available in the online environment?

This remainder of this chapter discusses potential transition towards online scholarly communications, and how by moving the whole process from research to publishing into an online environment not only helps with the serials crisis but also, and more importantly, the access/impact problem (Harnad et al. 2008). The serials crisis is looked at briefly before moving focus to open access and online dissemination and publishing, addressing how journal publication and open access can be shown to compliment each other.

2.1 The Serials Crisis

The serials crisis is the term given to the problem of libraries and other journal subscribers not having the funds to keep up with the number of Journals now being published (Panitch & Michalak 2005). Additionally, increasing journal prices and dwindling budgets mean that institutions are having problems maintaining their current subscriptions. Studies have found that the average yearly price increases on Journal subscriptions are in many areas, much higher than the world Consumer Price Index. In the UK between 2000 and 2004 price increases ranged between 27% and 94% (White & Creaser 2004), with a similar trend being reported in the US by Panitch & Michalak (2005).

Interestingly, the journals to see the highest price increases have been those with the greatest impact factor. This can have the the negative affect on libraries on limiting the

number of lower impact and start-up journals which can be purchased (Panitch & Michalak 2005, Odlyzko 2002). Thus this has a negative impact on newer researchers, who are struggling with the "Publish or Perish" paradigm and are unlikely to be publishing in high impact journals early on in their career.

Harnad (1995) analyses the scholarly communications market and asks why there exists a trade and subscription based model in an "esoteric" community? In a traditional publication market there are a small group of producers who are producing content for a large number of consumers. Scholarly communications is the opposite and tends to be very specialist; only written for a small number of fellow experts. The scholarly community is therefore an "esoteric" one where a low level of demand drives up the cost of items, such as printed conference proceedings and academic Journals (Harnad 1995).

Another parallel can be drawn with the common trade model of books in looking at the holders of the rights of the material. In the current environment, in order to get published authors have to sign over copyright on their works without return (Harnad 1995). Again in this situation, because the market is trade based, it is the publishers who are trying to glean back the money spent during the publication process. However on the occasion when they do make profit, none of this is fed back to the authors, unlike in the more widely known book publication market, where the demand is what influences the writing and publication of the book in the first place.

With the prices increasing (Dingley 2006), the serials crisis was viewed mainly as a library problem (Panitch & Michalak 2005), however a lot of suggestions have also been seen to move the cost away from the subscriber and onto the author (Panitch & Michalak 2005, Harnad 1995). This benefits many parties who have an agenda to see their works published including authors and funding councils. If the cost was incurred at this point it may open up a much bigger market of consumers.

Electronic publishing is proposed as another parallel solution. Based on the fact that most libraries are now searched from an electronic console, is it such a big leap to move to having a further button to allow you to download the work and print it locally if required (Panitch & Michalak 2005, Harnad 1995).

The answer to the question of how much money electronic only publication would save is still very much up in the air. While many publishers (Elsevier, Springer are just two examples) are now offering an electronic based service, this is offered as an additional service subject to a surcharge. Thus if a library owns a paper subscription, they can add the electronic one for an additional 10%, whereas the paper subscription can be added on top of the electronic one for an additional 25%. Thus the saving comes in between 10% and 15% (Panitch & Michalak 2005). According to Harnad (1995), in his own experience he could save between 70% and 90% by switching to electronic only publishing techniques. The key here was to not offer both types of publication as the decreased demand for hard copies will mean an increase in printing fees for shorter runs.

In addition, it may be possible to save further costs by cutting down on the amount of copy editing which is undertaken in order to push out well formatted documents when the authors' version may be suitable.

2.2 The Open Access Movement

Along with the serials crisis the "Publish or Perish" philosophy affects the area of initial publication. Researchers are interested ways in which their work becomes known, read and then cited in order to gain impact. With the serials crisis having a major effect on the number of journals researchers can now gain access to, the access/impact problem (Harnad et al. 2008) becomes a real issue. An author may not be capable of gaining publication in the few high end journals to which the majority of their audience are subscribed, whereas the journals that they are published in are not seeing high distribution rates. In parallel, it is estimated that there are now over 2.5 million articles published on a yearly basis (Figure 2.2), thus the problem is not just getting the articles online, but also making them discoverable and preserved for long term access.

Harnad et al. (2008) argues that simply by solving the serials crisis, by lowering the costs of all journals, won't solve the researchers "Publish or Perish" problem. With over 2.5 million articles published yearly, Harnad observes that being able to discover a researchers work in the first place, such that it can then be cited and gain impact, is a separate "Access/Impact problem".

With the proliferation of the internet and technology moving at a pace where processing speed and disk space are doubling approximately every 2 years (Moore 1998). Odlyzko (1995) predictions on using computing technology to not only store but also provide journals and articles to researches, is coming to fruition.

Due to the nature of the scholarly communications market place being esoteric (Harnad 1995), there is also a great deal of support behind the Open Access (OA) movement. OA is seen as one of the key ways to solve the article access/impact problem by allowing full-texts to be published online available for free download. This technique of publishing also directly benefits the author as it has also been shown that more citations are occurred towards open access publications (Brody 2006).

Originally called Free Online Scholarship (FOS) the Open Access Movement has a long history[5] in encouraging different methods of open access to supported by the journal publishers, institutions and individual authors themselves. By simplifying support for Open Access to easily classified Gold and Green routes, publishers are able to either make all their content open access (often and unfortunately applied with some delay factor i.e. 6 months) or allow authors to self archive their content openly elsewhere.

[5]Timeline of Open Access Movement - http://www.earlham.edu/ peters/fos/timeline.htm

Gold-OA refers to the former, fully open publication paradigm, while Green-OA is still highly useful even though more effort is required by the author (Harnad et al. 2008).

For OA publishing to succeed there has to be a great push for it, especially in the case of Green-OA publishing, both from the users as well as from institutions and funding bodies who initially fund the research. In the past few years a number of major bodies[6,7] have put in place policies, which encourage the OA publication of their funded research. These mandates are intended to make authors think more carefully about their publications including issues surrounding blindly handing over copyright to publishers. Such bodies are publicly funded and are tasked with distributing grants to scholarly projects and schemes. Since the money they are distributing is public, logically the outputs the public are paying for should, in turn, be freely accessible. As a result of many government recommendations in countries including the US and UK around 90% (based upon a study of just over 10,000 journals) are already Green-OA (Harnad et al. 2008) while 14% of all journals achieve the full Gold rating[8]. Similar results have also been seen in a study by Brody & Harnad (2004).

Along with funders and government bodies, a number of institutions[9,10] have also started mandating the OA publishing of content which is produced by its members. In the majority of cases this mandate is provided alongside an institutional repository, the purpose of which is to collect and preserve the content being submitted to it. Different clauses in the mandate will also dictate the amount of time an item is under embargo before it becomes open access. The number of institutions with a publication repository has also been growing steadily, however the number of full texts directly available through these repositories is still significantly less than the number of articles listed in these repositories (source: http://www.roar.eprints.org). This discrepancy is due to such repositories also being regarded as useful aggregators for metadata pertaining to articles which cannot be made freely available. If simply trying to locate the article is the aim, then this practice does go some way to helping solve the access/impact problem, however without the full text the article cannot be said to be open access. Commercial journals also publish a list of publications online as this may lead to individual article sales or better and thus increased revenue for the publisher.

Even with support for OA publishing on the increase, there is still a challenge in getting authors to support OA, specifically through the Green-OA route where the journal allows them to self-archive. The problem here is getting the author to self archive. Recent research by Swan & Brown (2004*a*) (with additional evidence presented in Swan & Brown

[6]Joint Information Systems Committee (JISC) Open Access policy - http://www.jisc.ac.uk/openaccess

[7]Engineering and Physical Sciences Research Council (EPSRC) Policy on access to research outputs - http://www.epsrc.ac.uk/about/infoaccess/Pages/roaccess.aspx

[8]Directory of Open Access Journals - http://www.doaj.org

[9]Harvard Open Access Mandate - http://osc.hul.harvard.edu/policies

[10]Princeton University accepted Open Access Proposal - http://www.cs.princeton.edu/ appel/open-access-report.pdf

2004*b*) found that the vast majority of authors who are currently non-OA would submit their work into an OA archive if mandated to do so. For the author, earlier publication in an OA archive has been demonstrated to provide a greater number of citations, thus assisting with the "publish or perish" problem (Lawrence 2001*b*, Antelman 2004, Brody & Harnad 2004, Eysenbach 2006).

2.3 Online Dissemination

With the World Wide Web (Web for short) now available in most countries worldwide, information on any subject can now be accessed with just a few key presses. Correspondingly if a piece of information is not available on the Web, this can now be disseminated without the need for prior permission or review. In this way online dissemination seems to contradict the tightly peer-reviewed world of scholarly communications. The challenge becomes how to join these two environments in order to take advantage of the quality control that peer reviewing provides whilst being able to utilise the ubiquitous coverage of the Web in every home and workplace.

One of the earliest examples of an online only repository attempting to merge these two environments is arXiv[11], started at Los Alamos National Laboratory in 1991 as an email distribution list for pre-prints. At this time it was focused only on High Energy Particle theory and had the email address hep-th@xxx.lanl.gov to which users could subscribe to obtain a feed of information about the latest publications as they were submitted. This service started as an entirely manually run operation however, it soon grew to a subscription of over 3600 users and started to expand its remit into other areas of study (Ginsparg 1994*b*). Ginsparg, the original developer of the arXiv service, predicted that paper based publication and its funding model would not survive with competition from the electronic realm. With xxx.lanl.gov (the then arXiv), ever expanding submissions began to be accepted via methods including ftp and the then young World Wide Web (Ginsparg 1994*a*). Ginsparg viewed electronic publishing as the future and asked not when it was going to happen but how quickly.

Fast forward 14 years, and arXiv is now ranked 5th in the world standings for digital repositories[12] and holds content pertaining to mathematics, physics, astronomy, computer science, biology and statistics. Access to arXiv is now almost entirely Web based and being indexed by all the major search engines makes this content reasonably easy to find. Ginsparg's predictions of an electronic only publishing environment have not come true yet however Lawrence's question of being "online or invisible" holds some merit with studies which look at the number of citations towards publications made available online (Antelman 2004, Brody & Harnad 2004, Eysenbach 2006).

[11] arXiv.org - http://arxiv.org/

[12] Ranking of World Repositories: Top 800 Repositories (2010) - http://repositories.webometrics.info/top800_rep.asp (Jan 2011)

ArXiv demonstrates how easy it is to create a niche service on the Web in order to communicate with a great many people, and assists with the access/impact problem while providing some level of online peer-review. The problem not addressed is that of finding material in a community about which a research has no prior knowledge. With 4 million new articles published online each year (2007 and 2008 figures from the OA repositories known by the Registry of Open Access Repositories (ROAR)) (Figure 2.4)), compared with the 2.5 million journal articles, the problem with finding the most relevant publication to any given search is compounded on the Web.

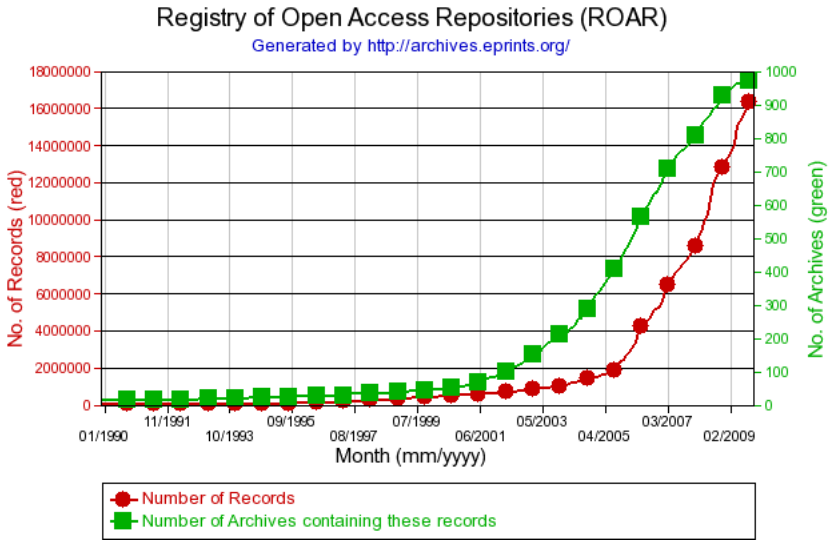

FIGURE 2.4: No. of Records in OA Repositories (source: http://roar.eprints.org)

Figure 2.4 reflects the growth in online availability of scholarly materials, however dissemination on the Web did not begin with digital repositories and online journals. It has taken a lot of work and realisation by people on how to host content online in a way which is able to merge the two communities of traditional journal and online publication. With less that 1000 entries known by ROAR (Figure 2.4), the growth rate shows the relatively young age of this practice. Conversely, it is likely that due to diversity and opportunities provided by the Web, many niche communities exist (of which those repositories listed in ROAR could be viewed as one) that are releasing their works online which are simply not known by this research.

In the remainder of this chapter looks at how sharing of scholarly information has moved from authors uploading articles onto dispersed websites, to the more federated process where traditional journal methodologies are being taken online. In the middle of these two exists the area of digital repositories which, while these could be seen as a competitor to online journals, can also be used to compliment them.

2.4 The Web - A Communication Revolution

In the early days of the Internet (the Web's infrastructure) Web browsers, search engines and carefully designed websites did not exist. At this stage FTP (File Transfer Protocol) was one key means for people to share materials including publications online. Using FTP, people could simply upload their files onto servers or other peoples computers such that they could see them. A lot of parallels could be drawn between FTP and techniques such as sending a printed copy or disk containing the material to a work colleague, FTP was simply faster ad more convenient. In parallel with FTP, email was being developed from existing intra-machine communication protocols such as SNDMSG (Peter 2010). By appending an @<host> to the persons username to whom a message was to be sent, an inter-machine communication protocol was developed. The first emails were sent in 1971 using custom applications before parts of the protocol were added to the FTP specification in 1972, thus allowing more widespread usage (Hardy 1996). Over the next two decades email evolved before seeing substantial usage as the internet became freely available to everyone.

The Web saw a major revolution 18 years later with the invention and subsequent widespread use of the HyperText Transfer Protocol (HTTP) and the HyperText Markup Language (HTML), originally proposed by Berners-Lee (1989). This enabled the addition of "splash" pages to your content which also linked to other content. The name "Web" comes from the very first browser invented to view the content and documents created (Cailliau 1995).

Although things like email and HTTP are relatively old protocols, the Web did not start to take off until the mid-nineties when more people became connected. Up until this point the number of websites could be sensibly counted, with there only being 2,738 in 1994 (Gray 1996), however with the ubiquitous coverage of the internet and home connections becoming cheaper the Web started to revolutionise the modern world (O'Neill et al. 2003).

In 1999, Lawrence & Giles (1999) estimated that the Web contained over 800 million pages, at the same time estimating that search engines only index the top 16% of these pages. One year later in 2000, Cyveillance.com (a business intelligence gathering firm) used their own proprietary system to estimate that there were two billion pages on the Internet as of July 2000. Additionally, they found that 7.3 million unique new pages were being published each day. Moving forward to 2005 and the search engine Yahoo claimed that they had 19.2 billion documents indexed, followed by Google in 2008 who hit a milestone of having knowledge of 1 trillion unique URLs (Alpert & Hajaj 2008). Each of these figures is significantly different due to the estimation techniques used. With Lawrence estimating that search engines online index 16% of the Web, Yahoo's 19.2 billion Web pages indexed can be used to implies that there were over 120 billion Web pages in 2005. Alpert's count of 1 Trillion URLs maybe a little generous, however

this figure may also account for pages in the non-indexable deep Web, estimated to be 400 to 550 times bigger than the indexable Web (Bergman 2001). Either way, there is little doubt that the Web is the fastest growing dissemination medium of all time. This is mainly due to the early decision to make it a free and open publication platform.

Looking at the growth in publication media over the years, it is possible to compare the growth of the Web to the total number of book publications. Here it is necessary to assess the number of websites, not the number of Web pages. This means that a site, consisting of a number of pages, is metaphorically equivalent to a book. An approximation of total number of book titles and publications for all time comes to 65 million worldwide[13]. This equates to approximately 1,100,000 new publications per year currently[14]. The number of websites surpassed the total number of book publications after only 10 years[15].

Admittedly there are a lot of websites out there which would never make it to book form, however it is interesting to draw the parallels purely on the basis of how difficult it can be to find the information you are looking for. Again this can be paralleled with how difficult it may be to find a book, of which there are so many less. Thus if you can always find something relevant on the Web, should these techniques be applied retrospectively to finding books?

2.5 Putting Things on the Web

Academics and scholars have been freely sharing their information with others for a great number of years. Before the internet came along this took the form of written letters accompanied by copies of published works which would be sent to peers and colleagues who may reference their work. A classic example would be Charles Darwin who wrote more than 15,000 letters in his lifetime on subjects including religion, gender and most famously science. Although this seems like a huge number of letters, this is only due to the pre-conception that letters are quite lengthy prose, where as if a person sends or receives an average of 42 emails a day[16], 15,000 could be sent in just over a year.

With the invention of the Web, 'self-archiving' content not only became much easier to publish and access, but also allowed there to be a much greater potential audience. Pinfield & Gorman (2004) states "The self-archiving of publications has the potential to revolutionise scholarly communication, making it more efficient and effective".

[13]How much Information? (2000) - http://www2.sims.berkeley.edu/research/projects/how-much-info/print.html

[14]UNESCO - http://www.uis.unesco.org/TEMPLATE/html/CultAndCom/Table_IV_5_Europe.html, collated at http://en.wikipedia.org/wiki/Books_published_per_country_per_year

[15]Netcraft October 2008 Web Server Survey - http://news.netcraft.com/archives/2008/10/29/october_2008_web_server_survey.html

[16]Survey Finds Workers Average Only Three Productive Days per Week - http://www.microsoft.com/presspass/press/2005/mar05/03-15threeproductivedayspr.mspx

Self archiving has the advantage that the author can still have control over both the publication medium (i.e. the website that their article is represented by) and also how long it remains there. The author is also able to retain the copyright on the work and even charge some nominal fee for access if they see fit. Charging is not currently a widely used practice, however authors may choose to enforce micro-payments, which then fund the long term storage and preservation of the article.

Self-archiving faces many problems, while it helps in the access/impact problem, maintaining a well indexed and high ranking site on the Web is not that simple. Also with a number of alternative mechanisms now available, including institutional repositories, these can present a more effective solution to support the online dissemination of scholarly communications.

2.6 Digital Repositories & Green-OA

Digital repositories typically take on the role of disseminating information to different parties by either providing a specific set of services or being designed to interact well with Web technologies and be easily indexable by search engines. They are primarily designed to take raw data, in this case publications from authors and enable management and dissemination of these without the author having to worry about writing Web pages and services themselves. There are many types of systems designed to take resources and perform this type of management[17], separated by the type of community at which they are aimed. Essentially all fall under the title of being Content Management Systems (CMS's). Handling different types of content and providing a different set of services is what separates them from other systems (Crow 2006).

In the area of scholarly communications online publication was initially about granting easier access to publications which are also published in some other "official" form. Other than the fact that at the time printed publications were the only form of scholarly publication which were rated by impact factor (see Chapter 3), they were also viewed as a well controlled and trusted medium against which to reference and cite. With the Web being a dynamic medium against which to cite, there is a risk that cited content can change or simply disappear (Dellavalle et al. 2003).

The Open Access movement went some way to help institutions and funding bodies alike to realise the importance of gathering together their research outputs. For both of these parties, having a record of the publications and output of researchers provides an excellent reporting tool and has been shown to help in scholarly research assessment exercises (Harnad 2006*a*, Day 2004, Carr & MacColl 2005).

[17]SPARC Repository Resources - http://www.arl.org/sparc/repositories/

Lynch (2003) defines an Institutional Repository (IR) as "a set of services that a university offices to the members of its community for the management and dissemination of the digital materials created by the institution and its community members". In the IR community this has become the main quotation which IR managers refer to when talking about their repository, however the reality of what the IR is actually used for can be quite different. Lynch's definition is broad enough to also encompass not only fully published and peer reviewed works but also the data and intermediary work which forms part of this or other research. Lynch's vision of an IR is one containing the "intellectual works" of an institution, extending beyond publications to include teaching material, source data and resources relating to the wider activities.

Many institutions now provide a repository in which research outputs can be gathered, as can be seen from those listed in the Repository of Open Access Repositories (ROAR). While some repositories have seen great levels of success, there still exists a barrier between researchers publishing their work as both a formal publication in a journal as well as in an open access institutional repository (Swan & Brown 2004*b*). In many cases this can be attributed to the confusion of the limitations of Green-OA publishing where authors are permitted to freely publish a pre-print of their work online. Typically a pre-print is a fully edited version which has not yet undergone the transformation into the final publishing template, thus in some cases the pre-print maybe easier to read due to being in the authors chosen style rather than that of the publisher. At the point of publication the amount of effort required for both publisher and pre-print publication may be high. At the time of publication, the researcher may view the journal publication copy as the important one and simply forget about the benefits of also depositing an open access version elsewhere (Swan & Brown 2004*a*).

From Lynch's broad definition (Lynch 2003), some confusion may exist with what an IR should accept, what it should do with the content and how it is preserved for future access.

Submission policies to IRs can also lead to confusion and lack of feedback being generated to encourage further submission by authors. In many cases an institution will want to have very close control over their external image and thus has a requirement to audit materials which go into their repository. This applies an often under-specified editorial buffer to the content being published in the repository, delaying the publication time and contradicting the reason to put the content in the repository in the first place. With a lot of effort being put in by the author to get their works published in a high controlled peer-reviewed journal, it seems a little backward to do further institutional level editing which may not be done by those who are experts in the area.

The cost of establishing an IR will also vary widely based upon the different types of supported content as well as the level of auditing which is required to take place on items before they are accepted (Jones et al. 2006). Compared to self-archiving, an institution

will often spend a lot of time carefully planning, specifying their requirements and testing prior to finally going live with a usable system. In some cases, the barrier of entry can simply be in setting up a piece of free software such as EPrints[18] on a machine with a suitable amount of storage to contain the intended resources. At the other end of the spectrum the amount of customisation required could price the institution out of actually getting a repository established. Alternatively, it could be that the number of people employed to maintain and attempt to gather materials could end up costing the institutional a large amount of money per item, something never likely to be revealed in formal reports, however with guide figures of $5 per item, some large repositories claim upwards of $200,000 a year in running costs (Schöpfel & Boukacem-Zeghmouri 2010).

At Southampton, in the School of Electronics and Computer Science, the EPrints Repository adopts an open injest and retrospective editing policy, more in line with Web publishing paradigms. It is up to the members of the school to audit the content of the repository collaboratively. With the repository ranked 17th in the world, according to the same measures which rated arXiv 5th[19], gives a suggestion that this type of policy can work well at an institutional level. As well as collaborative control, this success is also partly due to a number of services offered by the repository which directly benefit the authors and depositors, including well structured publication lists which can simply be included in other pages such as a researchers CV, or personal Web page. Another more recent addition to the EPrints software has been a way to track the number of citations that a publication receives (using the institutional subscription to services such as WoS), feeding this information back to the authors via existing interfaces. In addition, many other positive feedback mechanisms including citation counts and download metrics, outlined throughout this work, are also offered via the same platform[20].

A more recent problem concerning digital repositories for scholarly materials, is the decision as to whether these repositories should be purely subject based rather than institutional. If one of the key aims is for researchers to be able to disseminate their research to more people in their field then why should that field not have its own centralised repository. Should repositories be similar in nature to that of journal publications and only concentrate on a certain field of study or area of research? Perhaps the best way to answer this question is to look at open access and online only journals and their popularity.

[18] EPrints Software - www.eprints.org/software

[19] Ranking of World Repositories: Top 800 Repositories (2010) - http://repositories.webometrics.info/top800_rep.asp (Jan 2011)

[20] ECS EPrints - http://eprints.ecs.soton.ac.uk

2.7 Gold-OA and online-only journals

With journals typically held in the highest regard within the scholarly community (Fry et al. 2009). Having these as open access may prove beneficial both to help the serials crisis, whilst also increasing the availability of research and chance of gaining more citations.

With the Web making the cost of publication virtually free, the money saved through not having to print and distribute should be substantial. Harnad (1995) looks at the cost savings and the options being provided by current publishers who are making paid journals available online and concludes that these savings may not be being passed on. Publishers of "paid" journals have started making their material available in an online form for an small extra cost on top of the existing print subscription (around 10%), however an online only subscription costs about 75%-85% of the printed subscription and it is this amount of money which should be saved (Harnad 1995). The key benefit of being a publisher in the area of scholarly communications is that the material is already peer-reviewed, thus the main cost should then come in the editing and translating of the material into the publication format, which could be reduced by simply not performing this operation.

The perceived value is in the collating of information and publishing under the trusted banner a journal provides. Authors (and libraries) are likely to read and subscribe to high prestige journals, a model which dictates the desired publishing targets of authors. The Journal Citation Reports (JCR)[21], published each year be Thomson Reuters, represents a well established mechanism which uses the Impact Factor to calculate the prestige of each journal. Due to the overheads of collection and verifying all the information there is a charge to access the JCR information in different levels of detail. Thomson also have a well defined policy on entry for new journals into the registry, as such at the end of 2008 there were just over 6500 indexed journals relating to science.

In 2002, Testa & McVeigh (2004) looked at the impact which Open Access journals were having in the community (results also published in McVeigh 2005). At this time, the WoS index contained just 148 open access journals out of the 5876 total, equating to just 2.52%. In order to rank the popularity of these journals the Impact Factor was used to place each within a percentage bracket, thus the lower the percentile value the greater the level of impact of the journal. In total only 9 OA journals (of 587) ranked in the top percentile category. Figure 2.5 shows the result of this survey, here OA journals trend towards being in the less impact percentiles with the mean at 39.77%.

"Overall, 98 (66%) of the OA journals rank below the 50th percentile. Relatively few, around 6%, are in or above the 10th percentile." (Testa & McVeigh 2004)

[21] Journal Citation Reports - http://wokinfo.com/products_tools/analytical/jcr/

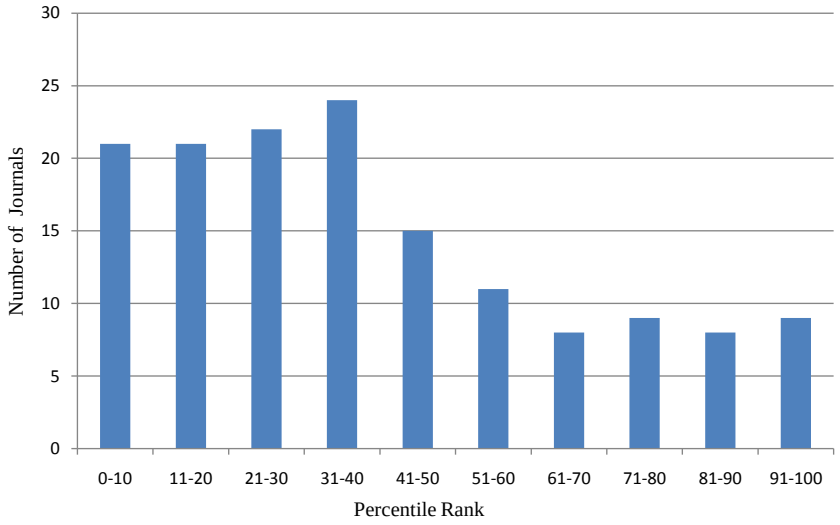

FIGURE 2.5: Distribution of Open Access Journals in WoS in 2002 (Lower percentiles are best) — Reproduced from data published by Testa & McVeigh (2004) and McVeigh (2005)

With open access being a relatively new concept back in 2002, even this result of having a few journals in the top quartile is a significant one. As part of this study, it was possible to repeat the same study and find there are now 355 open access journals listed in the JCRs (found by cross referencing records in the JCR with those contained in the Directory of Open Access Journals (www.doaj.org)). 355 out of the now total 6619 records in this index represents 5.36% of the index now being OA journals, an increase of 2.82% on the previous figure. Figure 2.6 shows the latest set of figures up until the end of 2008 and how OA journals are now spread amongst the rest.

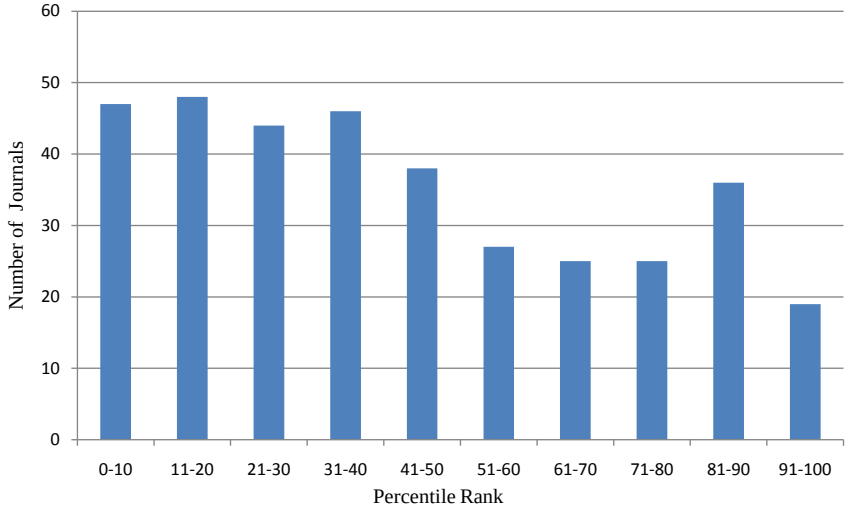

FIGURE 2.6: Distribution of Open Access Journals in WoS in 2008 - 355 OA Journals (self study data)

While there is no significant difference between Figure 2.5 and Figure 2.6, the fact that

as more OA journals come online they are managing to distribute themselves among the percentiles is quite impressive. If the barrier to entry into this market is much lower than for their printed contenders, then it could be anticipated there would be a greater number low impact journals added. However between 2002 and 2008 the mean percentile has fallen from 66.83% to 62.63% showing that the opposite trend is true.

To analyse this situation further it is necessary to take a look at the citation patterns. Testa & McVeigh (2004) hypothesised that open access articles would be cited sooner and to a higher level, due to the greater potential audience. They found that "OA journals have a broadly similar citation pattern to other journals, but have a slight tendency to earlier citations", a finding backed up by other studies (Brody & Harnad 2006, Eysenbach 2006).

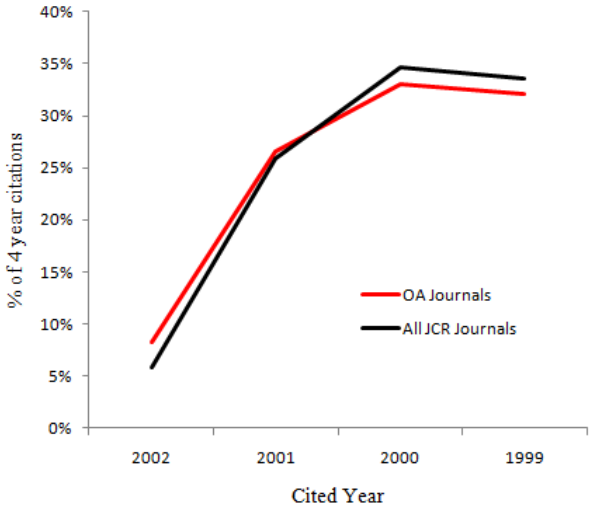

FIGURE 2.7: Cited Journals - Number vs. Age of articles cited by 2002 publications (extrapolated from McVeigh 2005)

Figure 2.7 is extrapolated from the findings of McVeigh (2005). By looking at only papers published in 2003, McVeigh analysed the citations and established if these were directed towards an OA or non-OA journal. Additionally the age of the cited publication was recorded and any citations towards publications older than four years were eliminated. Finally, by translating these citations into percentages of the total number of four year citations, a balanced comparison of citations towards both OA and non-OA publications can be made (Figure 2.7), showing that the citation patterns are broadly similar. Unfortunately the data behind this study is not as freely available for it to be updated, however it is possible to draw parallels with average citations each year that the different types of journal obtain.

From Figure 2.6, a good spread of OA journals (by impact) amongst all the non-OA journals is observed. Thus comparing the average number of citations each Journal receives on a yearly basis should reflect the average percentile figure of 62.63% found earlier, suggesting that OA journals will receive less citations on average than paid

journals. Using the data available from the Web of Science Journal Citation Reports, this is found to be true. Figure 2.8 shows this difference clearly with OA journals receiving only around 1250 citations a year compared to the 4000 of each non-OA journal.

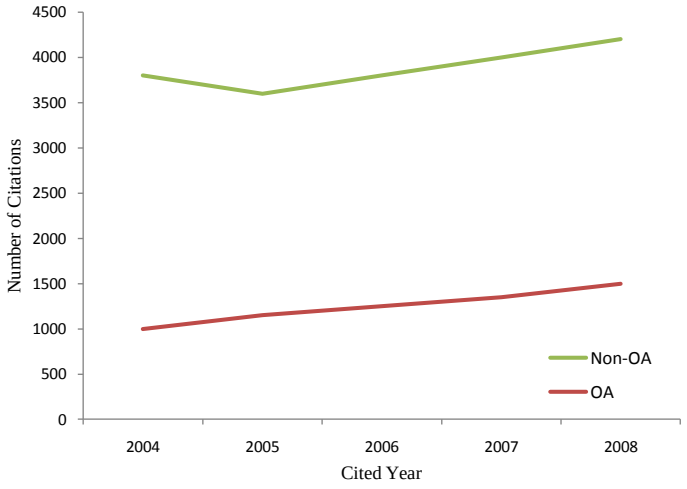

FIGURE 2.8: Citation per year, OA vs. Non OA Journals (source: Web of Science Journal Citation Reports)

Figure 2.9 shows that although OA journals receive less citations, the growth rate in the number of citations that OA journals are receiving is generally higher than their non-OA counterparts. Thus there is a growing popularity in citation of OA journals.

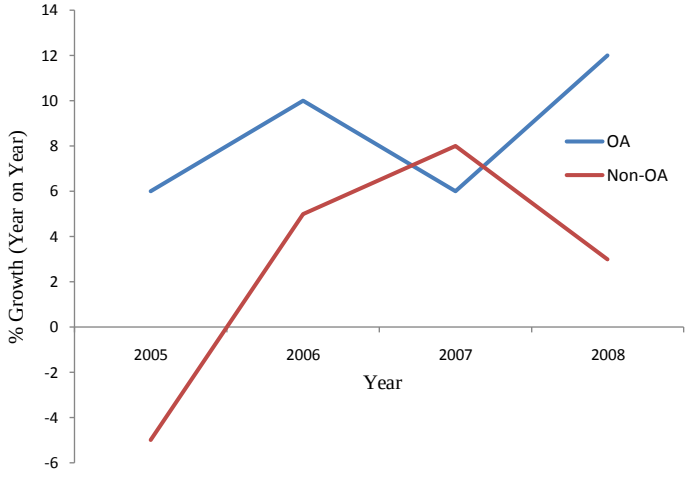

FIGURE 2.9: Citations growth rates, OA vs Non-OA Journals (source: WoS Journal Citation Reports)

Of those now among the top percentile of journals are the OA journals, the Public Library Of Science (PLOS)[22] have an author-pays model where authors are charged on acceptance of a publication, post peer-review. Few authors would incur this cost, instead

[22]Public Library of Science - http://www.plos.org/

it would be passed to their funding project or member institution, who are probably very willing to pay for publication in high impact journals. Other publishers have followed a similar model, such as Springer[23] who allow authors to pay a charge to allow open access to their article, and ISP[24] who cover costs from industry grants.

Open Access journals have good backing and power to influence the market through being discoverable via well established mechanisms such as the Web of Science (WoS) index. These are also likely to attract the most attention from the authors themselves due to this very fact. Some challenges still exist however in exposing these journals alongside others, such that authors can make a physical choice between publication in an OA and Non-OA journals. By cross referencing ISSN numbers from the WoS index with those listed in the Directory of Open Access Journals (DOAJ) it is possible to find those journals listed in WoS which are open access compliant. This facility along with many others, such as cost of publishing, are not listed in the WoS index. Although not listing these factors does level the playing field by solely emphasising which are the high impact journals, it does result in the high impact open access journals being challenging to find. These findings also concur with those of Swan & Brown (2004*a*).

2.8 Summary

Online dissemination and indexing has changed the way that we find and obtain access to data as a whole and this paradigm carries to the area of scholarly communications. Author self-archiving, such as that achieved through allowing publications to be downloaded from a personal website, started the movement towards open access online publishing and this has now developed into controlled and stable publication environments, such as digital journals and repositories. Open Access (OA) can often be seen as a side issue, however with deposition mandates on the increase (Harnad 2006*a*) and OA publishing leading to quicker and potentially more citations (Brody & Harnad 2006, Eysenbach 2006), support behind the movement is continual and positive.

There is still a long way to go to achieve widespread adopted of OA techniques. In 2004 Swan & Brown (2004*a*) performed a survey of both OA and non-OA authors to find out why these authors choose to publish in these different ways and reflect on how they felt it affected their research. The first major finding was that non-OA authors have only just heard of (in the last two years) OA publishing as opposed to those now actively publishing in OA journals. Although the OA authors state that the levels and quality of peer-review and publication vary little from non-OA publication, there is a definite worry in the non-OA field that publishing in OA journals may limit the potential impact of the work (77% of non-OA publishers believed this was the case). Section 2.7 found

[23]Springer - http://www.springer.com/
[24]Internet Scientific Publications - http://www.ispub.com/

that at the time of the survey only 207 OA journals were listed in the Web of Science (WoS) index, thus finding an OA journal in which to publish authors work could have been an issue, as reflected by over 50% of non-OA authors in Swan & Brown (2004*a*)s' report. From Section 2.7, it is possible to observe that the mean impact of an OA journal is roughly the same as a non-OA journal, a figure which could be used as proof when addressing the 77% of non-OA authors who believe that OA journals carry less impact.

Both reports by Swan and Brown (Swan & Brown 2004*a*,*b*) also look at the publication of articles in a digital (eprint) repository and find that 39% of non-OA authors have also deposited their work in a digital repository. Although not a large figure, this is a very positive sign if 45% of non-OA authors have only just heard of OA publishing. The main problems with publication in an OA repository lie in the politics surrounding the copyright of an article, the ability to publish the same article or a recent pre-print in a OA repository, and the effort involved in this process. There is still much education to be done surrounding author rights and abilities and by the repositories to lower the barriers and amount of effort required to deposit an authors work (Carr & Harnad 2005, Swan & Brown 2005).

Further findings from the author survey carried out by Swan & Brown (2004*b*) include 92% of OA authors stating that free access to scholarly research was important, there is clearly a general consensus to OA publishing. 56% also stated that they were concerned about the rising costs of journal subscriptions to their institution. Authors also stated that they would be prepared to pay for publication in one way or another — in most cases either via institution or project funding. Pay for publication could suffer one problem however in that currently journals want to be high quality to attract paying customers, if the pay model reverses, are some journals only going to be attracting those who can afford publication rather than the authors who have something valuable to contribute?

The Updated figures in Section 2.7 show that the numbers of OA journals are increasing steadily, which should lead to authors having more opportunities to publish in WoS listed OA journals in their field of study. Swan & Brown (2004*a*) see this as one of the major stepping stones to OA publishing. Having a greater proliferation of OA journals should lead both to more OA authors, as well as more non-OA authors being informed about OA publishing. According to Swan & Brown (2004*a*), 47% of now OA authors were informed of the opportunity by a colleague.

Much the same applies to author self-archiving which adds a further route to OA publishing (Swan & Brown 2005). Here 49% of the people surveyed stated that they had self archived at least one article and of these, only slightly more had used a digital repository for this purpose rather than a listings or personal website. Of those who have not yet self archived, 71% remain unaware of the option, concurring with the findings relating to knowledge of OA journals.

So while self-archiving and open access publishing are clearly viable options, there is still a change in publishing mindset to be achieved. This could be either through the acceptance of OA journals and OA publishing as a sustainable way of achieving impact, or by justification of the effort involved in multiple deposit and pre-print deposit. The latter of these methods is supported by the green-OA policies which 90% of journals now have (Harnad et al. 2008), but there is little reflection of it happening. In turn, if the highest impact journals were OA journals, then the rest of the non-OA journals may have to re-think their business model, shifting it away from the current "trade" methodology.

Odlyzko (1995) predicted back in 1994 that the models for publication would have to change in time and that the role of publishers, editors and libraries would decrease. As a result Odlyzko predicted that these companies would either have to change, shrink or disappear. This seems like bad news for an ancient (in comparison to electronic publication) economic system however, by looking at the exponential increases which cover both publications and technology (Moore 1998), the one thing which is not likely to increase at the same rate, is the amount of money which institutions have to spend on journal subscriptions.

Chapter 3

Bibliometrics

Bibliometrics began as the statistical study of written documents and was initially termed 'statistical bibliography' (Hulme 1923) before the term 'bibliometrics' was proposed by Pritchard (1969). Bibliometrics encompasses many techniques for gathering information from, and about, written texts through statistical methods such as term frequencies. Such methods can be used for creating thesauri, exploring the grammatical and syntactic structure of a text as well as measuring impact. It is the last of these which is the main focus of this sections exploration of bibliometrics. By looking at the bibliographies of scholarly literature, a vast network of academic papers can be constructed from the links created by citations and footnotes. From this data it has been found that, among other things, it is possible to infer something about the impact of a publication.

Analysis of published material to ensure quality is an idea which started back in 1955 as the brain child of Eugene Garfield (Garfield 1955). His idea of rating papers, based upon the number of citations that paper received, came to him whilst trying to find reliable information about an area of study about which he had little knowledge. This work led to the establishment of Garfield's *Impact Factor* (Garfield 1972), which is still used today as a measure of journal popularity.

Since Garfield's original work, many (including myself in the case of this thesis) have looked at different ways to rank publications. Each technique applied either processes a different corpus or source of data, or attempts to reduce the time required to establish a figure for the impact of a publication. The sources of data have changed from being solely based on journals, to include other publication mediums such as conference proceedings, academic publications and more recently the world wide Web. Each of these changes has brought about new ways to look at the area of Bibliometrics and classify it differently, leading to terms such as Scientometrics, Webometrics and the more generalised Infometrics. Figure 3.1 (from Bjorneborn & Ingwersen 2004) attempts to indicate

the overlaps between many areas of study of impact, all of which can be classified under the title of Infometrics.

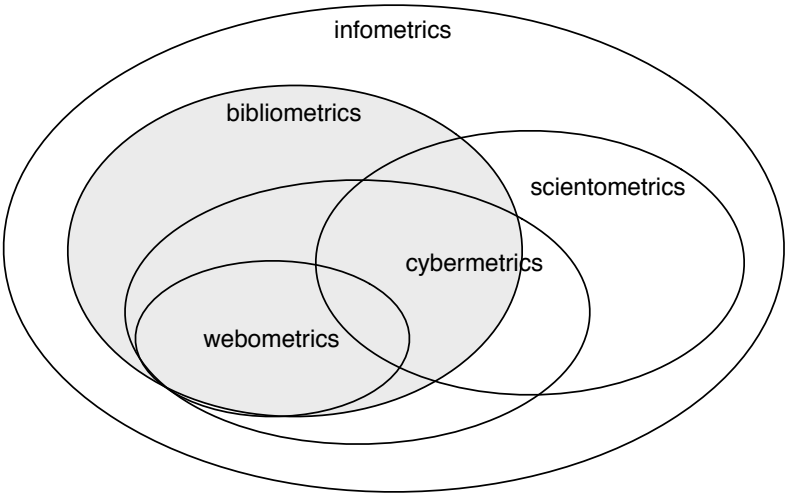

FIGURE 3.1: Relationships between types of Infometrics (from Bjorneborn & Ingwersen 2004) - Sizes of ellipses are for clarity purposes only

In Figure 3.1, each circle shows how influences have been taken from each of the previous research areas and then added to in order to create a more optimal solution for a new area. A more in depth look at the evolution from bibliometrics to webometrics is presented by Thelwall (2008).

This chapter looks at the development of bibliometrics since its inception and how this has evolved into what is commonly known today as Infometrics. The first part of this chapter focuses on early bibliometric techniques and the fundamental laws which form the basis of services such as the Web of Science (WoS). Importantly, it can also be demonstrated that many laws of bibliometrics are still relevant today and have influenced techniques to aid information discovery on the Web.

Metadata, "data about data", plays a key role in the study of bibliometrics relating to scholarly texts. This includes attributes such as the title, author and publisher. In bibliometrics one of the key pieces of metadata is the set of references which a paper contains. These references, or citations as they shall be referred to in this work, are the links between papers that authors establish to back up or introduce parts of their own work. Although in some cases citations and references are differentiated, broadly a citation is a reference to a source of information, often written in some shorthand form in the prose of a work. This shorthand version is then expanded in that works' references' section. Based upon this definition, the term citation will be used throughout this thesis when talking about references to other published works. Additionally, citable works are those which have been published in a citable form; footnote citations are ignored in the majority of cases.

A single citation represents the author exhibiting their intellectual honesty; it provides the reader an indication of the authors background knowledge and source for information. Conversely, a citation may also be used by an author to contradict points in their or others work; thus instances of negative citations also exist. Without negative citations, each one can represent a recognition by an author of the authority of another's work, and thus by deduction, the more citations a paper gets from different authors the more authority the referenced work may have. Collating together citation information from a number of publications enables analysis of impact by citation count, a practice commonly referred to as citation analysis (Garfield 1972).

Citation analysis works on the assumption that influential scientists and important works will be cited more than others. Thus highly cited publications are often regarded by consumers as influential and of a certain standard, suggesting the need for in depth review is not required. Studies have shown that citation analysis is able to effectively rate the impact of papers in the same way as the same set of papers being peer reviewed. Aksnes & Taxt (2004) and Meho & Sonnenwald (2000) both perform a similar study to each other where a set of publications were given to experts in the area, who were asked to rank them in order of perceived importance. The results produced mirrored the rank given by the citation counts of the same set of papers. While a positive correlation exists between citation analysis and peer review, it is not a perfect correlation.

Situations can be envisaged where a number of authors will intentionally make a negative citation in order to point out flaws in previously accepted works. Likewise, cited works may not yet be known to be of bad quality but are still cited. Thus Citation Count does not differentiate the context importance of the citing papers. A citation coming from an obscure paper has the same weight as one from a ground breaking, highly cited work (Maslov & Redner 2008). Oddly the idea of a weighted citation, where each citation source contributes a values of its perceived worth, is in widespread use on the Web, even though it was first recommended for use on publication material in 1976 (Pinski & Narin 1976). Pinski & Narin (1976) were the first scholars to note the difference between popularity and prestige, where a popular paper is highly cited and a prestigious paper is cited by other prestigious papers. When using citation data as part of critical assessment, it is advised that citation analysis techniques be used as an indicator and not to generate the process' outcomes (Moed 2009).

3.1 Citation Networks

The citation network of a particular publication grows over time as it is cited directly by other publications, shown here in Figure 3.2. Here the cited publication (shown on the left) is cited by four other publications. Citation links are represented as arrows linking the publications and time is shown passing from left (oldest) to right (newest).

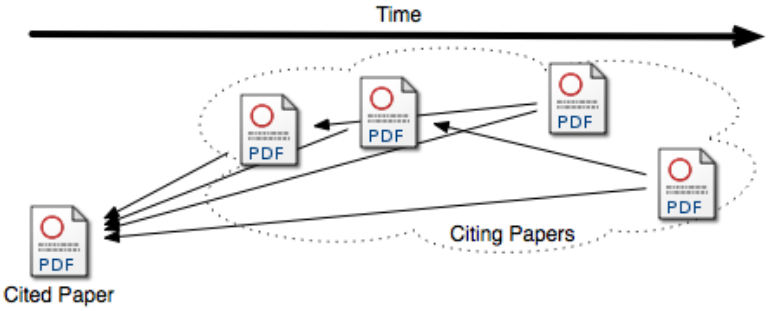

FIGURE 3.2: An example citation network

Although a citation network will continue to grow as new citations are obtained, there is a period of time which dictates the most accurate measurement for the maximum impact rating the article will achieve. This point, referred to as the peak citation rate, represents the time period when the article obtains the most citations in the shortest amount of time. In many subject areas the time taken for a publication to reach its peak citation rate is predictable dependant on how researchers in a subject area operate. For the majority of subject areas the time that an article takes to reach its peak citation rate is around three years (Moed 2005).

3.2 Laws of Bibliometrics

Bibliometrics commonly refers to three laws: Zipf's Law, Bradford's Law and Lotka's law and all three laws share a common relationship to the mathematical Power Law. The Power Law dictates that the frequency of an event will vary dependant on the power of some attribute. In order to observe which events are most significant, the Pareto distribution (a power law probability distribution), also known as the 80:20 rule, outlines the relationship between two variables, where few values occur with a high frequency (making up 80% of the samples), while there is a long tail of values which have low frequency. Using the 80:20 rule it is then easily possible to divide significant results from the long tail. Such an example can be seen in the distribution of people in the world. The majority of the population is focused into a small number of big cities while the minority which remain are scattered widely[1].

Figure 3.3 shows a Pareto distribution, outlining the 80:20 split and the long tail of low frequency results. This distribution is common in many areas of bibliometric study as is the case throughout this thesis. A more in depth history on the Power Law is outlined by Mitzenmacher (2004) while other uses and related laws can be found in Bookstein (1990).

[1]Percentage of global population living in cities, by continent (Guardian Data Blog) - http://www.guardian.co.uk/news/datablog/2009/aug/18/percentage-population-living-cities

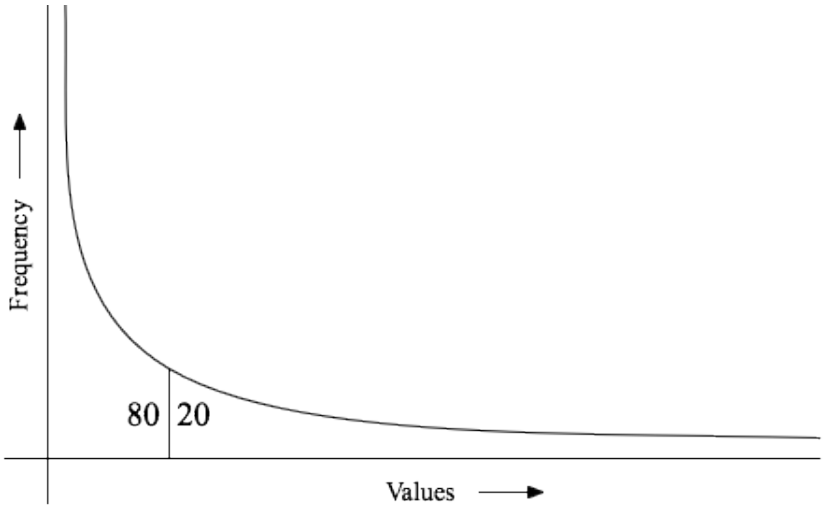

FIGURE 3.3: Example Pareto distribution exhibiting the 80:20 split point.

3.2.1 Zipf's Law

Zipf's work focuses primarily on human behaviour and the use of language (Zipf 1949, 1932).Although Zipf did not look at citations or impact of articles, he did look more broadly at the principle of occurrences, specifically the number of occurrences of each word within a given publication. Zipf found that the most frequent word will occur approximately twice as often as the second most frequent, the second most frequent twice as many as the third, the third twice as many as the fourth and so on.

Correspondingly, the rank of a word (by number of occurrences) times the number of occurrences will be constant (Potter 1988). Equation 3.1 shows a variation of Zipf's Law where the rank occurrence of a word (R_w) can be worked out by dividing a constant (in this case 1) by the word frequency (F_w).

$$R_w = 1/F_w \tag{3.1}$$

To demonstrate this principle, Table 3.1 shows the frequency counts for the top 5 terms which appear in this chapter. Furthermore, when the top 300 words and occurrences are plotted on a graph (see Figure 3.4) the relationship to the Power Law becomes clear.

Term	Frequency
the	581
of	396
a	289
to	247
and	193

TABLE 3.1: Top 5 terms in this Chapter

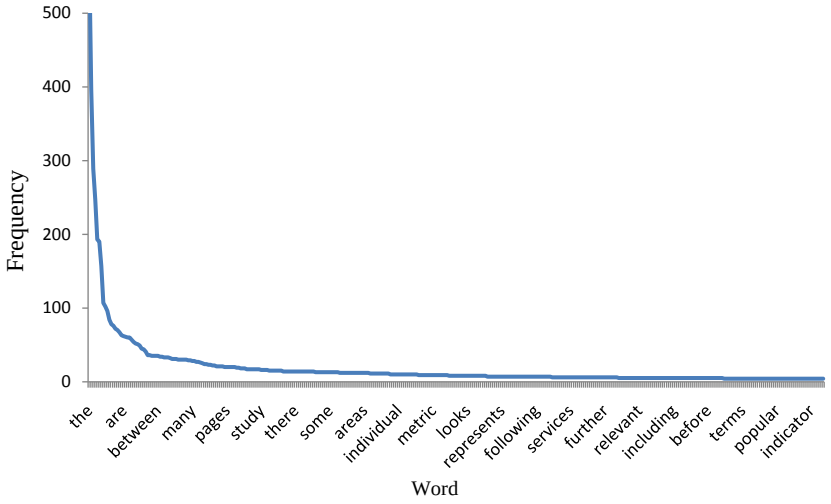

FIGURE 3.4: Term occurrence of words contained in this chapter demonstrating Zipf's Law

By taking Zipf's Law and applying it to citation count, the rank of a paper can also be found using Equation 3.1 and substituting word count for citation count (Redner 1998). Likewise, the same principal applies to citation towards journals indexed by the Web of Science. Using data from the Web of Science Journal Citation Reports (JCRs)[2], the top 80% of citations are towards the top 1389 journals with the remaining 5230 accumulating the other 20%. Figure 3.5 shows the Zipfian distribution present in journal citation metrics for the JCR dataset from 2008; note that for clarity this has been limited to the top 300 journals, plotting all 6619 at this scale would reveal a very long tail and cloud the result.

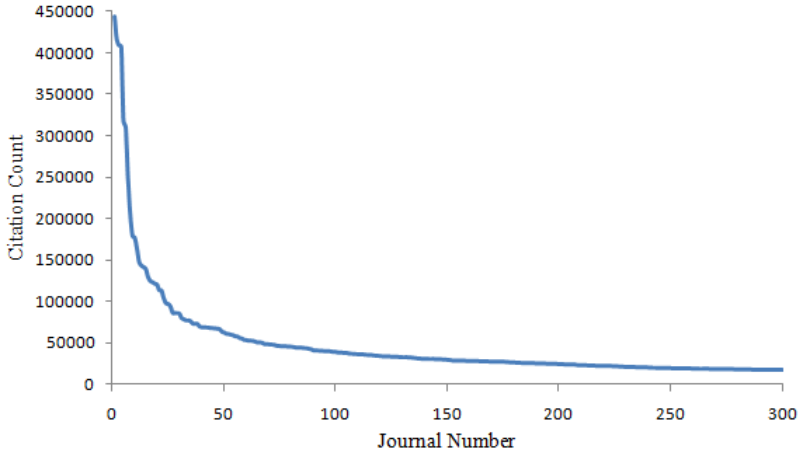

FIGURE 3.5: Citation distribution for journals in the Web of Science Index (2008 Journal Citation Reports)

[2]Journal Citation Reports - http://wokinfo.com/products_tools/analytical/jcr/

Figure 3.5 shows that Zipf's Law applies to many areas of bibliometric study, being relevant not just to individual articles as demonstrated by Redner (1998), but also to journal popularity.

3.2.2 Bradford's Law

Bradford's Law (Bradford 1934) looks at distribution of papers among journals giving a $1 : n : n^2$ relationship in article distribution. Using this distribution, Bradford is able to create a guideline which determines the number of core journals in any given field. This $1 : n : n^2$ relationship states that journals in a single field can be divided into three segments:

- 1 - A core set of journals, which contain approximately 1/3 of the total number of articles.
- n - A second set, containing the next 1/3 of all articles.
- n^2 - A third set, containing the remaining relevant articles.

Using this formula, to have access to all articles would require n^2 journals, however a much smaller number of journals (1) will contain 1/3 of the total articles. Bradford does not consider the impact of these journals and thus the $1/3rd$ of articles may in fact not be the core articles in high impact journals.

By looking at the number of citations towards journals, rather than the number of articles, it is also possible to find the same core set of journals from the available Web of Science data. In order for Bradford's Law to hold in this study, 1/3 of all citations need to be directed towards a small set of journals, the next 1/3 to a bigger set (n) and finally the last 1/3 to a set of size n^2. Using the Journal Citation Reports from 2008 as evidence (which list journal citation figures), it was found that 1/3 of all citations towards journals indexed by WoS were directed at only 116 of the 6619 journals. The next 1/3 were directed to the next 617, leaving 5886 journals to accept the final 1/3. While this does not hold exactly to Bradford's $1 : n : n^2$ relationship (there would need to exist over 380,000 journals), it is still the closest approximation that can be represented by a simple equation.

3.2.3 Lotka's Law

Lotka's Law (Lotka 1926) is another variation of Zipf's law looking at the frequency of publication by authors in a given field. Lotka states:

> The number (of authors) making n contributions is about $1/n$ of those making one; and the proportion of all contributors, that make a single contribution, is about 60 percent.

From this it is possible to deduce that only 15% of authors will have two publications ($1/2^2 \times 60\%$), 7% will have three and less than 4% will have published four or more papers. More recent studies by Egghe (2005) have found that a slightly more accurate formula can be deduced by plotting Lotka's Law as a Power Law distribution.

Changing Lotka's Law to conform strictly to Power Law principles would only incur a minor change. Taking 60% as the starting point raises the value at which the long tail starts, such that approximately 11% and 6% of authors will have published three and four papers respectively.

Recent studies have found Lokta's Law to have many uses; López-Munoz et al. (2003) uses Lotka's Law to analyse the participation, productivity and collaboration indexes related to authors. Wilson (1999) looks at the usage of Lotka's Law in the area of infometrics, realising that scholarly communications have changed as a result of the Web. Wilson's work re-enforces the overlap between bibliometrics, the study of written documents, and infometrics, the study of information in general.

3.3 The Science Citation Index

Garfield came across a problem which had a potentially simple solution. While looking for some authoritative and reliable information on a topic about which he had no knowledge, he realised that citations between works were one of the only ways of inferring authority. He recognised that the reader will follow the path of the author through their chosen citations, assuming that these cited works are a source for authoritative information (Garfield 1955). A problem exists however if you have no point in a subject area at which to start your search.

Garfield established that a good place to find citation information was in academic journal publications. Using the bibliographic information found in academic journals, namely the citations and references, Garfield established the Institute for Scientific Information, now called Web of Science (WoS)[3] . WoS collates, processes and stores citation data pertaining to academic journal publications. It then provides and applies a series of metrics, in order to produce a series of reports detailing the impact of the journals in its index.

From an early stage, Garfield realised how difficult the task of gathering all this information together accurately would be, mainly due to the variety of ways in which a citation

[3]Web of Science - http://thomsonreuters.com/products_services/science/science_products/a-z/web_of_science/

is presented (Garfield 1972). He realised the need for every author and publication to be disambiguated from each other, allowing for cases when names and titles become abbreviated or shortened. Garfield also realised how different subject areas used different types of citation technique, thus some citations would retain full journal titles while others would abbreviate or refer to a journal by a code name.

Web of Science, provides access to bibliographic information, author abstracts and cited references found in over 3,700 of the world's leading scholarly science and technical journals covering more than 100 disciplines. More recently, the WoS index has become available online in an expanded format covering more than 5,800 journals. The amount of carefully controlled and historical data stored in the index, has enabled its widespread use in measuring the impact of academic research, backed by many funding councils (Day 2004). More recently, improvements in automated data processing techniques have enabled a number of alternative services to become available including Scopus[4] and Google Scholar[5], each of which can have benefits in different areas of study as a source for citation data (Meho & Yang 2007, Falagas et al. 2008, Bakkalbasi et al. 2006).

The Web of Science index brings together data from many disparate locations and subject areas into one central store, thus providing a service to its users which is well specified and maintained. Services which provide this sort of specifically defined and reliable functionality, become used as sources of information key to the performance assessment of an institution, group or individual researcher.

Each year WoS publishes a set of Journal Citation Reports containing the up to date count for that year of the number of citations each Journal has received. By analysing this and combining it with other information these reports also list journal impact scores of various types. The following section introduces a number of the impact factors metrics applied in these reports before a comparison shows how all the metrics are in fact quite closely interrelated in their usage by the WoS.

3.4 Impact Factors

In Chapter 2 the Web of Science Journal Citation Reports were used to assess how Open Access (OA) journals were fitting in alongside the existing non-OA journals. The Journal Impact Factor was the metric used throughout this study and is one of many used in the JCRs. As well as Journal Impact Factor, this section introduces many of the other impact factors and the various properties of each. Broadly speaking impact factors can be categorised into groups, defined by what they are trying to classify. There are also two main types of impact factor, weighted and non-weighted. Weighted impact factors change the value of a score received from a citing article to apply some sort

[4]Scopus - http://www.info.sciverse.com/scopus/
[5]Google Scholar - http://scholar.google.com

of prestige score to each citation. Non-weighted metrics regard all citations as having equal value. Additionally, weighted metrics can also take account of link pollution; the process of linking between websites for no reason other than to try and gain popularity and prestige in search results.

Garfield's Impact Factor, the first Journal Impact Factor (JIF) to be introduced in this section, is designed to rank journals based upon the number of citations they receive per article. The key aspect of this algorithm is that it only considers articles published in the last two full years, and citations from the subsequent year, thus reflecting a fairly recent trend.

Eigenfactor is a similar metric to Impact Factor designed to indicate how long a researcher is reading a journal. It considers a five year time frame rather than two. The main difference to Impact Factor is that Eigenfactor is a weighted algorithm where citations from high ranked journals hold more significance. An Eigenfactor is calculated using a normalised cross-citation matrix (where journal self citations are removed) and an influence vector. In many cases the Eigenfactor and Impact Factor may resolve to be similar, due to the simple fact that popular journals get read and subsequently cited.

Although journal ranking has been around for a number of years, Seglen (1997), Lotka (1926) and Garfield himself (Garfield 2005), all question the applicability of using Journal Impact Factor metrics as a means of locating authoritative information. The metrics act as a good guide when ascertaining a journals impact, however should not be used (in their current form) as a method to deduce the impact of individual articles and authors. As per Zipf's and Bradford's laws there is expected to be a small corpus of highly cited papers in a number of journals. In a study, Seglen (1997) finds that a small fraction of the publications in a journal are responsible for the majority of the citations. Due to the intrinsic relationship between the papers and the journal in which they are contained, the remainder of the publications which are of lower citation count still appear as of high relevance.

Article Impact Factor (AIF) can be viewed as a possible way to break up the intrinsic relationship between high impact journals and their papers. Using AIF based algorithms allows individual impact scores to be calculated for each article regardless of publication medium or journal. Such mechanisms are becoming more applicable as electronic, open and distributed publication establishes itself.

Article Influence Score (as used in the Journal Citation Reports) sounds like an example of a AIF metric however, it is simply used to calculate the average article impact score from the eigenfactor score, thus it is still based on the journal impact and therefore a Journal Impact Factor (JIF) algorithm. Due to the free-form structure of the Web, AIF type algorithms are use when both rating sites and individual pages, PageRank the Hubs & Authorities techniques are looked at in Section 3.6.

Lastly, this section looks briefly at h-index, designed to rate an individual author regardless of the publication medium. Although all citations carry equal weighting, there is a threshold value for number of citations to separate those who publish influential papers, from those who simply publish a lot.

3.4.1 Garfield's Impact Factor

Web of Science uses Garfield's Journal Impact Factor (JIF), first introduced in Garfield (1955), to give each journal an impact score. The JIF equation, as shown in Equation 3.2, is calculated based upon a two year rolling period. Thus the JIF of a single journal is based upon the number of citable publications in that journal over two previous years ($\sum_{t-2}^{t-1} C_t$) and the number of citable articles published in those years ($\sum_{t-2}^{t-1} P$). By simply normalising these values, dividing one by the other, the Journal Impact Factor (I_j) can be calculated.

$$I_j = \frac{\sum_{t-2}^{t-1} C_t}{\sum_{t-2}^{t-1} P} \tag{3.2}$$

Garfield (2005) now realises many problems with this system and views it as a mixed blessing. Seglen (1997) also gives a critical view of the JIF, stating why it does not work on many layers. One of Seglen's findings enforces a method by which the JIF system can be abused. In this case a low rating author can gain a quick boost in perceived impact by getting published in a high impact journal. Seglen (1997) found that "15% of the articles [in three biochemical journals] accounted for 50% of the citations, and the most cited 50% of the papers account for 90% of the citations". Thus a low ranking author who is not in the top 50% of papers is achieving the same impact score even though they are accruing less than 10% of the total citations in that journal. Taking this one step further a journal can be accused of trying to raise their own rank through self citation (Fassoulaki et al. 2000).

3.4.2 Eigenfactor

Following on from the work of Lotka (1926), Eigenfactor (proposed by Bergstrom (2007)) provides a means to calculate the journals containing the top $1/3rd$ of papers. Logically these top $1/3rd$ will be those most read and subsequently most cited. To calculate Eigenfactor requires two matrices, one detailing the citations between journals (with self citations removed) over a five year period, and another detailing each journals influence (the influence vector).

A normalised cross-citation matrix (named H) is calculated from the matrix detailing the number of outgoing citations between journals (named Z), and normalising this by

the total number of outgoing citations from each journal. When considering a 5-year cross citation matrix Z, the entries for 2006 in this matrix would be:

$$Z_{i,j} = \begin{matrix}\text{Citations from journal j in 2006 to} \\ \text{articles published in journal i during 2001-2005}\end{matrix}$$

This number of citations Z is then normalised by the total number of outgoing citations from each journal $\sum Z_{k,j}$ to create the normalised cross-citation matrix as shown in Equation 3.3 (from West et al. 2008).

$$H_{i,j} = \frac{Z_{i,j}}{\sum Z_{k,j}} \tag{3.3}$$

The influence vector ($\pi*$) is calculated from the article vector (a_i). The article vector represents the number of citable articles published by each journal in the last five years (as per Impact Factor with a different time period). To normalise this vector the number of citable articles published by each journal is divided by the total number of citable articles across all journals.

Lastly in order to account for dangling nodes (those journals which are not cited), some further manipulation is performed between the cross-citation matrix (H) and the article vector (a_i). Additionally a probability score is applied in order to model the likeliness of someone reading each journal after following a number of citations. This probability score will be outlined later when introducing PageRank (Section 3.6.2), the webometric which first realised the technique.

With the normalised cross-citation matrix and influence vector calculated, it is the dot-product of these two matrices which provides a journal Eigenfactor (EF) score as shown in Equation 3.4 (West et al. 2008).

$$EF = 100\frac{H.\pi^*}{\sum_i [H.\pi^*]_i} \tag{3.4}$$

Equation 3.4 also demonstrates that the final score is multiplied by 100 to give a percentage.

Weighting the effect of citations helps to flatten out the effect witnessed by Seglen (1997). It also brings into question whether a "good" journal is one which only publishes high ranking papers from already established authors, or one which also contains less established newer work. An example mapping of this pattern is shown in Figure 3.6. This figure, taken from Bergstrom et al. (2008), plots article influence score against total articles in the area of Neuroscience. Here the classic Power Law Distribution is found, suggesting that the "Annual Review of Neuroscience (which has the highest article influence score) is the best journal in this way. However, when adding data pertaining

to eigenfactor, represented in Figure 3.6 by the size of each circle, the "Journal of Neuroscience" comes out best.

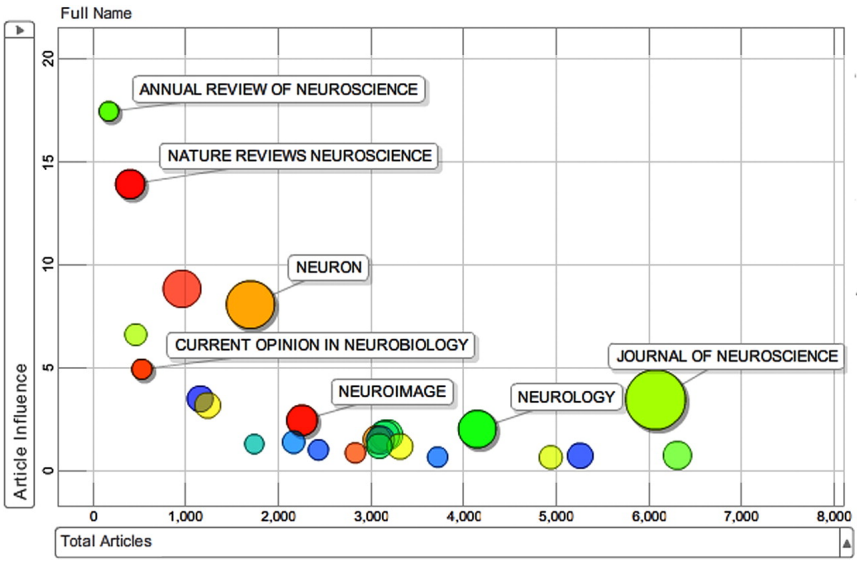

FIGURE 3.6: Article Influence Scores vs Total Articles vs Eigenfactor Score for 25 journals in the field of Neuroscience in 2007 (source Bergstrom et al. (2008))

Later, Section 3.5 looks at how factors such as the number of articles a journal has published influences its ranking compared to other metrics.

3.4.3 Article Influence Score

The Web of Science index allows the ordering of journals contained in its index by an Article Influence Score (AIS). While the name suggests some form of article based metric, Equation 3.5 shows this is not the case. This equation is in fact an extension to the Eigenfactor equation (Equation 3.4) and enables measurement of the relative importance of a journal on a per-article basis. Equation 3.5 (from West et al. 2008) shows how an AIS score is calculated by simply dividing the journals Eigenfactor score (EF_i) by the fraction of articles in the entire dataset published by that journal (vector a_i from Section 3.4.2).

$$AIS_i = \frac{EF_i}{a_i} \tag{3.5}$$

The fraction of all articles is normalised so that the sum of the articles from all journals is 1. Thus the mean Article Influence Score is 1.00 and anything scoring above that indicates a high impact journal and anything lower a below average impact. Article Influence Score has been included in order to draw attention to the fact that the influence is towards the journal, not towards the readers. The naming of this metric is misleading

as it could be interpreted as articles impact score in the wider community (granular level article rankings are not part of the metric family applied by Web of Science).

Section 3.5 compares this metric to others intended for ranking journals. Article Influence Score is just another mechanism that can be used to find which journal publishes the largest proportion of high impact publications.

3.4.4 h-index

Both Zipf and Bradford's Laws map directly onto the spread of citations in scholarly communications, which have been shown in this work to conform to the Power Law. The h-index takes the concept one step further and looks at the distribution of citations for a single author. Hirsch (2005), the creator of h-index, works on the hypothesis that influential author's are not those with lots of publications, which are all cited once or twice (thus in the tail of all publications), but those with a number of highly cited publications.

Hirsh realised that it is possible to calculate an author's influence by looking at the spread in number of citations among all the papers published by that author. In this way it is almost like taking the mean number of citations. The basic principle for this algorithm is as follows:

> *A scientist has index h if h of his/her Np papers have at least h citations each, and the other (Np - h) papers have fewer than h citations each.*

The h-index is designed to distinguish truly influential scientists from those who simply publish a lot of papers (or a few highly cited papers). Figure 3.7 shows the best representation of how h-index works.

Figure 3.7 shows 18 papers plotted which have accumulated a number of citations. They are plotted in descending order of number of citations and each point represents an individual paper. The idea is to draw a square on the graph ($h \times h$ in size) until the number of publications above the square is equal to the length of one of the sides of the square. In this example there are five publications above the square, which means that five publications obtain more than five citations each, thus the h-index for this person is five.

While this works well in many cases, it is still hard to differentiate between an author with a couple of very highly cited papers and one with a number of lower cited publications as demonstrated in Table 3.2. Here even though one author has 360 (or 800%) more citations than the other, the h-index's remain the same.

Hirsch (2005) suggests that h-index metric can be used as a guideline for awarding academic promotion, Nobel Prize winners and knighthoods. Moed (2009) puts forward

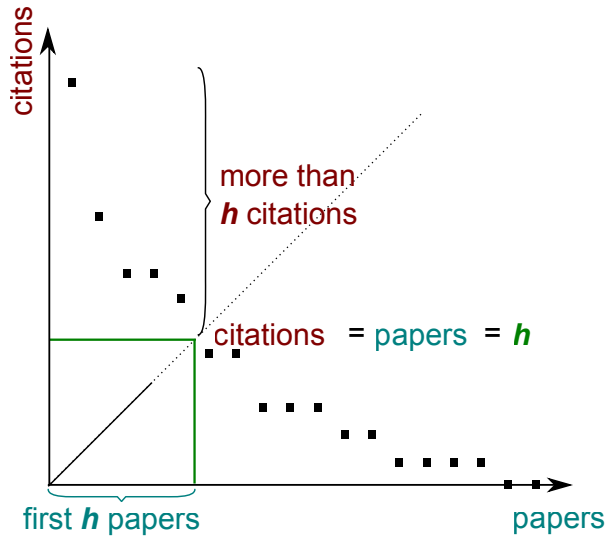

FIGURE 3.7: h-index for a set of papers with decreasing numbers of citations (courtesy of wikipedia user *Ael 2*)

Author 1		Author 2	
Paper	Citation Count	Paper	Citation Count
1	30	1	300
2	10	2	100
3	8	3	8
4	6	4	6
5	5	5	5
6	1	6	1
7	0	7	0
H = 5		H = 5	

TABLE 3.2: Two very different authors, same h-index

the case for h-index to only be used as an indicator, or combined with other data including average citation rate in each field of study. These combinatory factors should then help to avoid the situation from Table 3.2 becoming a factor.

3.5 Examining the Similarity of the Impact Factors using Web of Science data

This section looks in more depth at the Web of Science (WoS) Journal Citation Reports (JCRs) and the metrics used. Over a number of years the JCR reports have formed the basis of many studies into which journals institutions should invest. This investment is both monetary and involved, with institutions wanting to know what represents both

the critical mass of journals to purchase, and which should be publishing targets. For researchers who wish to improve their institutional standing, achieving a publication in a highly rated journal is a good step towards this aim, however as pointed out by Seglen (1997), publication in a highly ranked journal does not always lead to a high citation score and h-index.

This section carries forward the work started in Chapter 2, where the JCRs were used extensively to demonstrate the impact open access journals are having in overall scholarly publishing environment. By applying a number of different supported by the WoS index, this section looks at the influence these have on ranking the journals indexed by WoS. It is worth noting that WoS performs many quality control processes before it stores citation data, including the disambiguation of author and journal titles, as well as the removal of author self-citations.

In order to compare the different metrics the Spearman Rank Correlation Coefficient is going to be used. This coefficient takes the difference (d) between two ranked lists of the same items and calculates the correlation based upon the number of items (n). Thus to compare journal impact factor to number of citations, the ranked list for the two metrics is obtained, position differences for each individual journal calculated and then substituted into the Spearman Rank Correlation Coefficient equation (Equation 3.6).

$$p = 1 - \frac{6 \sum d_i^2}{n\left(n^2 - 1\right)} \tag{3.6}$$

The Spearman Rank Correlation result (p) will range from -1 to 1. Here -1 represents a perfect negative correlation, 1 a perfect correlation and 0 no correlation at all.

To show the similarities between the metrics used by WoS, a Spearman Rand Correlation is calculated between each combination of two metrics used. This will have the result of producing a correlation matrix, shown later in Table 3.4. In order to demonstrate this calculation from first principals, Table 3.3 shows the rank order of ten selected journals from the Web of Science index by both Citation Count and Impact Factor.

The ten journals shown in Table 3.3 represent the top ten journals by Citation Count selected from the thousands listed in the WoS index. The rank by impact has been translated from the raw position so that it lies between 1 and 10, for example the raw rank of publication 278424 is 131, which translates to a rate of 4. This process is required by Spearman as the size of both datasets being compared must be the same.

Along with the rank values, Table 3.3 shows the difference between the two ranks calculated by subtracting the citation count rank from the impact factor rank. Additionally we have squared this difference such that the sum total (88) and number of items (10) can be substituted back into Equation 3.6 as follows:

Publication ID,	Citation Count (rank)	Impact Factor (rank)	d_i	d_i^2
280836	1	2	-1	1
278424	2	4	-2	4
368075	3	3	0	0
219258	4	8	-4	16
27863	5	5	0	0
319007	6	6	0	0
10980121	7	10	-3	9
284793	8	1	7	49
36951	9	9	0	0
4637	10	7	3	9

TABLE 3.3: Example correlation data for Web of Science publications (Citation Count vs Impact Factor)

$$p = 1 - \frac{6 \times 88}{10\,(10^2 - 1)} \tag{3.7}$$

This evaluates to give a Spearman Rank Correlation Coefficient of $0.4\dot{6}\dot{6}$, which represents a positive correlation but not a very strong one. Table 3.4 shows a much stronger correlation is found when looking at the entire dataset rather than just 10 items. Additionally, Table 3.4 shows the full correlation matrix between each pair of algorithms when looking at the data taken from the Journal Citation Report for 2008. Due to the intrinsic link between all of the metrics (as outlined earlier in Section 3.4) many of the metrics do relate very closely to each other showing correlations greater than 0.7, including Citation Count and Impact Factor.

An exception to this positive can be observed when looking at the number of articles per journal. Here more articles does not lead to greater journal impact. The high correlation between articles and citations is most likely a result of lower impact journals not attracting enough interest to justify the need to expand and publish more articles. This relation would also hold for the comparison between Article Impact Score and Citation Count.

	Citations	Impact Factor	No. of Articles	Eigenfactor
Impact Factor	0.706			
No. of Articles	0.706	0.340		
Eigenfactor	0.683	0.780	0.284	
Article Impact Score	0.937	0.765	0.696	0.750

TABLE 3.4: Spearman Correlation of Metrics Applied by Web of Science

Looking back over the history of the Journal Citation Reports, year on year the top 100 journals contain publications such as Nature and the Physics review journals. The one surprise comes when ranking the journals by Impact Factor and find that the top

result is an Open Access journal. This journal — A Cancer Journal for Clinicians — provides 19 citable articles gaining a total of 7522 citations. By publishing a few very highly cited articles, means this journal fits well with Bradford's law into the group of journals (1) who are publishing the top 1/3 of papers.

3.6 Discovery and Ranking on the Web

From an early stage in the growth of the World Wide Web, it was apparent that managing and creating routes to access content was going to be a problem. Much like the problems which faced Garfield, it was hard to find a starting point from which reliable and authoritative content could be located.

In the early days of the Web, specific websites, early search engines and content aggregators (of which one example still exists online[6]) were relied upon as starting points to guide users around the very confusing and overwhelming amount of pages that existed. These services were often narrowly focused, and offered only links to a limited number of websites or companies who had paid to have their sites appear at the alongside search results or in prominent places on commonly visited Web pages. An easy way to make money, accompanied with people's reliance on such services, meant that search engines became big business; leading to high levels of competition between the various providers (Gandal 2001).

Early, and more successful, search engines would "crawl" (a term to represent the sequential processing of a number of websites) a number of websites known to them on a regular basis in order to "index" their content. Early methods of indexing included the popular Term Frequency and Inverse Document Frequency (IDF) among others. IDF is used to find documents which contain distinctive keywords compared to the whole corpus, so popular words like "the" are eliminated as they appear so often. With complex indexes built, search engines could begin to process the results and return them in some sort of order. Again this could be done using term frequency and compare the located frequency to the average rank the resource found (Salton 1987).

It was not until the late 1990s, when computer based technology began to spiral, that the use of "bots" and "spiders" to crawl the Web autonomously became more widespread. These spiders are essentially software services which index the Web by following links (in the form of hyperlinks) between pages and index whatever information they find.

Links on the Web and citations in publication can be viewed as very similar concepts, as they are both links to other available information. Simply counting these links in order to rank websites suffers from the fact that anyone can publish a link; there is no review process to go through when publishing a Web page. Another way to increase rank is to

[6] http://www.w3.org/History/19921103-hypertext/hypertext/DataSources/WWW/Servers.html

use common search terms out of context on a Web page, thus a false result appears in amongst the genuine articles[7].

While the indexing and discovery of Web pages is still performed in much the same way, methods to rank the returned results have evolved, thanks in part of bibliometrics. Rather than count every link as a vote of one, as is the case with publications, Pinski & Narin (1976) were the early pioneers of the weighted citation. Their basic premiss was that for a journal to be "influential" then, recursively, it must be cited by other influential journals.

Work by Kleinberg (1999) (Section 3.6.1) and Brin & Page (1998) (Section 3.6.2) reenforces the importance of weighting citations introduced by Pinski & Narin (1976). While Kleinberg (1999) extends the work by Pinski & Narin (1976) splitting publications into a set of Hubs and Authorities, Brin & Page (1998) abstract this one layer further in their PageRank algorithm and make everything a first class object, where each object is equal in position and meaning. PageRank is thus an example of an AIF algorithm as it does not differentiate between journals and publications, whilst still maintaining that to be high impact an article should be cited by many other high impact articles. In practice, PageRank is not strictly about pages on the Web, as it still performs aggregation around sites and keyword matching to provide likely pages from high ranking sites as the result of searches.

3.6.1 Hubs and Authorities

Hyperlink-Induced Topic Search (HITS), more commonly known as Hubs and Authorities, algorithm (Kleinberg 1999) attempts to rank pages on the Web through the use of a very simple principle:

> A good Hub is one that points to many good Authorities; a good Authority is one that is pointed to by many good Hubs.

Kleinberg (1999) makes the observation that the amount of relevant information available on the Web pertaining to a single topic is growing rapidly and in a way beyond the scope of human processing. By coming up with a way to distil a broad topic area, the HITS algorithm, they aid to collect together a list of "authoritative" sources on a topic.

A Hub is a site which contains a high number of out-links (links to other websites out of its domain), and a good Hub points to lots of good Authorities. An Authority is a page containing a lot of in-links from other sites, some of which are subsequently classified as Hubs. The resulting ranking is used by the search engine to order the results which are displayed back to the user, usually with the high ranking authorities first.

[7]Britney Spears' Guide to Semiconductor Physics - http://britneyspears.ac/lasers.htm

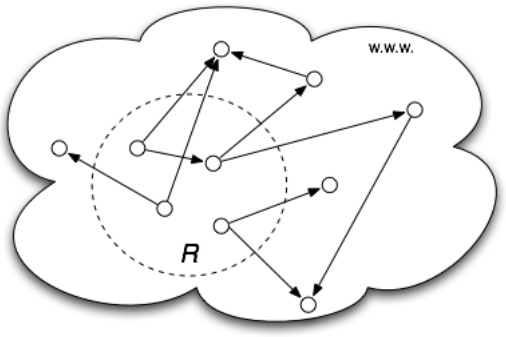

FIGURE 3.8: Search Results R showing out-links (arrows) between pages (circles) on the Web

The HITS algorithm is query dependant and Hub and Authority scores are calculated from a large corpus of data (websites) which are returned as the result of a search (R in Figure 3.8). Two algorithms, one to calculate the degree of authority of that page and one to calculate the hub-ness, are then run upon each page in the result set R. Equation 3.8 and 3.9 show the Authority score and Hub score algorithms respectively. Each of these are then used iteratively over the set of pages p in result set R. For each page p in R which links to a page q, a score a is given representing the pages authority (Equation 3.8) and a score h for the hub-ness (Equation 3.9).

$$a_p \leftarrow \sum_{q:(q,p)\epsilon E} h_q \tag{3.8}$$

$$h_p \leftarrow \sum_{q:(p,q)\epsilon E} a_q \tag{3.9}$$

At the end of each iteration the Authority and Hub scores are normalised such that their squares sum to 1. Scores with larger values are viewed as being "better" Authorities and Hubs respectively.

As adoption of this algorithm, and other similar algorithms based upon number in and out-degrees (Botafogo et al. 1992, Carrière & Kazman 1997), began to rise it became apparent that pollution of search results by creation of false positives would be an easy way to attract users to a website.

3.6.2 PageRank

PageRank (Brin et al. 1998) is based upon the links which exist between pages and sites on the Web, however it does not view all links as having equal weight. In PageRank, the score that a page receives, as the result of gaining a citation, is the rank of the

citing page divided by the number of other pages it also cites. Applying this technique improves accuracy of results in a manner similar to Hubs and Authorities, where a high ranking Hub which links to a small number of other sites may increase the rank of that site. Conversely low ranking "link farms" which simply link to a lot of sites will have very little affect on the rank.

The PageRank algorithm, shown in Equation 3.10, is applied iteratively over a citation network to rank each page. The basic principle is that the PageRank (PR) of every page is calculated from the PageRank of the pages which link to the one in question. Additionally this donated PageRank score is shared evenly between all the links (L) that this page contains. Iteration is required due to each pages' PageRank depending on the PageRank of all the pages which provide links. Before the algorithm is first run, the rank of each page is set to $1/|V|$ where $|V|$ is approximately the number of pages in the system. The final part of the equation ($\propto$) is necessary to model a random surfer who will not follow every link between Web pages.

$$PR(n) = \frac{1- \propto}{|V|} + \propto \sum_{p_j \in M(p_i)} \frac{PR\left(p_j\right)}{L\left(p_j\right)} \tag{3.10}$$

As the PageRank of each page depends on the PageRank of all the pages which provide links; to calculate an accurate PageRank for a single page may take several iterations over the Web graph, something which is huge in size. On the Web there will also exist many circular paths; these are paths which eventually end up back at the start, something which cannot occur in scholarly publications. These paths are created as content on the Web is dynamic and not static; a simple example would be a news site giving a link to a related article and that related article then being updated to reflect the fact that there is a related news article. So as not to follow too many links (and potential circular routes) when ranking pages, PageRank follows a link graph for a while before modelling what a "typical" Web user would do and skips to somewhere else on the Web and starts the process again. This is known as the random surfer principal.

The random surfer principle in PageRank dictates that PageRank, much like a real user, will not always follow a link on a Web page in order to traverse to the next page for processing. To model the random surfer behaviour PageRank uses a damping factor, represented in Equation 3.10 as $\propto$. This factor represents the probability of someone following a link, and from research performed by Brin & Page (1998), it is recommended to set this figure to 0.85 (representing a 15% chance of following any given link).

The damping factor can be justified in its usage by relating PageRank back to the basic bibliometric laws of Zipf and Bradford. By using these laws, it is possible to discover that the top few journals and publications will gain 80% of the citations, while there is a long tail of the rest. PageRank utilises this principle through the random surfer model by operating on the assumption that 15% of links will be enough of a proportional study

to easily find which websites are the highest rank and linked to by a lot of other high ranking websites. By applying PageRank in this way, iterations should be performed over the majority of high ranking sites rather than getting lost in the long tail.

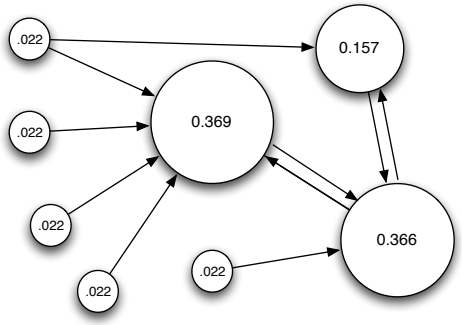

FIGURE 3.9: PageRank results (five iterations) over a simple closed network

Figure 3.9 shows a closed network (meaning this is the entire network) of eight pages and their PageRanks after five iterations. For the purposes of simplicity the random suffer damping factor has been removed from the PageRank algorithm in this example. Each node was given the starting value of 1/8 as there are eight nodes and the result after five iterations is shown by Figure 3.9. On such small networks, only a small number of iterations is required for the rank order to stabilise. By investigation, Brin et al. (1998) found that PageRank computation time was scalable in log_n, meaning that the number of required iterations remains feasible, even on huge datasets.

If PageRank is applied to publications and the damping factor maintained, the PageRank of a paper is generated from the PageRank of the papers p_j, which cite the original paper divided by the number of papers (L), which p_j cites. Many studies have already examined the potential use of PageRank within scholarly and citation networks with mixed correlation results depending on the specific test or subject area being considered. Chen et al. (2007) applied PageRank in order to assess the relative importance of all publications in the Physical Review family of journals, finding a positive correlation. When applied to the fields of Biochemistry and Molecular Biology, Ma et al. (2008) found PageRank to be very highly correlated with current impact measures used in these areas. Following on from Section 3.5, Dellavalle et al. (2007) suggests applying PageRank to journal ranking in order to introduce weighted citations.

PageRank is one of the metrics applied in this work to a real network of scholarly publications; results of which are compared to those of other new and existing metrics performing the same operation.

3.7 Bibliographic coupling and co-citations

Bibliographic coupling is the technique by which publications are related through reference lists, term occurrence and authorship relations. Bibliographic coupling via references is a technique first described by Kessler (1963) and exists when two publications both reference a common third publication or term.

A co-citation is very similar to bibliographic coupling but relates two publications via the reference list in another different publication. A co-citation occurs when two or more publications are cited by a single source publication. The relationship between the co-cited publications becomes stronger as they are co-cited together by many other publications.

Figure 3.10 develops Figure 3.2 from earlier and demonstrates how a network of co-cited publications is constructed. The cloud of four citing publications all cite at least two papers each, thus creating co-citation links between the originally cited publication and the co-cited papers. These are not the only three co-citations, as two of the directly citing papers also cite two of the other directly citing publications, thus there are five co-citation relationships established between the subject paper and the other publications; a citing paper can also be a co-cited paper.

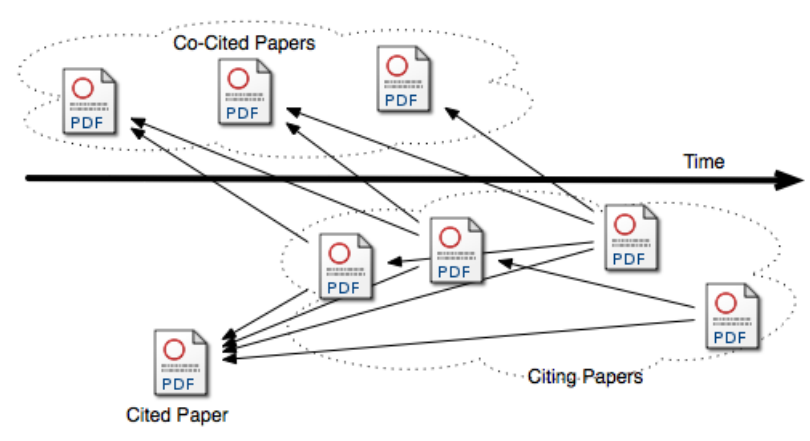

FIGURE 3.10: An example co-citation network

A citation network is a directed network, where links exist only in one direction between any two nodes. Additionally, this is a hierarchical network, as no older publication can cite a newer one. A co-citation network is an example of an un-directed network as the link between two publications exists in both directions - publication A is co-cited with publication B, thus, by definition, publication B is co-cited with publication A. Due to the fact that this relation can only be made by a third publication (C), the co-relations cannot be established by publications A or B. This observation means, that a co-citation network can still be used with algorithms usually designed solely for application on directed networks, as an equivalent network can be constructed.

Typically, a scholarly publication will cite a good number of other articles, thus each single citation will connect the subject paper to a number of co-cited neighbours. Logically this network of related publications also builds at a much greater rate than the citation network, in a shorter time period.

Also significant is the fact that a co-cited publication will always be older than the directly citing publication that establishes the link. Logically in order for a publication to be cited, it must be citable at the time of writing. However, it is anticipated that a number of the Co-Cited publications will be further through their publication life cycle and may have an established citation rate. It is these more established publications in addition to all co-cited publication which may offer a more stabilised rank score to any algorithm wishing to use this information as some form of indicator.

Co-citation analysis is traditionally used to relate two objects together by saying that they are linked in some way. As an example, if two publications are highly co-cited then it can be inferred that these publications are related and potentially core material to this subject area (Small 1973).

This thesis investigates the potential to use the co-citation as a means to ascertain the impact of an individual publication. Further it is hoped, due to the fact a co-citation network builds quicker, containing more data than the citation network, that this judgement of impact can also be made over a shorter time period than that required by existing methods based upon the citation network.

3.8 Conclusion

This chapter has outlined how the field of bibliometrics has developed, gathering statistical information from resources. In the early years these resources encompassed only written texts, and bibliometrics formed an essential scientific method by which authoritative sources of data could be found. This realisation by Garfield (Garfield 1955) that researchers follow citations, much like people follow links on the Web, shaped the future of bibliometrics research. As a result the area of bibliometrics is just as important now as it was back in the mid 1960s, especially when trying to find information in the ever expanding world of journals, both printed and published electronically.

The Web of Science index is proof alone of just how important bibliometrics are in ranking scholarly publications. In this chapter many of the techniques used to rank journals in the index were outlined and then analysed to show how similar these seemingly different sounding metrics are.

Methods of publication are evolving, already a number of digital only journals being listed in the WoS index, and these are not low impact. With more publications moving

to an online form, there is a clear opportunity again to expand the field of bibliometrics in order to discover new techniques to rank these publications and articles.

The latter sections of this chapter moved away from journal based metrics to look at individual article or resource based techniques. With nearly all the WoS metrics dependent on the existence of a journal it is fair to say that this model does not carry entirely into the online environment. With author self-archiving and institutional repositories now both receiving strong backing, ranking information found in these environments should be viewed as equally important.

Lastly, new sources for citation based metadata were briefly address, namely the co-citation. Here was observed the increased rate at which co-citation data is obtained and how this relates to more established publications. Co-citation data is already in widespread use to ascertain and deduce existing and new subject areas. The aim of this thesis is to examine the potential of using co-citation as a surrogate indicator of later impact.

The study of bibliometrics has provided some essential measures without which academic research would not have progressed as effectively. Thus the continual study could lead to even greater enhancements in these areas in the near future.

Chapter 4

Web Based Metrics and Early Indication Metrics

In scholarly communications, authors reference background and related work by means of a written citation. Bibliometric techniques analyse these citation networks in order to curate an impact score for each publication, the simplest of which is a Citation Count. Citation Count forms the basis of many analysis techniques designed to indicate different behaviours, the problem is understanding these factors. Many metrics utilise Citation Count in a manner which misleads many authors and readers in this way including the WoS Impact Factor (Garfield 2005). Impact Factor is measuring the readership base of Journals, indicating which are most read, not which contain the best material. Naturally, good scholars will want to publish their work as widely as possible, a statistic measurable in part via the Citation Count, and this means aiming to publish in high impact journals. Thus high impact journals will naturally attract high quality work which is thoroughly peer reviewed.

The Web provides an open platform for publication and dissemination on which links provide a similar concept to citations. Each resource and page is assigned a URL (Uniform Resource Locator) and it is these locations which are embedded into links on pages, to link from one website to another. It is these links which form the basis of Web based metrics (webometrics), which aim to help process the vast amount of information on the Web; much like bibliometrics is designed to help in the scholarly communications field (Thelwall 2008).

Web based publication tools such as blogs are changing the way in which people look at publishing on the Web (Blood 2004). Originally designed to hold a web-log of links a person found interesting, blogs are now a publication space where people exude their thoughts on topics, providing links to sites that backup or relate to the topic of the blog post. The importance of links has also been emphasised through the growth in

popularity of Wikipedia[1], which is now growing into a network of articles each similar to a professionally produced scholarly publication. This is partly due to verifiability policy[2] which dictates that statements made within each article have to be backed up with citations where possible. Historically the Web has lacked peer review, however with comments and editorial policies affecting blogs and sites like Wikipedia, the quality of content is naturally getting better.

The Wikipedia model is particularly interesting as anyone can be author or edit a page, thus many believe that the quality of the information would be mediocre. Chesney (2006) performed an empirical evaluation of Wikipedia. He found that by giving a series of articles to both experts and non-experts in the areas related to those articles, that the experts reviewed them as more credible than the non-experts did. Of the articles reviewed only 13% were found to contain errors. These results suggest that there is still a lack of confidence in Wikipedia as a source of information even though the articles are of high quality. It is also easy to contradict this confidence in Wikipedia as doing a search (on any Web based search engine) for any number of topics will tend to reveal the Wikipedia article among the top 10 or 20 results. This shows the number of citations (links) to Wikipedia which exist, as well as the ease at which these citations can be created.

With the huge amount of resources available on the Web, search engines play an essential role in the discovery of applicable content. It is these aggregators which have a real need to utilise infometric algorithms in order to rank results. As with traditional bibliometrics, the realisation that links are important has led to technologies being developed that help track and process these. Additionally users are able to track links and downloads of their own material, providing a surrogate measure of popularity.

Webometrics covers the study of techniques which can be used to analyse materials and resources available on the Web (Thelwall 2008). Figure 3.1 from the previous chapter has already demonstrated how webometrics take influences from other infometric methods. With the amount of content available on the Web, researching metrics able to process this amount of material was simply a logical progression. Unlike the field of scholarly communications, the Web provides a very granular platform for dissemination of knowledge. On the Web, individual resources can be given a download or hit score indicating more likely readership, something not possible with printed materials. As well as download statistics, webometrics covers areas such as link and citation analysis, search engine evaluation as well as purely descriptive studies of the Web. This chapter examines many established and emerging webometric techniques which can be applied to scholarly information disseminated via the Web. Use of such metrics is designed to indicate the subsequent citation impact of the same publication.

[1]Wikipedia - http://www.wikipedia.org/
[2]Wikipedia Verifiability Policy - http://en.wikipedia.org/wiki/Wikipedia:Verifiability

4.1 Download Statistics

Download statistics are the only type of metric presented here which does not rely on links existing to or from a page on the Web. Download statistics are simply a count of how many times a file hosted by a Web server has been downloaded. If a website does not contain files and all content is on the Web pages themselves, then page visitation (or hit count) is used in place of download count. Although both imply some user interaction with the content, neither are able to give very detailed analysis on the level of interaction such as how much the user read and for how long.

Download metrics are now a widely used mechanism for finding out how popular resources are on the Web. They see large scale use in the open source community, including within digital repository software, as well as in large companies such as Apple who use them to promote how popular their 'App Stores' are.

The advantage download count over citation is the drastic decrease in delay between initial publication and first usage data. Work by Brody & Harnad (2006) analyses the effectiveness of using download metrics as an early indication metric for publications, which are made available in an open access archive. They conclude that download counts can provide a good early indicator of subsequent impact, even on the basis that only 10-20% of scholarly communications were available to download online at the time of the study. Brody (2006) finds a 0.4 correlation between download metrics and eventual impact of articles by citation. Although not a perfect correlation, this early work demonstrates that download metrics can be used as an early indicator and have potential to be refined.

Like citation metrics, download metrics can also be manipulated for the authors own gain. Through repeated downloading, an author may be able to influence their own ranking. To achieve the same result with citation metrics is much more challenging. In this case, an author would need to publish an additional peer reviewed article that includes a self-citation to the previous work. To help deal with polluted metrics on the Web, programs such as awstats[3] are available which process each request in an attempt to remove false information. Many of these softwares also provide the user with much more information about the requests including:

- The Country where the request came from
- The agent or browser type which made the request
- Referer (the place the request originated)
- Search terms used to locate the site

[3] http://awstats.sourceforge.net/

- Duration of visit

The agent is one of the most important pieces of information, allowing the removal of non-human agents, including search engine "crawlers", which would otherwise heavily influence the results. Even after redaction of most of the irrelevant information done by the software itself, further manual processing may still required to remove local factors which may influence results, such as author self download.

4.2 Reader Pathway Metrics

Reader pathway metrics, a concept applied to scholarly communications by Bollen et al. (2005), examines the research process of a scholar. Specifically it is looking at the pathway taken by each scholar which results in the reading of one or more full text articles. When comparing this Reader Generated Network (RGN) with the Author Generated Network (AGN) of citations, a strong correlation was found. In order to better compare this type of metric with other types, Bollen constructs a four quadrant graph, shown by Figure 4.1, that illustrates the different factors which can be used to judge impact, each with a number of examples listed.

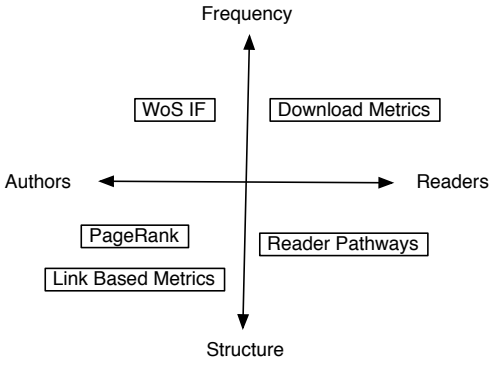

FIGURE 4.1: Metrics Types on the Web (from Bollen et al. 2005)

The four axis F (Frequency), R (Readers), S (Structure) and A (Authors) refer to actors and statistics which when combined in different ways, represent different metrics. Three of the quadrants are reasonably easy to explain starting with FA. Frequency-Author, of which the WoS Impact Factor is the perfect example, represent the mapping between authors and impact though the counting of citations (Section 3.3). SA (Structure-Author) contains all citation metrics which look at the links between papers including the PageRank algorithm (see Section 3.6.2). RF (Reader-Frequency) covers aspects including download statistics (Section 4.1) linking a reader to which publication they have downloaded and potentially read.

The last section, linking Readers to Structure (RS) is the hardest to obtain information about in a complete manner. RS information is that relating to the paths that readers

(rather than authors) take between materials; it is the study of which citations a reader actually follows, rather than what citations an author gives. Bollen et al. (2005) studied RS patterns through gathering and analysing a series of Web logs taken from several journal hosting services. Using the download statistics obtained from these sites a readers path through the publications could be calculated, resulting in a high correlation being observed between the number of downloads of publications which were cited by the previously downloaded publication. Bollen et al. (2005) concludes that download based Reader Generated Networks (RGNs) followed closely to that of citation based Author Generated Networks (AGNs) and thus in turn this network ties closely to that of the WoS IF. He found that people following a number of links (in this case citations) to find subsequent material, a practise common on the Web.

It makes sense in a peer reviewed network that the reader would trust the author of a paper to cite relevant high impact. Following this citation, they themselves may choose to read and perhaps cite the same paper in their own work. Logically, the consumer is likely to follow these links for a number of possible iterations. As a continuation of this work it would have been interesting to see if they could work out how many iterations are typical for a reader. From this a damping factor could be calculated and compared to that calculated by Brin & Page (1998) when looking at reader pathways on the Web and the PageRank algorithm.

4.3 Linkback

Linkback is the colloquial term for a notification system which alerts authors when someone links to their content on the Web. This is a practice similar to a journal report, which a publisher may email an author to inform them of downloads and citation count figures. Linkback is designed to be a live technology, able to inform authors of instant changes.

This section introduces three implementations of Linkback technology: Refback, Trackback and Pingback. Refback covers the technique of recoding the referrer of the request (as already covered in Section 4.1), requiring minimal technological support. Trackback and Pingback rely on support to be provided by both citing and cited servers, which while making them more complex than Refback this does have the advantage of making them even earlier impact indicators. Additionally each subsequent technique was designed with the intention of stopping users from generating artificial data for personal gain; a technique also referred to as gaming the system. Link pollution, the practice of linking only for the purpose of gaining standing, is an example of gaming the system attempted on the Web. In scholarly publications, gaming the system is harder due to time delay, but indexes including Web of Science still choose to remove author self citations.

The idea behind Linkback is to report who is linking to what content on the Web. Additionally, if the link is prestigious enough, many services will display a list of linking sites in order to inform readers of other possible related resources available on the Web. As with each metric however, once the algorithm and technique is known, pollution can become a real problem and Linkback is no exception. With false links being so easy to generate and report, techniques other than Refback have not become widely used despite their potential.

4.3.1 Refback

Refback uses the HTTP referrer header (Fielding et al. 1999) to inform the cited server where a client originated from, if anywhere. Each time a link is clicked a Web browser can choose to report the readers previous location and if the source and destination are both provided by the same server, then a reader pathway can be deduced. Refback only works in the circumstance where the reader has clicked a link and not pasted the address in from somewhere else.

Refback information informs the author of two things. Firstly that a link exists on the origin site (something unknown until it is clicked the first time) and that a reader chose this pathway to get to the desired content.

Refbacks can be forged easily as they form part of the HTTP header that is sent to the cited server as part of a HTTP GET request. Currently, there is no commonly used way to counter false Refback information that is able to ascertain how genuine the client and the link are.

4.3.2 Trackback

While Refback relies on a link being clicked before the author is informed of its creation, Trackback provides a technique for automated and instant notification. This is effectively the equivalent of a newly published paper resulting in the instant informing of all cited authors to the presence of a new citation. This form of notification could be used, along with other bibliometric techniques, to help inform impact and build timely social connections between scholars (Matthews et al. 2009).

Trackback works by the citing server sending a Trackback "ping" to the cited server which it then logs. The ping request takes the form of an HTTP POST request and like Refback these can be easily polluted by sending made up requests at a "cited" server. One of the best ways to avoid Trackback pollution is for the cited server to not look like it does anything with the ping data. The main aim of spammers is to get a return citation on a higher ranking website such that it may gain more hits itself. If the Trackbacks are

simply counted and not displayed publicly than there is no incentive for a spammer to send false Trackbacks.

Unlike static citations in scholarly literature, a link on a website can be removed, and it is the cited servers responsibility to manually check the continued presence of the inbound link.

4.3.3 Pingback

Pingback is the most complex of the three Linkback techniques and requires support from both the citing and cited server to process Pingback "pings" (Langridge & Hickson 2002). Pingback extends the principals behind Trackback in an attempt to prevent spamming.

Pingback uses a specific x-pingback HTTP header along with an XML-RPC (remote procedure call) script. When the cited server receives a Pingback "ping" it is required to process the XML, and navigate back to the citing server and process the content to ensure that the link to itself exists. Pingback enforces the requirement that the cited server must check that the link to it exists, it even provides the method by which this should be done. Of course the link may subsequently be removed, thus the citing server may wish to re-check it still exists.

Within the scholarly communications field, Pingback could be a very useful technique to implement on citation hub websites such as Citebase and Citeseer. Such websites perform the automated extraction of citations from scholarly literature. Pingback could be used to both inform the authors of a new citation and also allow the authors to ratify that citation, thus potentially providing a simple disambiguation technique.

4.4 Mention-It!

Mention-It![4] is a new idea introduced by Tim Donohue at the Open Repositories 2009 conference in Atlanta where it won first prize in the developer challenge event. Mention-It! is a simple program which searches across many providers on the Web to find mentions of a specified string, such as a paper title. By specifically searching only search engines which support RSS or ATOM exporting (thus designed to be harvested), Mention-It! is aimed at being a search tool for locating mentions of a chosen topic on social network sites such as personal blogs and Twitter.

As Mention-It! is searching for references to a specific item or topic, these have to exist, thus pollution of results is harder than with Refback and Trackback. The main problem

[4]Mention-It! - http://code.google.com/p/Mention-It/ (Jan 2011)

with Mention-It! is the sheer number of results which are likely to be returned for a single phrase. Conversely if Mention-It is searching for a specific URL, this essentially represents a Pingback search without the instant notification. However, when searching for the publication title or author name then there is a distinct chance of overlapping with other unrelated results.

By studying the application of these techniques on scholarly materials, it is possible to see potential usages in different scenarios. The remainder of this chapter introduces many of these and concludes on which are applicable for use as early indication metrics when dealing with scholarly communications.

4.5 Web Based Metrics and Digital Repositories

Using Web based metrics with Digital Repositories is beginning to gain more widespread adoption and research as to its relevance. With many research councils now mandating that their funded research must be made available openly in an online repository, a greater number of full texts are becoming available online (Brody & Harnad 2006). With this comes new ways in which research assessment scores can be gathered which fits better with the modern ways in which people are now using material produced as a result of research projects.

The School of Electronics and Computer Science at the University of Southampton which develops and uses the EPrints repository software, has been gathering and processing webometrics for some time through logging usage of their digital repository named ECS EPrints[5]. This section looks at how these metrics can be used in conjunction with existing software, before addressing some of the ways in which this data can be processed specifically for a digital repository.

4.5.1 Download Metrics

A number of modules are installed an enabled on the ECS EPrints repository including awstats (introduces in Section 4.1) and data has been collected for a number of years. This section looks solely at the data collected in 2008, including number of accesses, number of downloads, client types and referees. The output from the awstats software focuses on time based accesses and not page specific accesses, therefore only date ranges can be looked at, hence the reason to choose the year of 2008 as the sample range.

From awstats, in 2008 the repository handled over 600,000 unique visitors each visiting on average 1.39 times. On each visit a user would visit an average of nine pages and download about 1.5Mb of data. This much data transfer per user suggests a high number

[5]ECS EPrints Repository - http://eprints.ecs.soton.ac.uk

of users who visited actually downloaded an item available from the repository. Over the course of 2008 almost 1.2TB of downloads were made from the repository and this does not include search engine crawlers (which identified themselves as crawlers); that by themselves totalled nearly 650Gb in data transfer.

Figure 4.2 gives a breakdown of bandwidth used to transfer different types of files from the archive in 2008. From our total of 1.2TB for the year, about 40% (XML + HTML) of this consists of page views and exports to clients, which are not identified as crawlers or computer agents. These figures equate to 1.6 million page views and if this is aligned with the 1.4 million images downloaded, it would be logical to say that these were Web pages which also contained images for the purposes of style.

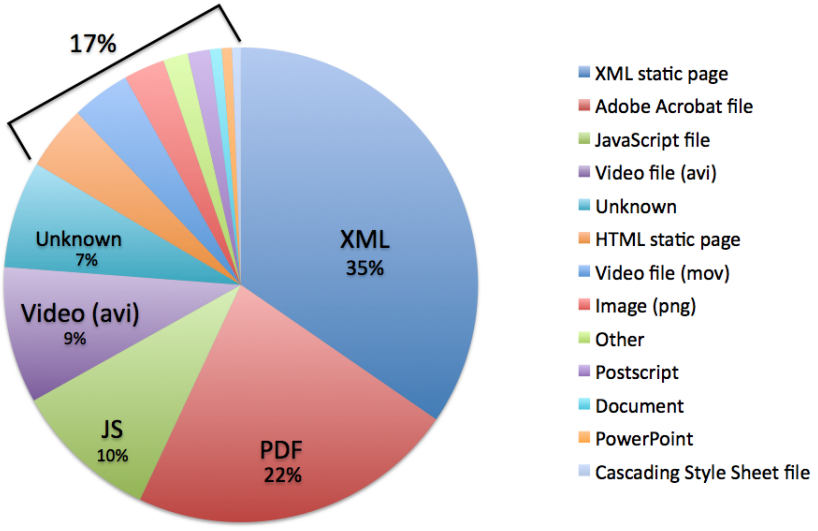

FIGURE 4.2: Downloads by Total Bandwidth (EPrints ECS in 2008)

Figure 4.2 shows the second largest amount of bandwidth was used in transferring of PDF documents. This equates to 275Gb in downloads (about 1/4 of the total for the year) with a total of just under 442,000 requests. From the 1.5 million Web page accesses, it can approximate that of these accesses, around 1/3 may have resulted in content being downloaded. However, due to the fact that content could be downloaded directly, it would be inappropriate to draw that conclusion without further data about where requests to content originated (Section 4.5.2).

In order to establish the amount of actual content that is being downloaded from the repository, it is necessary first to find out what types of content there are in the repository. By making use of the Repository of Open Access Repositories (ROAR)[6] service we can find out a repository's Preserv profile. The Preserv profile is the result of scanning all the hosted content within the repository and finding out the file types, shown here in Figure 4.3.

[6]Repository of Open Access Repositories (ROAR) - http://roar.eprints.org

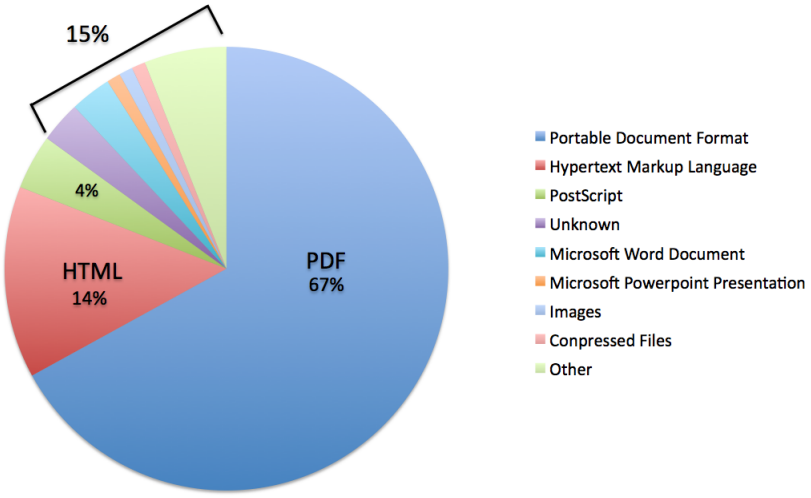

FIGURE 4.3: Preserv File Format Profile for EPrints ECS

By factoring out those formats not included in the Preserv profile, Figure 4.4 shows the number of each file type downloaded from the repository. This shows that PDF equates to 71% of the total downloads from the EPrints ECS Repository.

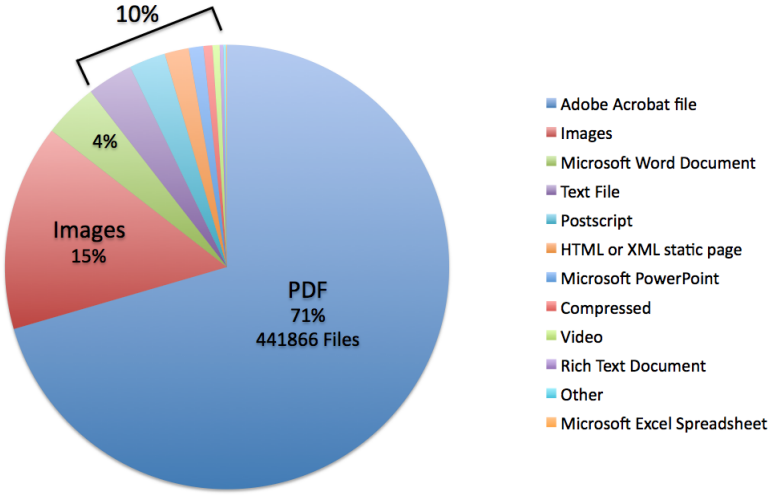

FIGURE 4.4: Downloads of Preserv Profile Formats (EPrints ECS 2008)

Interestingly the percentage of downloads which are PDF (from Figure 4.4) almost matches the percentage of PDFs which exist in the repository as stated by the Preserv profile (Figure 4.3).

4.5.2 EPrints and IRStats

With EPrints responsible for organising and disseminating the contents contained within it on the Web, there was a logical progression realised in constructing a statistics module to process the Web logs produced by awstats. As in Section 4.5.1, the level of processing

provided by awstats is very basic in nature and focuses on how many hits the website achieves rather than on downloads of the hosted content.

For the purposes of early indication metrics it is of greater relevance to find out how many people are downloading the full texts from the repository and where these users are coming from. IRStats[7] acts as a processor for the logs collected, analysing them in greater depth and with more context about the repository environment in which it is operating.

Figure 4.5 shows download metrics for a single publication which give an instant indication of the impact of this publication on the Web. By displaying these graphs on the abstract page which relates to the publication, a user can gain an impression on how authoritative this article is.

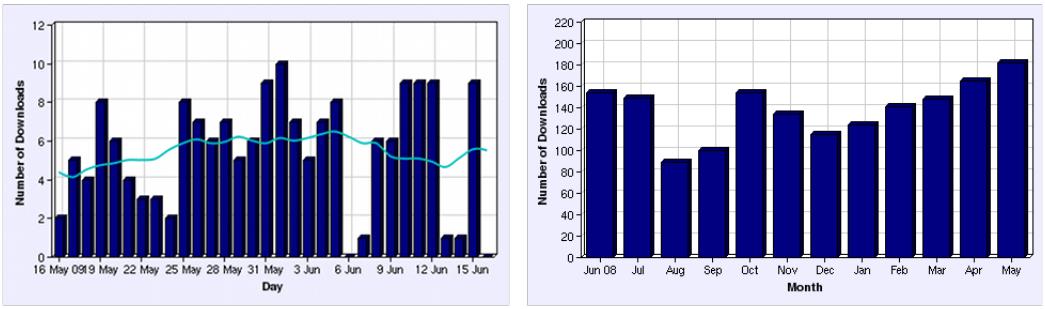

FIGURE 4.5: IRStats: Daily and monthly download graphs for a single publication

This work became a core part of the work by Brody & Harnad (2006). They proved that download metrics provide an effective early indication metric showing a 0.4 correlation with Citation Count.

With the ECS repository averaging 30,000 downloads a month, as tracked by IRStats, this represents 360,000 downloads in the year after false positives (bots and crawlers etc) are removed.

Figure 4.6 shows the global distribution of downloads demonstrating the widespread nature of the internet as a publication medium with a total of 58 countries represented (30% of the countries in the world) in 2008.

4.5.3 Evaluating Refback on EPrints

From Figure 4.2, the possible conclusion was drawn that about 1/3 of the visitors to the ECS EPrints repository downloaded a full text file such as a PDF document. 1.5 million Web page hits suggests that a large number of the users may be downloading an article via clicking a download link contained on another page in the repository. If this

[7]IRStats - http://wiki.eprints.org/w/IRStats

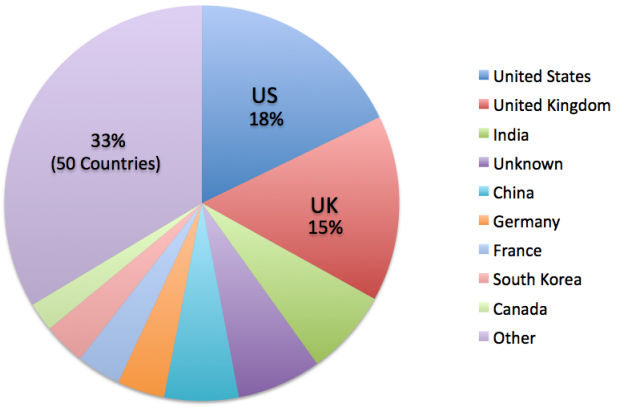

FIGURE 4.6: IRStats: Countries downloading from EPrints ECS in 2008

is the case, then they may also be noticing graphs such as that displayed in Figure 4.5 which, in turn may be influencing whether or not they download the article.

In order to analyse a users behaviour it is necessary to look at the Reader Pathway (RP) metrics which are provided via Refback. There are two possible ways to analyse RP metrics on any website. One looks at only the direct referrer to the eventual download of a full text, the other tracks the readers pathway back to the initial entry point to the repository. IRStats performs the second type of RP metric looking specifically at the users session from the point when they entered the repository environment.

From awstats it is only possible to find the direct referrers used each time an access is made to a part of the repository (Figure 4.7). this figure demonstrates that the majority of referrers are apparently from within the repository, represented on the left of Figure 4.7 by Direct / Bookmarks, which would match the earlier hypothesis relating to number of page views and downloads.

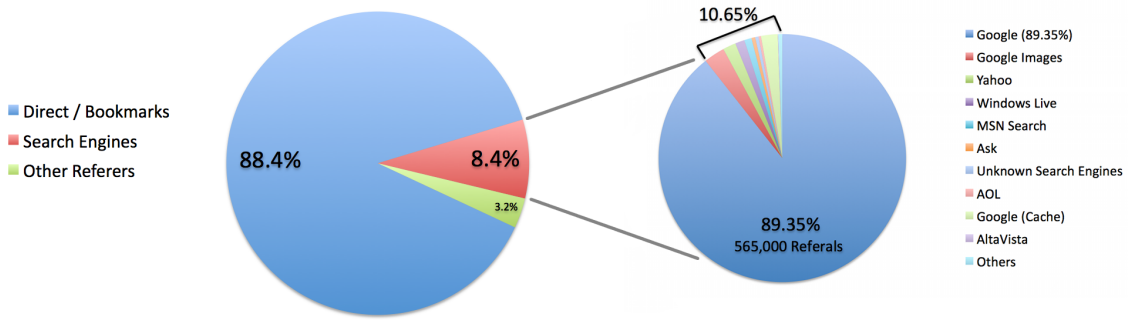

FIGURE 4.7: Repository Referrers for EPrints ECS in 2008

Taking the data from awstats reveals a total of over 7 million referrers which does not match the 1.5 million page hits from earlier. Additionally without session data, it is not possible to conclude which of these referrers resulted in a full text download.

By limiting the data, a process performed by IRStats, it is possible to list only the session referrers that resulted in one or more full text downloads. Figure 4.8 shows both the direct referrers and the session referrers which satisfied this condition. Clearly dominant in both cases is the search engine Google, even when 25% of referrers unknown (mainly due to browser compatibility and privacy extensions). Wikipedia provides a nice 3% of external references, meaning that some of the publications are cited on wikipedia as sources of information. Internal pages to the University (Abstract Pages, EPrints ECS and Other ECS) account for around 8% of the full text downloads with other external websites making up about 13%.

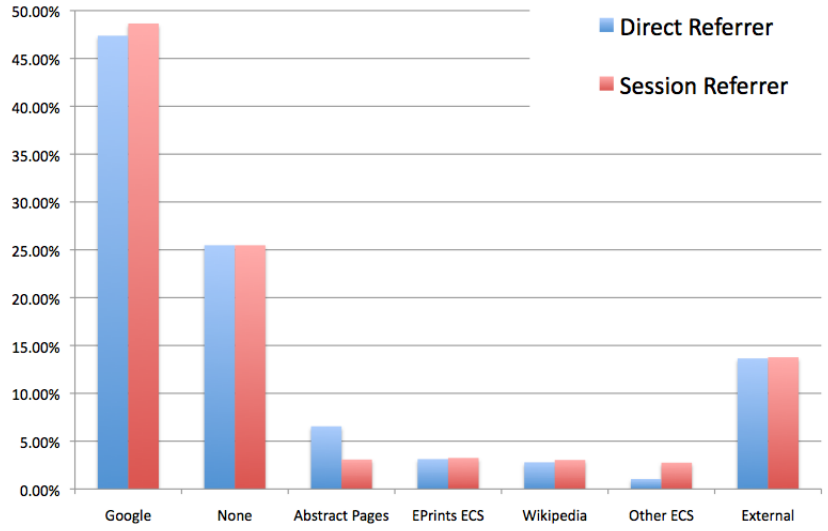

FIGURE 4.8: IRStats: Direct referrers vs Session referres

The only change of note between the two sets of results (direct and session referrer) is the number of people who download from an abstract page within the repository is reduced once the origin of their session is found. In the majority of these cases the persons session started at an external search engine.

With the combined school and repository Web presence only providing 8% of the total references to the eventual downloads of fulltexts from the repository, it is clear to see just how important search engines are in discovering content on the Web. This finding re-enforces the importance of revealing the data and metadata relating to a publication for services like a search engine to index. Conversely this also suggests that spending more time on making a repository website look good than revealing the data, is not advisable, the majority of users will never see the website according to this result.

4.5.4 Evaluating Mention-It! on EPrints

Mention-It! searches for "mentions" of an item on the Web. Specifically the idea is to track early impact by finding out how much a publication is being talked about in the community on social websites, blogs and services such as Twitter.

Many challenges exist for Mention-It!, the greatest of which is what to search for? With most of the Web lacking in context, i.e. a computer cannot tell for sure what someone is talking about without URI or URL references. Unless just searching for these links, it is hard to know if search results are relevant. This is particularly the case with publication titles, where some can be short and use common phrases, while others are relatively unique due to length and structure.

The Mention-It! software is designed to use a number of search services which are able to export their results in a processable form including Google, Wordpress and Technorati blog searches, as well as the social services Twitter and Friendfeed.

In order to do a quick evaluation of the usefulness of such techniques it was decided to set up a Mention-It! search for the ECS EPrints repository, this search looked for direct links and title mentions to the publications contained within repository. The search was not limited to any particular publications thus mentions to any in the repository were sought after. Figure 4.9 shows the number of mentions to publications, split down by month of publication, including searches for both paper titles and URL searches. Although this is only a quick glance at this service, a significant result would be a visible trend in age of publications which are being mentioned highly.

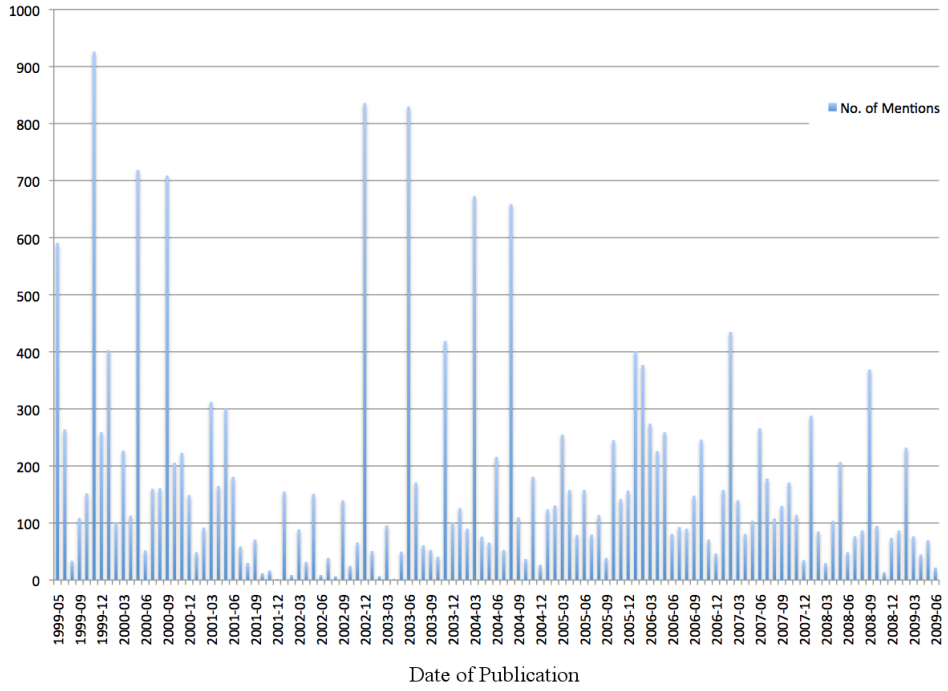

FIGURE 4.9: Mention-It!: EPrints ECS - Titles and URL mentions

While no consistent trend is demonstrated in Figure 4.9, there are a few peaks in the graph which are larger the older the publication gets. This suggests that the repository contains a number of established high impact articles which are producing many links. Further examination found this not to be the case, although there are some publications

which account for a significant amount of mentions these, as predicted, were the ones which use common phrases for titles.

In order to obtain a more accurate impression of mentions a second search was done which was limited to URL mentions only; meaning any results have to contain absolute links to the publication in question. Doing this search returned only 238 mentions for over 34,000 publications, which did not constitute a significant enough result to process it any further. This did serve as a good experiment to examine the problems with such services and demonstrate how the level of interconnectivity on the internet affects what people browse for. It also highlights the problems with simple searches compared to those which process results in depth, this is particularly the case in circumstances where people use link shortening services, when sharing links via short messaging services like Twitter.

Lastly, social networking data, including status updates, are not persisted very well on the Web and thus any search of these services can only return the most recent set of results. Any service wishing to process these results will have to perform a regular search and cache the results for next time.

It is clear then that although using a concept like Mention-It! seems like a good idea, the usage needs to be controlled and kept within the bounds of a limited context. For this reason it was decided to perform a similar experiment during a live event, in this case the 2010 Open Repositories Conference (or10)[8] was chosen. The or10 conference has a strong twitter presence among the delegates, allowing the feed of messages being tagged with the conference hashtag (#or10) to be logged and processed. With the conference program input into a database, along with the tweets, processing was done to attempt to match the tweet to a presentation of a publication. Due to knowing when a publication was being presented, along with the presenter name, links and title of presentation, some simple text based processing could be done to classify a tweet as belonging to that presentation.

The aim of the study was to identify the most popular presentation as well as look at which of the parallel track was the most popular. In addition, due to the limited context, it was also possible to examine who were the most prolific tweeters and identify which sessions these tweeters went to. The output graph from this study is shown in Figure 4.10 and significantly, in just two days, over 834 tweets were collected that corresponded to a particular publication.

Figure 4.10 represents a more realistic current use for Mention-It! where more context is known about both the publications and the specific search required. Significantly, the results of the search, from the conference where the publications are being presented, are only likely to be about the publications in question. Further to producing the graphs,

[8]Open Repositories Conference 2010 - http://or2010.fecyt.es/

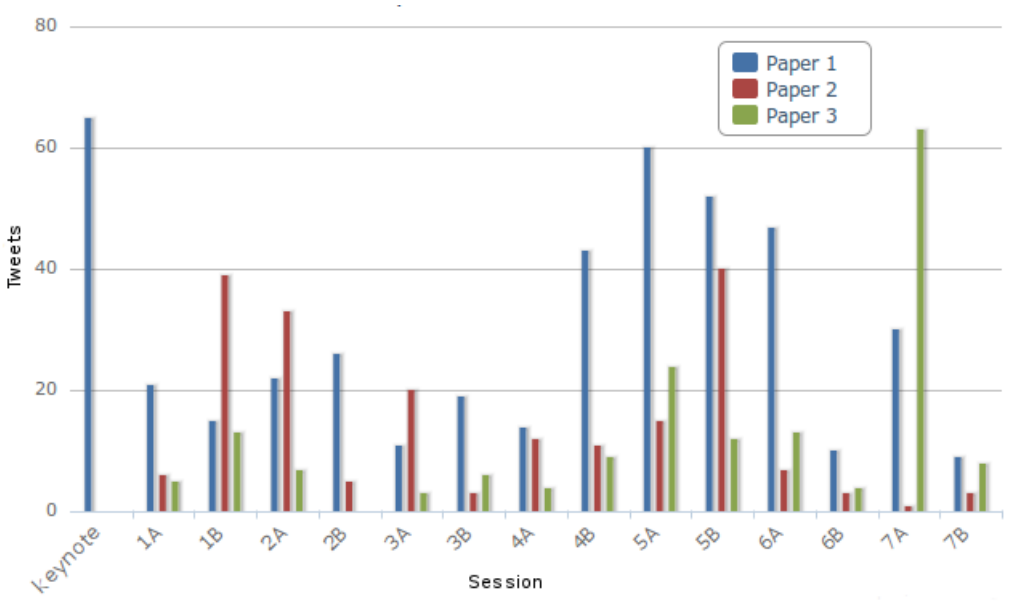

FIGURE 4.10: Mention-It! Live: Twitter study of a conference

software was also written[9] to archive the matched tweets in perpetuity alongside the repository record for that publication. Thus the author and subsequent readers can read the comments made during the presentation itself.

4.6 Conclusion

In an exponentially expanding environment of information, such as the Web, the need to be able to filter data to find resources which bear relevance to searches is critical. Early techniques employed on the Web included pay-ads and basic citation counting mechanisms similar to that applied in the area of scholarly communications and bibliometrics.

Brin & Page (1998) realised a problem with link pollution on the Web and devised PageRank, a metric which applies a measure of prestige to links, as the solution. To this day, Google and the PageRank algorithm remain the most popular search engine on the Web. While PageRank demonstrates a technique for effectively filtering Web pages or publications that also takes into account and eliminates false positives, it is not a very good early indication metric. This is due to a network of citation links having to be constructed prior to PageRank having any affect, something looked at in more depth in Chapter 5.

Download counts have proved effective as early indication metrics and show good correlations against later Citation Count (Brody 2006). There are a number of problems with download counts though, not only can they get polluted, but the statistics can also become distributed if the resource being downloaded is hosted by many providers.

[9] EPrints Twitter Harvester, alpha 1 - http://files.eprints.org/563/

Technologies including as Linkback and Refback allow collection of statistics relating to usage of materials, with potential use as an early indication metrics in the field of scholarly communications (Matthews et al. 2009). These techniques allow users to be automatically informed when links are created between sites, potentially creating new social connections within the community.

A shift in social publishing is meaning that people are now publishing links on sites such as Blogs and Twitter feeds. There is an opportunity here to employ technologies such as Linkback in this new environment to analyse these new networks. However this work is in its infancy compared to many of the other techniques. Mention-It! attempts to eliminate the need for people to link to actual URLs by looking for mentions of a publication by title, however this technique suffers from not being contextually aware. Distinguishing between authors has already been shown to be a problem and until more context aware linking is used, there is very little information you can take seriously from a Mention-It! search. By applying the principals behind Mention-It! to a small, tightly controlled event, a strong potential use for the technology was found. By looking for trends in what people are mentioning, such techniques could be used to predict which publications from a conference are subsequently going to become highly cited.

Early indication metrics are important to help attempt to judge later impact and can act, in some cases, as an indicator as to whether work in certain areas should continue. The example uses of Trackback, IRStats and Mention-IT! outlined in this chapter all work best as early indication metrics if focused on a single service, such as a digital repository, an event, or small field of study. In these controlled networks where context is well defined, there is a clear opportunity to expand beyond citation and download metrics into social metrics, as well combining the well controlled publication networks with social networks, for the purposes of discovering hot topics and potential high ranking future publications.

Over a longer duration, there is still no replacement as yet for the respect gained through a citation in a journal publication. The significant difference is the amount of effort required in publishing a paper containing this citation. It is the aim of this work to take a look at authoritative sources of information, such as citations, and examine the possibility of establishing an early indication metric based upon this controlled network of information. This requires combining many of the techniques used on the Web, with the network of citations and co-citations established in the scholarly communications network.

Chapter 5

Co-Citation Metrics

Bibliometric and Infometric techniques all rely on the corpus of data, both directly and in-directly, resulting from the publication of a number of resources. As the quantity of resources being published on many platforms grows, so does the need for techniques to automatically process these in order to provide some sort of guide as to which materials are deemed important. Like many areas of research, bibliometrics can benefit from the incredibly rapid development of information sharing on the world wide Web. Before the Web, companies including Thomson were heavily relied upon (as they still are) to carefully collate together information indexes from which information pertaining to journal impact could be calculated. With the Web now transforming into a "web of data", these indexes are becoming easier both to gather information from, as well as disseminate from. These aspects, along with the enhanced abilities available to process this data, allow the generation of entirely new networks of data.

This chapter looks at new ways to process this data with the aim of examining whether it can be used in publication ranking. More importantly if it can be used for ranking, does any new technique provide any benefit over those which already exist. Chapter 3 looked at the key role that ranking plays in the modern day environment both in the area of scholarly communications as well as on the world wide Web. Especially important was the key observation that in an open publication environment, such as the Web, traditional ranking techniques, such as the Citation Count, have the potential to become easily polluted. This is likely to reduce their usefulness compared to algorithms which take this into account.

Chapter 2 outlined how scholarly communication has changed, showing how scholars are now eager to share their work openly online in the hope of achieving a greater and more widespread impact. While some see self archiving and online dissemination as an assured way to be cited (Harnad 2006*b*), there is still a need to consider carefully where to publish in order to gain highest visibility. The most important stage of the scholarly life cycle, for an author, is that of gaining citations; the respect and trust of the

community for their work. In order to gain citations an author should attempt to get their work as well known as possible, typically by attempting to publish it in a medium which has a high readership.

Today, publication in a journal which is indexed by services like Web of Science is the key route to get an article noticed by the research community and parent institution to the author. To some, it seems strange that an index designed to help librarians with the serials crisis (Section 2.1), is now used to rate the authors themselves (Seglen 1997, Hecht et al. 1998). Even Garfield himself is wary of the problems with the impact factor metric and he addresses these in the same manner as Hecht et al. (1998). Garfield suggests that people misunderstand the term "Impact Factor" and acknowledges that people use the Journal Impact Factor as a measure of an authors' prestige. He stresses that an authors publication and citation count should be used as a primary indicator, with Impact Factor only used as a surrogate indicator for the most recently published articles, with no guarantees offered (Garfield 2003). On the issue of whether a journal Impact Factor should be used as an indicator for impact of an article in that journal, Garfield is unclear.

Looking in more depth at the Journal Citation Reports can raise a few questions about how a high impact journal becomes high impact; was this due to citations alone or was it more than that? Also how does a new journal become high impact? Once a journal is high impact it is logical to deduce that increased readership will in turn increase interest in publishing in this journal. Chapter 2 addressed many of the aspects which go some way to answering these questions and new journals are likely to emerge which focus on a new area of research.

Seglen (1997) makes the contentious point that the peer review process that controls the quality of publications is a lottery. If the absolute expert is not the one who reviews a publication submitted to a conference or journal, they may take into account secondary criteria, such as reputation of authors, institutions, journal prestige. Potentially a high impact author could submit a paper containing complete gibberish and it would still get published in a highly cited journal.

Moed, one of the worlds leading bibliometricians, has many issues with the way the impact factor is now used. In Moed & Van Leeuwen (1995), the main observation is that a journal impact factor is used as a surrogate for impact of an article published within it. Along with Seglen (1997), who concurs, they observe that it is the articles Citation Count which adds up to the journals impact and because of this the same cannot translate in the opposite direction. Moed & Van Leeuwen (1995) also point out that many journals publish what the WoS index terms "non-citable" documents, which are removed from the total article count when calculating Impact Factor. The problem is that the citation towards these non-citable documents are not removed from the same calculation. This means there now exist more citations towards less articles, rather than

a normalised count for both. As outlined earlier, Seglen (1997) finds that 50% of articles contained in a Journal account for 90% of the total citations, however it was not found if this 50% of articles includes those which are non-citable.

Using the Impact Factor of a journal as a surrogate measure of later impact is also something considered by Levitt & Thelwall (2011), who look at combining this with the current Citation Count. By combining measures in this way and applying carefully considered weightings to the score each factor contributes, Levitt looks for a technique capable of predicting later impact of individual articles. Although not directly addressed, looking at the early citation patterns could potentially reveal if an article is one of those in the top 50% of cited articles as detailed by Seglen (1997). There is a risk however that an article may take some time to establish itself and become one of these articles (Levitt & Thelwall 2008*b*).

For an author, in order for their work to be cited and/or published it needs to be discoverable, thus the place of publication is important. If Impact Factor is regarded highly at an institution as a measure of performance, then an author may have to limit their possible audience in order to keep their job. In an ideal world, publishing in a high impact journal would provide the best route to obtain promotion and a high number of citations. However there is no guarantee the amount of effort involved in this process will lead to a particular paper becoming one of the most highly cited 50%. Conversely, publishing somewhere where the article is more likely to be read and cited might not provide a high journal Impact Factor to count towards future employment.

Green and Gold Open Access publishing techniques (Harnad et al. 2008) provide a potential solution to the accessibility issues introduced by expensive journal subscriptions, whilst maintaining a potentially high Impact Factor. Both provide a means for the author to provide free access to their works, usually via the Web. At this point the article is now available on the Web alongside other open access articles and indexes of approximately 2.5million other journal articles which get published each year. Discoverability now becomes the major issue and webometrics are important as a method to help people locate relevant information.

In order to discover information, people typically refer to hubs of information or search engines as a starting point, including high impact journals, library catalogues and services like Google. On the Web, search engines are clearly the most important target in which information needs to be listed in order to be discoverable (as proven by Figure 4.8). From this hub(or search engine) users may follow a number of citation links, creating a pathway to their intended content. Kleinberg (1999) introduces the notion of Hubs and Authorities on the Web as a means to rank material, where a good Hub links to a number of good Authorities and likewise a good Authority is linked to by a number of good Hubs. If Google is considered as a type of Hubs, the Authorities to which they link must remain relevant as a large part of their business model relies upon it.

Webometrics (Thelwall et al. 2005) hold an important role within search engines in processing websites, in order to establish which are good authorities for information. The resultant ranking is then combined with text matching techniques against the users query to return high ranking relevant results. Chapter 3 introduced some of the techniques used on the Web to index and rank Web pages. The basic idea is to make a presence discoverable, by describing it well in an easy to read form (both by humans and computers). The last stage of the process it to hope that others link to the content produced; a concept which maps directly to citations in scholarly publishing, where the more a publication is cited, the more discoverable it becomes. On the Web, links represent citations between works and metrics designed to process these can be used to establish a websites rank score.

While metrics such as Impact Factor and Citation Count can, and should, be used to help guide choices online, these metrics have been designed for a uni-directional graph model where no cycles exist between nodes. Due to the temporal and static nature of scholarly communications, a previously published article cannot link to a newer one, something not true on the Web where sites change constantly. Without peer-review, the Web is an open platform on which to publish links, increasing the threat of false linkages intended for the purpose of increasing rank. For these reasons techniques such as PageRank apply weighting factors to each link and website that significantly decrease the influence of these false link hubs.

PageRank (Brin et al. 1998) has already been suggested by some as a good alternative algorithm for ranking publications (Chen et al. 2007) and this is due to the particular characteristics of the algorithm. While Chen et al. (2007) believes this to be a positive effect, Sidiropoulos & Manolopoulos (2006) still view it as having downsides in bibliometrics. However, both realise the potential usage of alternative metrics when considering different data sources available in disparate fields of study.

This chapter looks at the characteristics of many different ranking algorithms, generalising these as much as possible in order to allow thought about different use cases. A theoretical network of publications and citations between these is introduced, to which a number of different metrics are applied and results analysed. With the Web providing access to a greater amount of richer data and computing power enabling this to be processed in a timely fashion, extending beyond the simple citation network in order to find new sources of ranking data becomes possible. This chapter introduces the co-citation network as one such source of information and examines the implications and theory of ranking based upon this. This leads to the introduction of the CoRank algorithm which takes influences from both citation metrics and Web based metrics with the aim of providing some benefit over both.

With the key concepts of what defines a successful outcome, this chapter looks at the theory behind the CoRank algorithm and how it developed from PageRank. Some

testing is performed on a theoretical network of publications before moving on to look at how CoRank performs in a live environment. During the theoretical tests the techniques used to compare the various metrics are introduced; this allows the construction of the testing plan which then outlines the structure for the remainder of the chapter. Firstly then, it is necessary to address what criteria will be used when comparing metrics.

5.1 What constitutes a "Better" metric?

In order to compare one metric to another, it is first necessary to define the conditions of testing such that one can judged to be "better". Additionally these conditions should take account of any environmental factors. Garfield realised that by looking at the way people build a social scholarly network, he could apply a quantitative measure to the quality of publications in this network, to help new people in each field quickly judge the social standing of each available resource (Garfield 1955).

The 'needle in a haystack' problem (that of finding one item of valuable material among millions) gets exponentially worse on the Web, where the number of resources available is so huge that finding a set of resources in a subject area is difficult enough even before attempting to rank them in some sort of order of relevancy. Chapter 3 outlined some of the solutions to this problem that have become the accepted ways by which resources are filtered and returned on the Web.

Both Citation Count and PageRank have become widely used and valuable in today's modern society, used by services including Web of Science and Google. This thesis looks at other techniques which could potentially add benefit to these existing indicators. In order to add benefit, a number of desirable properties of any new metric have to be outlined, each of these should also be measurable in some form against current metrics. The following points in this section address some of the conditions which define an effective result of a publication search.

Finding the most significant publications

While Citation Count is able to help indicate how popular any given publication is, it does not help indicate the significance of this publication. Citation Count provides a good indicator, which can be used as part of any significance assessment, however the time taken in establishing a stable Citation Count score can delay this judgement. Any new metric should allow substitution with Citation Count for use in significance assessments, potentially with the added benefit of allowing these assessments to be carried out earlier in a publication or author life cycle.

To measure this condition, a correlation calculation is going to be used to examine the similarity between Citation Count (the currently accepted metric) and any new metric. In addition, this correlation is going to be examined over a number of periods in time

to help discover how quickly each new metric becomes correlated. A highly correlated metric will be capable of revealing the most significant publications, as defined by current measures.

Finding the core corpus of information

Here the Citation Count metric, or the journal rankings, could be used to find the core set of articles and journals which publish popular articles in a certain area. Like Citation Count, on which journal ranking relies, it is understood that the publication conditions are well controlled. The journal is aiming to only be publishing high impact articles which in turn maintains the journal impact.

In addition to the correlation test, publication rank information will reveal if highly ranked publications, remain highly ranked after an initial time period. Any core corpus of information should remain highly ranked over long periods of time. There is a potential for early indication metrics to fail on this test due to initial activity (such as social hype and early downloads).

Finding the most recent, significant articles

By specifying a boundary condition relating to time, the delay incurred for a publication reaching peak citation rate means Citation Count is unsuitable to use for finding the most recent, significant articles. For scholars, having early access to scientific findings is key to not repeating experiments or spending unnecessary time on research. Currently conferences, project meetings and word of mouth are the main mains in which such early research can be discovered in a reliable fashion. A good early indication metric should aim for the same level of reliability and capacity to identify more recent, significant publications.

This is perhaps the hardest requirement to fulfil while also satisfying the two previous conditions. To measure success against this criteria, the same correlation and rank data can be measured over time and not just at a static point. Measurement over time can be achieved through having a target to aim for, in this case the target will be set by the Citation Count metric. If any new metric demonstrates a high correlation or similarity in rank order to the target at an earlier point in time, it will be satisfying the condition of finding these significant articles sooner.

If any metric satisfies this last condition, logically it would also rank higher an amount of more recent publications. This condition can be tested by recording the average age of the top $n\%$ of publications in each ranked set, with the ideal result seeing a small number of more recent publications revealed. Conversely, if a new metric is ranking the majority of new publications highly, then it is unlikely to be satisfying the previous conditions of being a better metric.

5.2 Building an Artificial Publication Network

In order to develop and test both new and existing metrics a small artificial network was constructed against which their basic properties can be analysed. Using this network, it is then possible to visualise both the citation and co-citation relationship on which each metric is based including CoRank, the proposed new metric. The artificial network has been created to represent a number of publications at different points in their life cycles and has been created in such a way that a clear rank order by citation count does exist. Additionally, the network has been carefully constructed to mirror a plausible citation network, while also creating a good basis on which to demonstrate the key features of a co-citation network over a plain citation network. By augmenting this artificial network with additional figures and explanations this allows the clear explanation of the algorithms analysed in this section. With a set of metrics outlined, including the CoRank algorithm, these will be applied to the artificial network in order to examine relative performance and access the initial suitability of the CoRank algorithm.

After each of the metrics has been applied, this section looks at techniques which can be used to examine each to find any "better" metric. Establishing the age of the articles in the artificial network allows the older, highly cited articles to be distinguished from the most recent, potentially significant articles. Using this information, combined with the rank order obtained through application of each metric, enables the construction of a series of tests, which can be used to demonstrate that CoRank exhibits some benefit over existing metrics.

Figure 5.1 shows the carefully constructed artificial network consisting of 18 resources which are all interlinked in some way. Although the resources and links could represent different types of directed network (e.g. a decision tree), in this work all the nodes represent publications and the links between them are citations.

Figure 5.1 shows an artificial network contains a total of 18 nodes, each of which has been assigned a node number, shown alongside each node. These numbers become particularly important when analysing the network with regards to obtaining a ranked order for the nodes by different metrics. Further to the 18 nodes there exists 32 links between them which represent citations.

With the artificial network outlined and the citation links established all the information required to apply each metric has been gathered. The result of applying each metric will be a ranked list of publications, each of which can be compared to the resultant rank orders produced by the other algorithms, in order to find the similarities and differences. By looking at the correlation between these sets of results, the first two important aspects defining a good publication metric can be analysed. These aspects include the ability to be able to find the core and high respected publications from a large corpus of information.

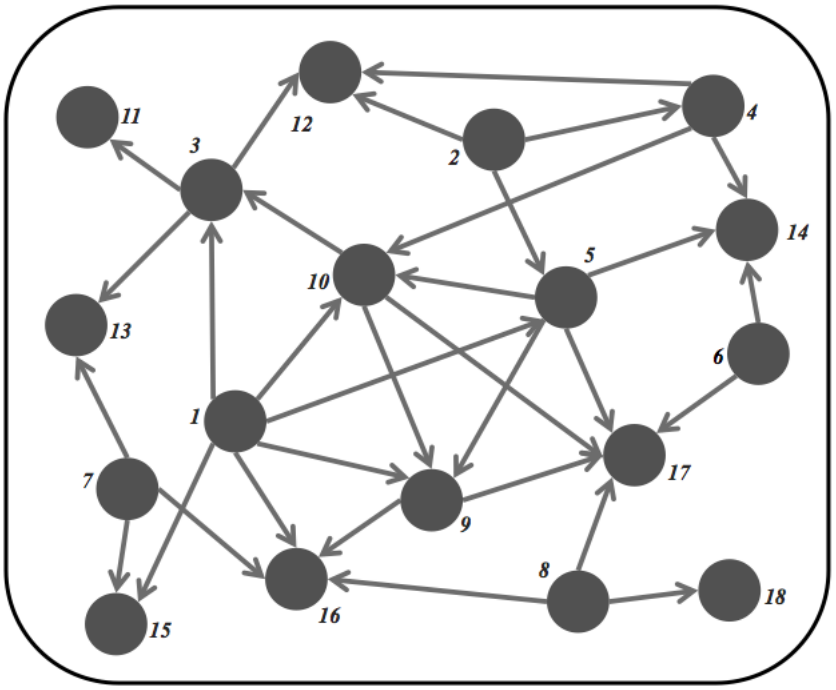

FIGURE 5.1: An Artificial network of related nodes (publications)

To analyse correlation, the Spearman Rank Correlation Coefficient (introduced in Section 3.5) can be used to return a correlation result ranging between -1 and +1, reflecting a negative (or indirect) and positive (or direct) correlation respectively. The Spearman Correlation Coefficient (Equation 3.6), is a non-parametric test, making it well suited when dealing with a ranked list of publications which do not exhibit a normal distribution (as outlined in Section 3.2 which explained the various principals surrounding bibliometrics and how Bradford's Law shows that citation metrics conform to the Power Law).

In order to examine if any of the metrics are revealing any significant recent publications, requires that each publication in the artificial network be assigned an age. This can be achieved by making the assumption that all publications will get cited at some point in their life cycle, thus the inverse means that those which are not yet cited are the newest. This is a large assumption to apply on such a network, however this does represent the most relevant approximation. Logically a new publication is most likely to be un-cited, with the opposite true for the oldest publications. Following on from this, those which are cited by others which are cited themselves, are thus the oldest. Figure 5.2 demonstrates the result of applying this assumption to the artificial network of publications. Here publications marked with an "O" are the oldest, "E" for those which are established and "N" being those which are new publications.

Using this information allows examination of age of the top ranked publications identified by each algorithm and hopefully a small difference in the results can be demonstrated.

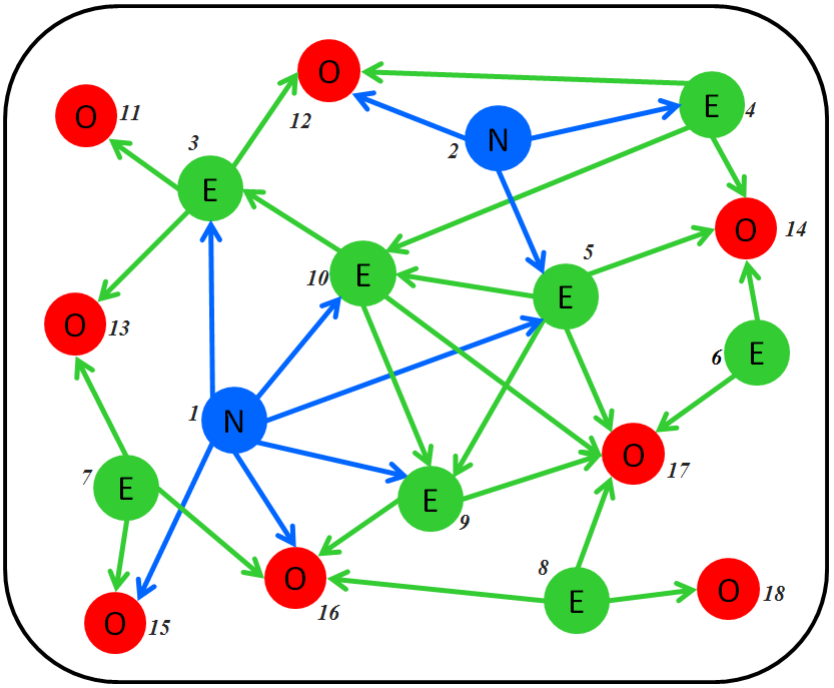

FIGURE 5.2: Publication Ages: O - Oldest, E - Established, N - Newest. Age is inferred from citation patterns

It is expected that maintaining a high rank for the older more respected publications whilst also identifying the up-and-coming publications will not be an easy task.

5.3 Co-Citations

Section 3.7 briefly outlined how a co-citation network will grow at a much greater rate than a plain citation network. A co-citation is created when two or more publications are cited together in the references section of a single other paper. Thus every time a publication is cited (singularly) it is likely to be co-cited many times.

A co-citation establishes a bi-directional link between the publication involved, thus creating an un-directed network of links. Figure 5.3 shows a simple network of directed citation links and the bi-directional co-citation links.

A number of the algorithms, including PageRank, utilise the properties of a directed network, where links have to be explicitly created in each direction between nodes. In a co-citation network, bi-directional links are implicit, however unlike on the Web these links are created by a third party, thus it is possible to apply PageRank to such networks (Perra 2008). PageRank for undirected graphs has been used in many bibliometric studies, including text summary (Wang et al. 2007) and sentence extraction (Mihalcea 2004). If a bi-directional link could be created by a single linking party on the Web,

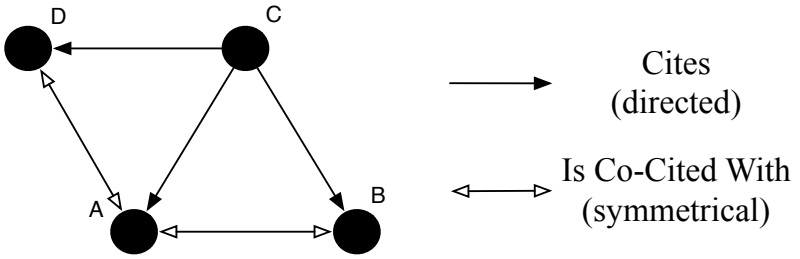

FIGURE 5.3: Directed citation network and the resulting un-directed co-citation network

then PageRank would not work as every page would be linked together and thus the favouring and weighting would not have any affect. Since the creation of co-citation links is dependent on a third party each link is just as relevant and a weighted algorithm will still work. Figure 5.4 demonstrates this by showing a directed network of co-citations with publication A; such a network can thus be used by each algorithm.

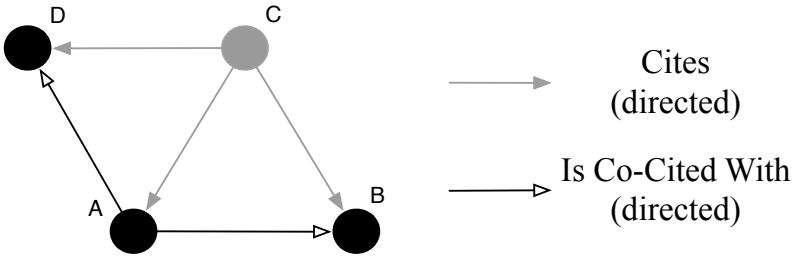

FIGURE 5.4: A directed co-citation network focused on publication A

In the case of the artificial network (introduced in Figure 5.1), 32 citations turn into 82 co-citation relations. This is limited to cases where a co-citation only links one publication with one other (${}_nC_2$). It is not uncommon to say that a single publication is often co-cited with two others, but this would cause this calculation to return all combinations of co-citations. With the average number of citations each node in the network gives being less than two, this is not a very large graph when compared to scholarly communications where typically each publication often cites 20 or more information sources.

Figure 5.5 shows the co-citations which exist for node 9 in the example network; here the three citations from nodes 1, 5 and 10 have provided a network of seven unique co-citations. If non-unique co-citations, where node 9 is co-cited with the same paper but via different citing nodes, are also counted then node 9 in fact has nine co-citations. In a co-citation network the non-unique co-citations can be used to suggest a stronger relationship existing between the publications involved and are often used to join publications to fields of research.

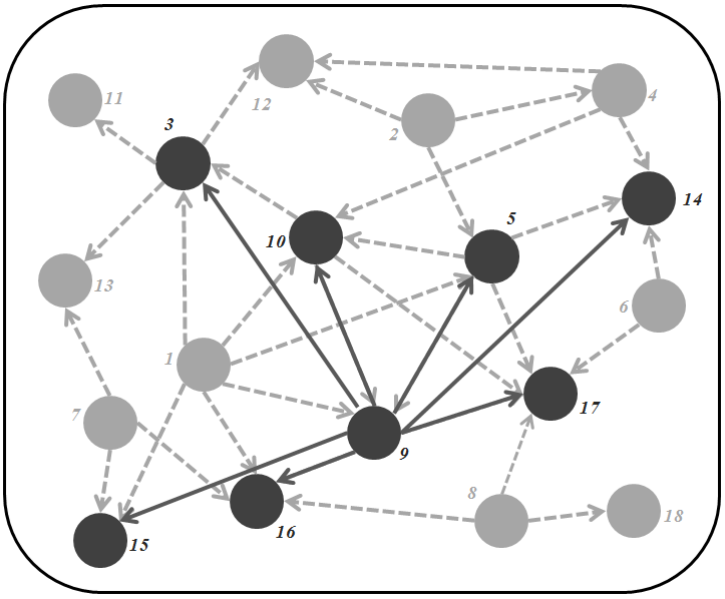

FIGURE 5.5: Example co-citation network centred on Node 9. Other than node 9 all dark nodes represent co-cited nodes.

Further, it can be observed in a co-citations network that many of the publications being co-cited are more established (older) then the citing publication. In Figure 5.2, take publication 9 for instance; this publication is cited by publications 1 and 5, however as a result of these citations, a co-relation is established with publications 14, 15, 16 and 17; four of the oldest publications in the network. This figure also shows how a citing publication can also be a co-cited publication in the case of nodes 5 and 10.

5.4 The CoRank Algorithm

The CoRank algorithm looks to utilize three key features of the co-citation network for the purpose of ranking.

- There are more publications in a co-citation network.
- These publications are likely to have a well established citation network.
- A co-citation network can include a publication twice, suggesting a stronger relation.

In order to take advantage of all three features, looking at the co-citation count was discarded due the close relationship to Citation Count (presented later in Sections 6.7 and 7.5) and the field of study (Section 8.2). Additionally, a simple co-citation count

does not take into account the prestige of the co-cited publication, something intended to remain in this thesis.

Prestige of a co-cited publication is maintained via initially basing the CoRank algorithm on the PageRank algorithm (outlined in Chapter 3). The key difference is that the data sources for the rank scores are the co-cited publications and not those which are directly citing. This means that a much larger number of links are considered when calculating the rank of a publication compared to those algorithms purely based upon direct citations. This is illustrated by Figure 5.5 which demonstrates how the three citations obtained by paper 9 produce nine co-citations with papers including 16, 17, 5 and 10 twice. The CoRank algorithm, shown by Equation 5.1, is based upon the principals of the PageRank algorithm, as such it is simply a variation of PageRank algorithm where the input data has changed rather than the algorithm itself. The key difference is that in CoRank, the rank gained from each co-cited publication is the CoRank of that paper $CR(p_j)$ divided by the number of co-citations (Co-Links) $CL(p_j)$ that paper has with other papers. In Equation 5.1 the paper p now represents a paper with which n is co-cited and not cited by. With opportunities to use the CoRank algorithm on other co-related data, including on the Web, the damning factor ($\propto$) remains and the number of co-citations is represented in all further algorithms as the number of co-links (CL). The equation is simply one level removed from the plain PageRank algorithm, using indirectly related publication, rather than directly related publication scores.

$$CR(n) = \frac{1- \propto}{|V|} + \propto \sum_{p_j \in M(p_i)} \frac{CR\,(p_j)}{CL\,(p_j)} \tag{5.1}$$

Unlike Citation Count and PageRank, in the CoRank network it is possible to obtain the CoRank of a paper more than once. Similar concepts can be seen in many online shopping websites. Such shops look at buying trends to see what products consumers buy together, if the relationship between two or more products is a strong one (they are bought together often) then this will become a recommendation for future users. The same principal can be mapped onto reader pathway metrics where a digital repository could look at what users are downloading over a short period of time and make recommendations in a similar fashion. Logically, this could then result in both publications being co-cited in a subsequent paper, making the relationship stronger again. CoRank simply takes this concept further and looks at how these relations can be used for ranking.

5.5 Four Metrics: Examination through Application

This section looks at the application of each metric; Citation Count, HITS (Hubs and Authorities), PageRank and CoRank to the artificial network of publications. Citation Count is currently the default metric for use in the area of scholarly publications and will be the reference metric when performing initial comparisons between the various metrics trialled in this section. In order to find alternative metrics to Citation Count, it is first necessary to examine the behaviour of any new metric against the accepted standard metric, as a significantly different metric is not likely to be that beneficial or see quick adoption.

5.5.1 Citation Count

Citation Count is the simplest of all the metrics, calculated by adding together all the citations towards each publication and recording this as the total. The publication with the highest citation count is thus the most highly ranked. Figure 5.6 shows the citation count of each publication and each node has been scaled to reflect its citation count. Therefore the larger the node, the higher its rank. Node 17 is the most cited, followed by node 16 and then nodes 9, 10, 12 and 14 sharing three citations each.

This section will be constantly referring back to the result obtained through application of Citation Count on this network, thus this result, shown by Figure 5.6, is acting as a control. In this figure the nodes have been scaled to show which are the most (biggest) and least (smallest) cited. By leaving the nodes scaled by Citation Count in all subsequent results, the top five or so nodes by each algorithm can quickly be compared to the Citation Count (accepted metric) result, enabling a visual comparison between each metric. While the size of the node will remain constant, the numbers contained inside each node will then represent the position of that node according to the algorithm being trialled.

Citation Count is a basic measure of popularity resulting in many of the nodes in the artificial network are ranked the same, i.e. the same number of citations. While this is not so significant in a publication network due to the fact that impact is a boundary measurement (e.g. a publication is typically regarded as having very few, a good number or a large amount of citations) not a comparative, this will have a greater effect when comparing ranking algorithms. The starting point for looking at the theory of how different ranking algorithms affect the network is to look at a number of existing algorithms, including PageRank and analyse how these compare to Citation Count. Consequently this will allow a direct comparison of Web based to traditional publication metrics.

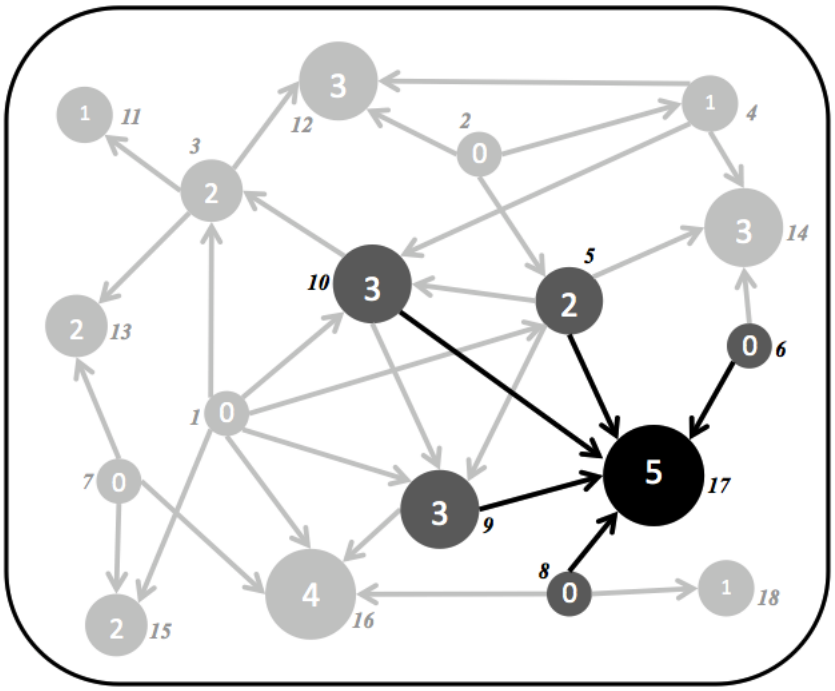

FIGURE 5.6: Citation Count applied to artificial network — Nodes scaled and internally labelled with their Citation Count

5.5.2 Hubs and Authorities

This is the first of the Web based metrics introduced in Section 3.6.1. Hubs and Authorities (HITS) (Kleinberg 1999) consists of two interacting recursive algorithms, one which calculates scores for Hubs and one which calculates Authority scores. Thus a good Hub is something which points to good Authorities, thus is calculated from the Authority scores, and a good Authority is pointed to by good Hubs, thus indirectly depending on the Hub scores.

Using HITS to locate individual high ranking articles can be achieved by only considering the results of the Authority score calculation. Figure 5.7 shows the Authority result of five iterations of the HITS algorithm over the artificial network.

With each node initially receiving a score of one, the algorithm is required to be iterative. The number of iterations depends mainly on the initial values used for each node (here $1/|V|$ where $|V|$ represents the number of nodes) and a bit of experimentation. Brin et al. (1998) found the number of iterations required for their PageRank algorithm (a metric similar in requirement to HITS) to converge to be linear in log_n. At the time their experimentation over a link graph of 322 million links converged in roughly 52 iterations with half this number of links requiring 45 iterations. With only 32 links in the test network it was found, by experimentation, that five iterations was perfectly suitable for the results of HITS, PageRank and CoRank to converge and stabilize.

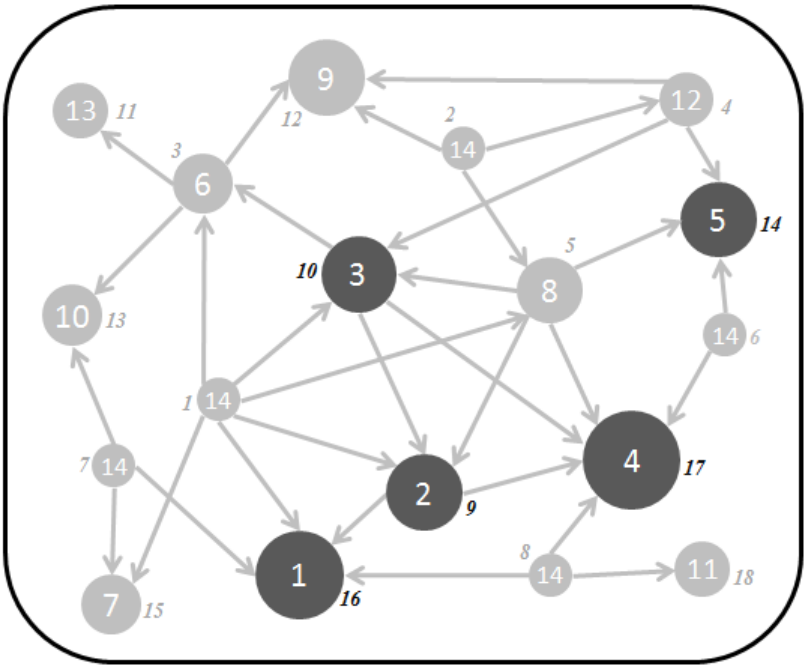

FIGURE 5.7: Nodes ranked by Authority Score — Internal numbers represent rank position while size remains an indicator of citation count standing

In Figure 5.7 the nodes remain the same size as dictated by their Citation Count, thus the bigger the node the more citations it receives. The authority rankings are represented by the numbers contained inside the nodes (unlike in Figure 5.6), thus node 16 (the node numbers are outside the nodes) is ranked as the most authoritative. By citation count the most cited nodes were 17, 16, 9, 10, 12 and 14 all of which received three or more citations, compared with citation count the order by authority scores changes to 16, 9, 10, 17, 14. Only node 12 is missing from this list of the top five, as even though node 12 receives three citations these have not be deemed to be from strong Hubs.

5.5.3 PageRank

PageRank (Brin et al. 1998), explained in Section 3.6.2, works on the basis that each link in the network does not obtain an equal weight. The introduction of this weighted system, where a rank is dependant on the rank of the linking item, is seen as an ideal mechanism through which false positives can be handled whilst maintaining a high position for important articles.

Although in a peer reviewed environment false positives are rare, as publication mechanisms become more open and the amount of available material grows, the suitability and necessity to consider such factors may become apparent. On the Web, if the rank of a Web page was calculated by simply adding up all the links towards a Web page then

it would be relatively easy to simply publish new pages, which provide links purely for the purpose of increasing rank.

In the context of this study, comparing the performance of PageRank against citation count will show how the two algorithms are related and help to place CoRank, which is based upon PageRank.

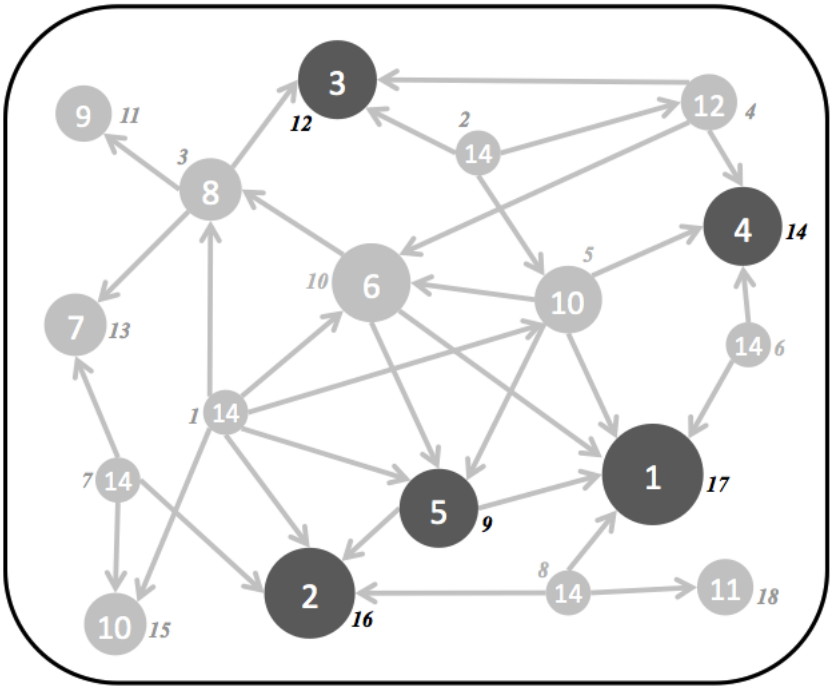

FIGURE 5.8: Nodes ranked by PageRank — Internal numbers represent rank position while size remains an indicator of Citation Count standing

Figure 5.8 shows the results of applying PageRank to the artificial network for a series of five iterations (the minimum number for the rank positions to stabilise on this network). From this, it can be seen that the results are very similar to that obtained by Citation Count; the top six including all the nodes which have a citation count of three or more. With PageRank being a more complex algorithm than Citation Count, a more finite rank order of papers is obtained separating those that have the same Citation Count. Figure 5.8 also demonstrates how PageRank works. Both nodes 10 and 12 are cited by three nodes of equal total weight (here defined as total citation count), the difference between the two is node 3. Node 3 cites node 12 and is also cited by node 10. Due to the high PageRank of node 10, the rank of node 12 is increased via the direct citation from node 3, thus demonstrating perfectly the iterative nature and dependencies of PageRank.

Node 10 is also newer than node 12 (from Figure 5.2) thus it would be interesting to see if PageRank reveals newer publications; one of the aims of finding a "better" algorithm. This is something very unlikely however, as each publications PageRank is based directly

upon the citing publications PageRank recursively, thus PageRank should take longer to establish than Citation Count.

To obtain the result shown in Figure 5.8, a number of iterations were required. From the work of Brin et al. (1998), who looked at ways to optimise PageRank to be computable in reasonable time, it was discovered that computation time was scalable in log_n. This meant that for an exponential amount of citations, rank order can still be computed in reasonable time. In their implementation Brin and Page also discuss the removal of dangling links, these are links to nodes which do not have any outgoing links (an example can be seen between node 5 and 14 in the artificial network, where node 14 has no outgoing links). Upon examination of how initial removal of these links would affect the results, no discernible difference was found. This is perhaps why there is no mention of this concept in the subsequent publication on PageRank (Brin & Page 1998), possibly also due to the negligible difference in compute time to find and remove them verses leaving them in.

When applied to the number of resources on the Web Brin and Page added a damping factor d to PageRank, which aims to model a "random surfer" who will not continue to follow more than about 15% of links (when $d = 0.85$). When PageRank is applied to links on the Web, a starting place is chosen to start and then only 15% of links are followed before starting again somewhere else in the network. Completing a full iteration for every page on the Web is seen as both intractable and also not required for the results to be accurate; if the website is popular enough then the likeliness is that it will get linked by enough people to become picked up in the 15% of followed links. Brin et al. (1998) realised that the damping factor could be used to favour some websites over others, however this mechanism would only be useful to users who wished to create their own custom search preferences.

15% may also be a good indicator for the number of citations a reader may choose to follow between scholarly communications, such as that outlined during the Reader Pathway (RP) study looked at earlier (Section 4.2). Bollen et al. (2005) examined how readers of publications follow from one publication to another via the reference list. Although Bollen did not calculate the damping factor, it would not be a great surprise to find if a reader followed around 15% of citations in each publication. In the case of the artificial network this damping factor will have no effect as PageRank is being applied in it's eigenvector form where a number of full iterations over every node will be completed.

Since the artificial network only details 18 publications, there is no need to selectively follow links and a full iteration can be easily undertaken. PageRank scores for all of the nodes can be computed without requiring the random jumping between them, therefore whatever the damping factor is set to, it will have no effect on the rank order result.

5.5.4 CoRank

Figure 5.9 shows the result of applying the CoRank algorithm to the artificial network of 18 publications. As with the PageRank a damping factor of 0.85 has been maintained and five iterations performed.

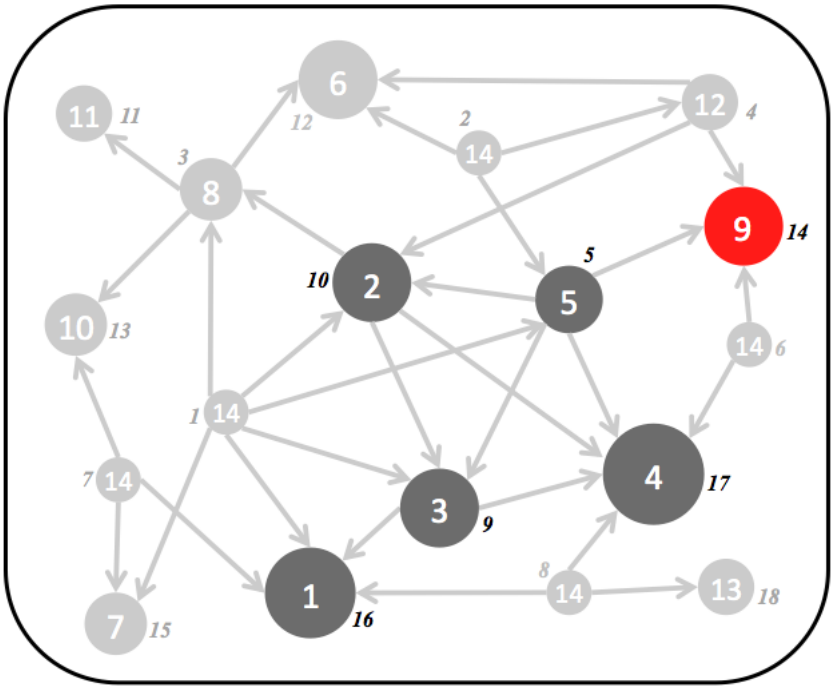

FIGURE 5.9: Nodes ranked by CoRank —- Internal numbers represent rank position while size remains an indicator of Citation Count standing

In Figure 5.9, the result of applying CoRank is not that much different to that found with Citation Count and PageRank metrics, with the exception of nodes 10 and 14. Previously node 14 had been highly ranked however in this instance, the co-citation network for node 10 consists of many links with nodes 9, 16 and 17 which are ranked highly themselves. Conversely node 14, while linked to nodes 9, 10 and 17, is only co-cited with these nodes once (twice for node 17). This shows that node 10 is more strongly related to these nodes, picking up their CoRank on multiple occasions, resulting in the higher rank.

In order for CoRank to be considered further, it must first satisfy a number of conditions which suggest that it provides some benefit without loss of core features maintained by other metrics. If CoRank is to reveal a number of more recent publications, it must do this while not demoting significant old publications. From the results in Figure 5.9, CoRank gives node 16, one of the older publications, the highest rank while the remainder of the top five consists of newer (established) publications. Out of all of the algorithms trialled, CoRank reveals a number of less old publications near the top of its ranked list, however this is all caused by a shift of only two nodes. This represents a

hopeful indication of stability and potential application for CoRank to applied against a real dataset.

5.6 Evaluation of Initial Results

One of the criteria that defines a good metric, is to not be significantly different from current metrics but to be more accurate, in less time. Using the Spearman Correlation Coefficient (as applied in Section 3.5), enables the comparison of each metric by analysing the ranked list of results produced by each. Table 5.1 shows the results of applying Spearman to each combination of two algorithms.

	Auth Score	5x PR (0.85)	5x CR (0.85)
Cite Count	0.93	0.96	0.92
Auth Score		0.89	0.95
5x PR (0.85)			0.88

TABLE 5.1: Spearman Correlation Rank of artificial network results

With no real drastic difference between any of the results in Table 5.1, all of the algorithms applied to the artificial network have returned a very strong positive correlation to each other. This result re-enforces the similarities between the existing algorithms, as well as providing a good indication that the theory behind CoRank is valid and may prove to be a beneficial new metric.

In order to better see the relations between the algorithms, Figure 5.10 shows a chart plotting the Spearman Correlation of each algorithm to the order obtained by Citation Count.

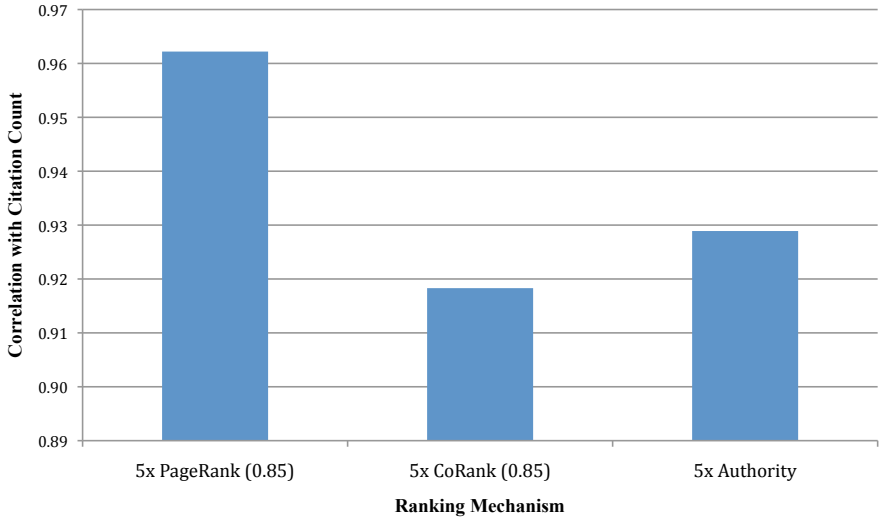

FIGURE 5.10: Correlation between ranking algorithms and Citation Count

Although Figure 5.10 looks to be showing a large difference between the various metrics, the y-axis has been limited to show only a small range of values. This has been changed specifically to emphasize the differences between the algorithms when applied to the artificial network. It is expected that these small differences will become very important when each algorithm is applied to a larger set of data.

At this stage, there is nothing to suggest that CoRank is going to produce significantly different results from the other three algorithms. With one of the conditions of producing a 'good' algorithm fulfilled, the bonus characteristics, such as age of publications can be addressed.

Table 5.2 examines the relative ages of the top six publications as ranked by each algorithm. Here the top six publications are categorised into the three different age brackets and a percentage figure is assigned. This table also shows the precise order of publication by age (via the order column) where the letters representing age (**O**ld, **E**stablished and **N**ew) are given. Any equal ranks are depicted with brackets.

	Order	% old	% established	% new
Citation Count	OO(OOEE)	66	33	0
Authority Score	OEEOOE	50	50	0
PageRank	OOOOEE	66	33	0
CoRank	OEEOEO	50	50	0

TABLE 5.2: Percentage comparison of publication age per algorithm, top six publications

Table 5.2 shows that the behaviour of Citation Count and PageRank is exactly the same for the first six top ranked publications. CoRank selects roughly the same publications for its top six (it discards node 14 in favour of node 5), but it ranks them such that more established (rather than old) publications are revealed ranked higher.

This result also reflects the earlier correlation graph (Figure 5.10). Again HITS and CoRank demonstrate a similar behaviour to each other, revealing a number of more recent publications higher in their ranking.

5.7 Conclusion

Having completed an analysis of four algorithms, including CoRank, on an artificial network of publications and the citation links between them, it is possible to conclude based upon this network, that CoRank is not significantly different in performance to citation count or PageRank, but holds the closest relation thus far to the HITS metric. This is evident both by the correlation between the two algorithms as well as the age of publications within the top third of the results.

It is interesting to see that the HITS algorithm behaves in a manner consistent with the suggested outcome, that of revealing high impact articles sooner in the publication life cycle. While it was anticipated that CoRank should achieve this outcome, with the same 50/50 split between recently established and older publications, it was not expected that HITS, which shares many characteristics with PageRank, would obtain this result.

The strong correlation between all of the algorithms also suggests that none of the publications in the artificial network have been considered to be providing false citation information. If this were the case, both PageRank and CoRank would show significant differences from Citation Count. Again this is something which may become more evident in a larger publication network where a number of lesser cited publications may be ranked more highly do to having a good number of high prestige citations. Theoretically publications should not provide false citation information, but indexes such as Web of Science have still historically decided to remove author self citations from their statistics.

It is important to emphasize that although a number of possible conclusions have been drawn in this section, these have all been based upon a artificial network. The artificial network was created purely to demonstrate how each algorithm works identify what data is required to apply each metric and analyse the results. By expanding this study to a real network containing millions of citation links, demonstrating how each algorithm works will be impractical. Further it is not just the application techniques which will carry forward, but also the methods used to analyse the algorithms performance, including age of publications revealed and correlation between each set of algorithms.

When applying PageRank and CoRank to a full citation network, false positives will be replaced by citations from publications which are themselves not cited. These publications form the long tail of authors who have published a single article and those which have achieved very few citations. In this respect, PageRank and CoRank should result in providing a form of h-index to a publication. Rather than more citations being interpreted as an indicator of high impact, high impact publications are those which receive citations from other high impact (or prestigious) publications.

It is due to the desire to maintain this weighting factor that assigns prestige to citations, that CoRank takes many influences from the PageRank algorithm and has so far proved itself to be of some worth against an artificial network. CoRank, like PageRank and many other metrics, could also be applied to many other data sources and graphs of information.

It this chapter CoRank was demonstrated to be a ‘good’ metric, based upon an artificial network, and exhibited many improved properties over existing metrics. With the explanation and theoretical study of the algorithms performed, it is now necessary to apply CoRank to a real and extensive citation network collected from genuine literature.

Chapter 6

Applying CoRank

In the previous chapter, the CoRank metric was introduced and examined against an artificial publication network. By comparing the behaviour of CoRank to several other algorithms on this network, it was found not to be unsuitable for use as a metric to examine scholarly publication networks. This chapter subsequently examines how CoRank performs when applied to a real dataset. In order to genuinely test the effectiveness of CoRank a substantial dataset is required. Careful selection of this dataset should ensure a number of metrics including CoRank can be applied in order to examine the relationships between these fairly.

The artificial network, used to examine the applicability of CoRank, modelled a network of the citations between publications at a single point in time. This meant that the only form of early indication analysis possible was through establishing the approximate ages of the publications and examine which metric is revealing the most recent publications among those ranked highly. The down side of this is that there is no way to confirm that these publications become ranked highly later in their publication cycle by pure Citation Count.

CoRank has been designed with the intention of being an early indication metric for impact. Thus, CoRank is not considering the impact of a recently published directly citing publication, rather the co-cited publications which have existed in the publication network for a larger amount of time. Additionally, since each citation establishes a good number of co-citations to other publications, the impact network is also much larger, quicker. To analyse if CoRank is an effective early impact metric requires a network of publications that can be tracked throughout their entire publication life cycle. This way both the impact by CoRank during the early stages of the publication life cycle, and the final impact by the highly regarded Citation Count metric, can be found and compared. Additionally, the potential benefit of the other metrics introduced in Chapter 5 have not been ruled out, thus these are also going to be applied in the same way as CoRank.

In a live environment, publications can be tracked over their life cycle by recording rank positions each month, and as months pass, compare the results. This means that to track a publication over a three year life cycle, either involves performing an evaluation in real time over three years, or locating a dataset which can be retrospectively backdated. In order to save time and to avoid the need to start again in the case of an error, the latter of these options was chosen.

Following a short study of available datasets, the Citebase[1] dataset was chosen as the best source of information. With over 308,000 publications indexed along with all their citations, the Citebase dataset not only provided a good range of data, but also allowed for this data to be easily backdated to re-establish past states in Citebase. The first section of this chapter examines the Citebase dataset and looks at the number of citations and co-citations indexed over a period of three years. This data enables the calculation of how many iterations of algorithms, including PageRank and CoRank are required to stabilise the rank scores for every publication in the dataset. Additionally this gives an early indication on the time required to apply all the metrics to every snapshot, which subsequently has a heavy influence on the system designed to apply each metric and output the results.

The fastest way to compute any results is to load all the relevant data into very fast memory close to the processor and not rely on a high number of disk reads. This became the core requirement of the system tasked with loading and applying each metric on the Citebase dataset, named the "Co-Ordinator". The Co-Ordinator, built as part of this work and outlined in Section 6.4, is a modular system where a number of algorithms can be defined, each requiring a number of datasets. The Co-Ordinator manages the whole process by gathering together the datasets and applying the metric in memory, without the need for high levels of disk access. Additionally the Co-Ordinator is designed to be a simple system able to run on a number of architectures, thus allowing it to be developed on a smaller platform, before being transferred onto a much more powerful platform in "the cloud" for fast, reliable application of each algorithm.

Section 6.2 outlines the testing plan for processing of the results with Section 6.6 providing an overview on the high level interfaces built on top of the Co-Ordinator which allow results to be gathered and processed quickly. Again these interfaces, designed to be as flexible as possible, allow a number of parameters to be set which change the input data, process performed and module used for outputting results.

The bulk of this chapter (Section 6.5) presents the results from the first application of each algorithm upon the Citebase dataset. Much like in Chapter 5, the output of each test is analysed individually, paying close attention to the results produced. This initial application is applied to a small sample of the Citebase subset. In Section 6.6 the testing is opened up to cover a much larger set of publications, thus allowing the results to be

[1] Citebase - http://www.citebase.org

cross checked and ratified. Additionally, in this section the results obtained by each metric are overlaid on top of each other to provide a much simpler way to evaluate the results obtained. Lastly in this chapter, the results are evaluated further against the characteristics of the Citebase dataset.

6.1 The Citebase Dataset

In order to examine the performance of CoRank against the other metrics, a real dataset is required. The Citebase citation registry, designed specifically for the purpose of indexing citations over time, provided the perfect candidate. This registry, developed as part of the Open Citation project (OpCit) (Hitchcock, Brody, Gutteridge, Carr, Hall, Harnad, Bergmark & Lagoze 2002) was one of just a handful of open citation registries available at the time, along with NEC's Research Index (Lawrence et al. 1999) and CERN's Document Server (Claivaz et al. 2001). The role of Citebase in the OpCit project was to help demonstrate the benefits which come from not only open access publishing, but also open access citation data.

Today, Citebase still extracts references from the larger Open Archives Initiative (OAI) disciplinary archives, namely arXiv[2], CogPrints[3] and BioMed Central[4]. These citation lists are then associated with the OAI metadata record for the document in which they are identified and stored. This data is then used to build a "classic" citation database of document citations from which the forward citations are calculated, i.e.:

- Document B cites document A
- Document C cites document A
- Document A is cited by documents B and C

Importantly, although Citebase is harvesting from a limited number of registries, the citation lists are not filtered to contain only those citations to documents about which Citebase has a record. Therefore, although Citebase may only contain 308,000 (as of February 2007) publications, this equates to over 3 million citations.

Citebase was designed to be a "Google for the refereed research literature" (Hitchcock, Brody, Gutteridge, Carr, Hall, Harnad, Bergmark & Lagoze 2002), and to the results of each search a number of different ranking algorithms can be applied. Due to the controlled nature of the network, Citebase could quite realistically apply a Citation Count algorithm in order to rank the results. Going beyond this, its makers wished to

[2]arXiv - http://arxiv.org/
[3]CogPrints - http://cogprints.org/
[4]Biomed Central - http://www.biomedcentral.com/

allow a number of other rankings to be available including hit count, which calculates the number of downloads. Applying temporal information to the ranks also allows users to order a search by popular downloads over a particular period.

Due to the fact that Citebase infers forward citations, the Citebase interface for a single record also lists other interesting facts about each record still not present in modern day services, such as Google Scholar. These include:

- Graph of Articles Citation/Download History
- Citation list for this publication
- Top 5 citing publications
- Top 5 co-cited publications

These services not only allow users to see the current state of a publication, but also its history, citation lists and with whom this publication is co-cited. The top five citing and co-citing publications refer to those publications relating to this article which are themselves highly ranked. Where publications are co-cited many times, this count can be summed to provide a co-cited score. The top five are those with the biggest co-cited scores and, in the event of a tie on co-cite score, they are then ordered by rank. Hitchcock, Woukeu, Brody, Carr, Hall & Harnad (2002) discusses the relative merits of having all of this information available and its usefulness to the targeted customer base. In Citebase, indexing the co-citation allows information to be presented to the users, informing them about closely related publications to the current one; the typical use for the co-citation.

Having this network of co-citations already calculated (as a by product of the forward citations being inferred) means that the Citebase registry is a perfect source of data needed for testing CoRank. More importantly, because Citebase has the OAI (Open Archives Initiative) metadata record relating to the citations, it is possible to retrospectively back-date the Citebase dataset to a previous state, containing only citations relevant to that point in time. This provides a means to obtain many years worth of data relating to the growth of the citation network for each and every publication allowing testing of each metric against an entire publication life cycle.

In order to execute a series of metric tests upon the Citebase dataset, a snapshot was obtained containing all of the data up to and including February 2007. Some details about which are as follows:

- **No. of Publications:** 308023
- **No. of Citations:** 3.37 million

- **No. of Co-Citations:** 46 million

The distribution of citations for publications in Citebase is shown in Figure 6.1. Here a Power Law distribution can be observed, something expected from a dataset of this type.

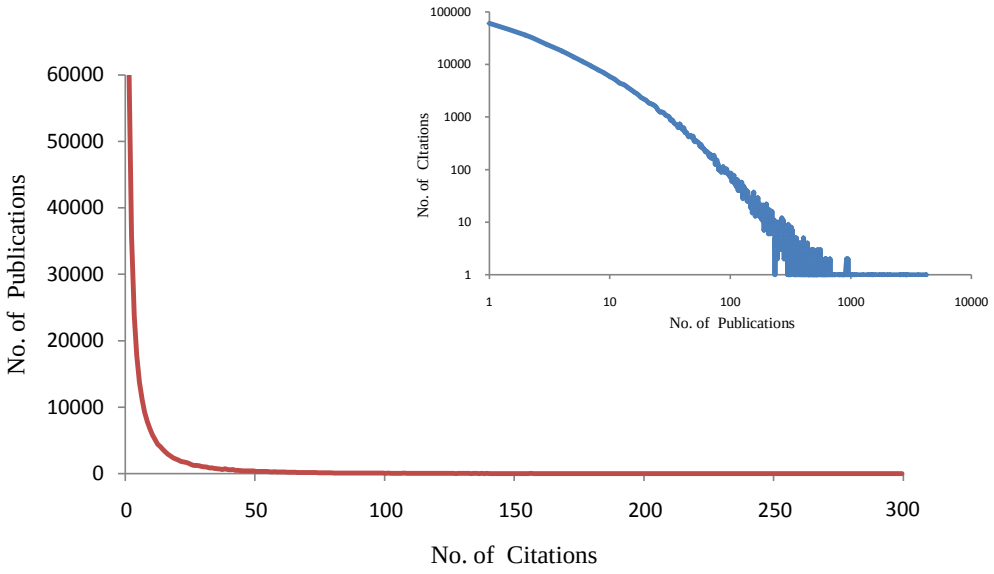

FIGURE 6.1: Distribution of Citations in Citebase

Approximately a fifth of the publications in the dataset have only one known citation, while one publication has over 4000 citations. There are in fact 459 publications which receive over 300 citations which, for clarity, have been removed from the main graph in Figure 6.1. These are included in the sub graph, which shows the logarithmic distribution representing the same data. Although the Power Law plotted as a logarithmic distribution should show a linear gradient, there are a small number of publications which receive exceptionally high numbers of citations, something which is common when plotting Power Law citation graphs.

6.2 Test Strategy

To best test the effectiveness of a number of publication ranking algorithms, over time, with the aim of finding high impact publications sooner in the publication life cycle, a dataset is required which can be retrospectively dated. This makes the testing, evaluation and fine tuning a far easier process than having to wait three years to test one algorithm on live data.

The Citebase dataset offered the ideal dataset, having been used for similar experiments previously (Brody & Harnad 2006), containing publication information pertaining to pre- and post- prints from many research areas such as physics, mathematics, information science, and biomedical science.

Starting with a snapshot of Citebase taken at the end of February 2007, 36 snapshots of the Citebase database were generated; one per month dating back three years. The latest snapshot was taken on the 1st of March 2007 and contained all data up to and including the end of February 2007. As Citebase logs the date that each publication is added to its database, snapshots could be generated by removing any publications added subsequent to the date required. Doing this resulted in a new Citebase database being created for each snapshot generated, each of which was stored in mysql[5]. Figure 6.1 gives an indication to the growth of Citebase showing the number of records and citations each snapshot held at that point in time.

Snapshot	No. of Publications	No. of Citations	No. of Co-Citations
March 2004	199107	1.94m	24m
September 2004	216031	2.16m	27m
March 2005	234593	2.39m	30.6m
September 2005	253691	2.63m	34.3m
February 2006	279304	2.85m	37.6m
September 2006	293613	3.17m	42.7m
Feburary 2007	308023	3.37m	46m

TABLE 6.1: Number of Publications in Citebase Snapshots

Table 6.1 demonstrates that Citebase contains a growing corpus of publications and has successfully harvested a number of citations from these. The average number of citations Citebase obtains from each publication grows steadily from 9.8 to 10.94 over the three year period, demonstration a growth in the number of citations scholars are listing. At the end of the three year period, Citebase has over 3.3 million citations indexed; around 13 times the number of publications.

Using a simple factorial calculation (${}_nC_r$) to deduce how many co-citations Citebase contains finds that the number of co-citations can be roughly calculated as being 200 times larger than the number of publications, growing as it does from 120.29 co-citations per paper to 149.29 over the three year sample period. Thus the network of co-citations is, on average, 15.5 times larger than the network of citations. In this instance a co-citation is defined as a combination (in any order) of just two publications, which are both cited by a single other publication; combinations of three or more publications have not been considered. Using this data, each paper in Citebase is cited an average of ten times and co-cited 135 times, demonstrating the drastic difference in the size of the citation and co-citation networks.

[5]mysql - http://www.mysql.com

6.3 Iteration Requirements Verification

As in Chapter 5 the number of iterations required by each algorithm on the link graph can be calculated partly from the work of Brin et al. (1998), and also through application. From the two sets of figures presented in Section 5.5.2 earlier, Page and Brin found the number of iterations required grew in a linear fashion proportional to log_n. Taking their work and doing the calculations backwards finds that the equation to work out exactly the number of iterations is roughly $6log_n$. Applying this to the Citebase network containing 3.37 million citations (from the most recent snapshot), 3.37 million can be substituted for n with the result suggesting that at least 39 iterations are required.

Verification of this result is recommended however, this is a straight forward process to test that the basic three algorithms (HITS, PageRank and CoRank) stabilise their rank order within 40 iterations. A stable rank, is one that does not change substantially between one iteration and the next. Thus to test that each algorithm stabilises within 40 iterations, after each iteration the rank order is recorded and compared with the previous order. Figure 6.2 demonstrates that all three algorithms stabilise within the specified 40 iterations, here each point on each plot line represents the correlation with the previous point. Thus as soon as the correlation is constantly one, the algorithm results can be said to have converged.

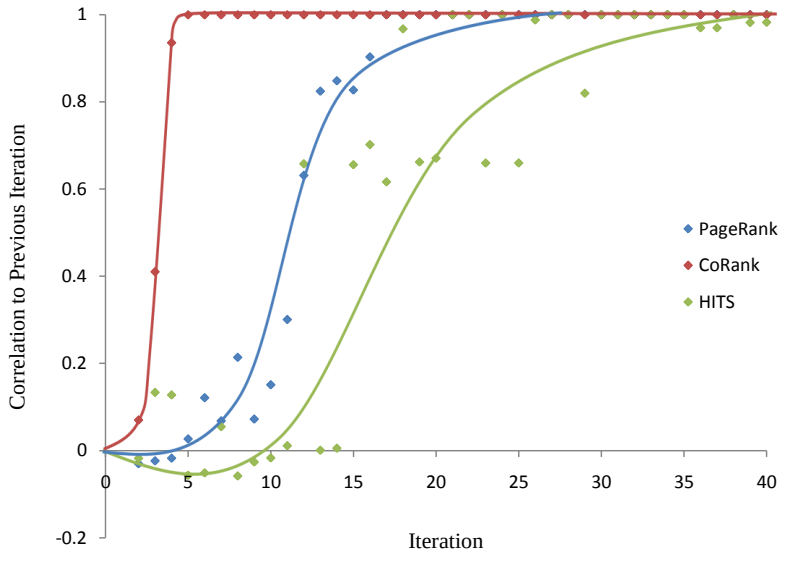

FIGURE 6.2: Iteration verification for application of core algorithms

Figure 6.2 reveals some interesting characteristics of the three algorithms in question. PageRank stabilises in around 18 iterations, while CoRank is a lot quicker. This rapid stabilisation is most likely due to the larger network of co-citations, compared to citations, being considered by CoRank. As HITS is a two part algorithm — Hub score is based on Authority score and Authority score is based upon the Hub score — it is not

surprising that this takes longer than PageRank to stabilise. This figure demonstrates that it is very important to consider the data required by any iterative algorithm, to judge how long it may take to stabilise rank.

Trialling this on the most recent Citebase snapshot (that containing the most links and thus the most likely to take the longest to stabilise), found that 40 iterations was a good approximation. Although in a lot of cases a lesser figure could have been chosen, it was judged that performing extra iterations to ensure stability is less expensive (due to the marginal difference in computational power required on the Citebase dataset) than potentially obtaining inaccurate results.

6.4 The Co-Ordinator

With both a number of snapshots and algorithms having to be processed it was decided that a generic system would be built, capable of systematically processing large amounts of data and also analyse it to obtain the various results. Figure 6.3 gives a broad overview of the system. The Co-Ordinator, the name for the system, consists of three main parts, the snapshot, metric and results processors. The first of these creates the individual snapshots of Citebase from the master snapshot provided. The master snapshot is a database dump of Citebase, containing all the information pertaining to the publications in an already normalised form including a table which lists the Citation Count for each publication. The snapshot processor generates previous snapshots in time, it does this by first removing all references to publications added after the date required, then re-calculating the Citation Counts and links between publications. The key table in each snapshot contains the citation information relating a source publication to all the papers it cites.

With the snapshots obtained the next part of the system to be developed was required to do the ranking itself. After various experiments with a few algorithms on a single snapshot, it was found that constantly querying the database to find citation and co-citation information was intractable. If one query has to be performed a citation is looked up for a single publication, about 3.3 million are required queries to find all the co-citation data. Further, a significant number of additional queries are required to write back ranking data to the database pertaining to a single publication. Operating in this manner would mean processing times spanning days to calculate the results corresponding to one algorithm.

To help solve this problem the algorithm methods were separated from the metrics processor itself (as shown in Figure 6.3). This means that the metrics processor, although the main part of the system, could now focus on reading the data into data structures optimised for the algorithm methods, followed by the saving of subsequently generated rank results.

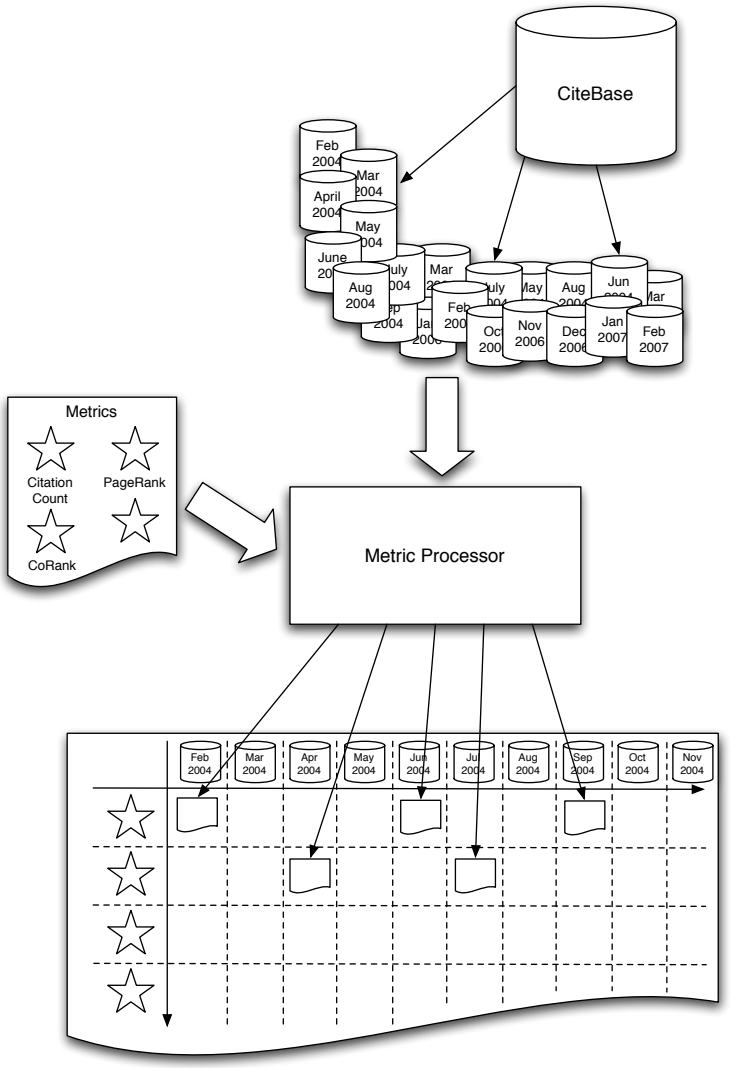

FIGURE 6.3: Architectural view of the Co-Ordinator's Snapshot Processor

When writing the data, the metrics processor creates a new table relating the snapshot to the algorithm being executed. This new table contains only two fields, the publication Citebase ID and the rank score, for example publication 12345 could have 43 citations, or a PageRank of 0.21555.

The metrics processor is outlined by Figure 6.4. Here the data importer and indexer builds the data structures required to execute a number of algorithms. These in-memory data structures were optimised such that data pertaining to each publication could be referenced quickly. Information regarding publication IDs of citing and co-cited publications, as well as the total count of these publications, was designed to be accessed quickly with computation requirements kept minimal.

Figure 6.4 shows two of the main data structures required for the majority of metrics. These are then used by the metrics processor which builds the output data structure

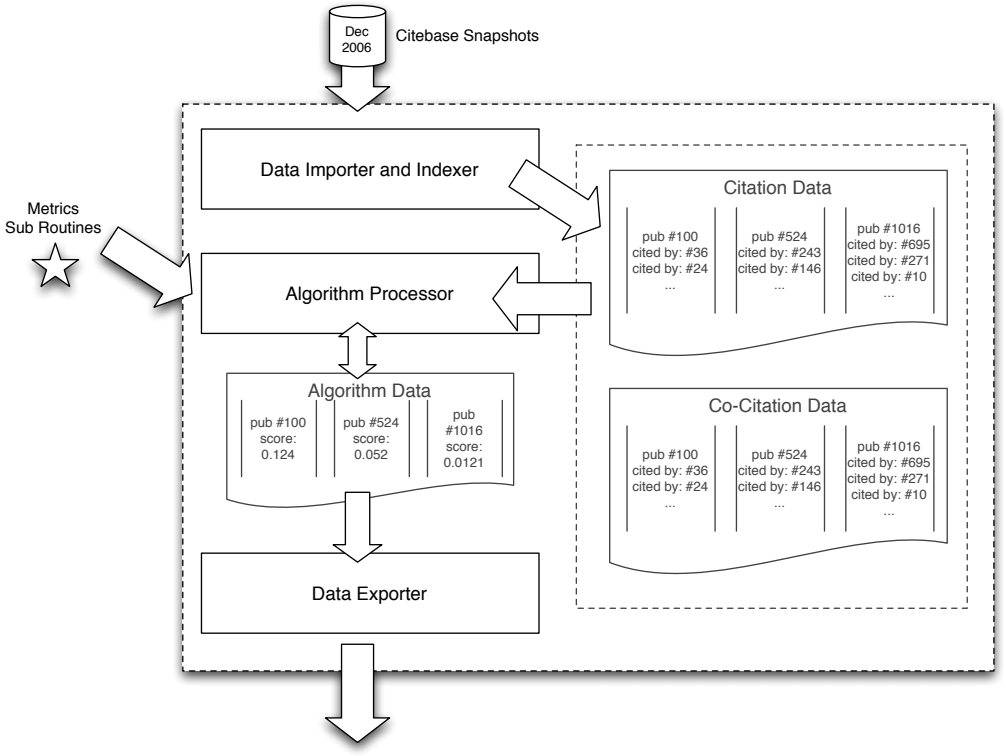

FIGURE 6.4: Detailed view of the Co-Ordinator's Metric Processor

(another hash table) containing the publication IDs and their rank scores. The metrics data is stored in a similar way to the input data as many of the metrics are iterative and depend on this data for subsequent iterations (as shown by the bi-directional arrow between the processor and data in Figure 6.4). Once the metric processor has finished with each algorithm for the appropriate amount of iterations, it hands the resultant data to the data exporter and is ready for the next operation.

With the algorithm processor optimised to perform calculations quickly from in-memory data, the longest part of the whole process is building these in memory data structures each time. Previously each algorithm would begin executing almost immediately, however it would take a number of days to perform all the database read and writes required to gather information pertaining to the publications and citation network. By taking a number of minutes to index this data in-memory, this bottle neck is avoided and each algorithm then takes only a few minutes itself to process. In testing 40 iterations of the PageRank algorithm took nine minutes to process on a machine with four cores and 8Gb of RAM, a minute of this time was used to load the data.

Finally, the data exporter takes the data from the algorithm processor and serialises this into a single query, which commits the data into a table in the results database corresponding to the input snapshot.

In order to speed up the processing further the whole system was deployed in "the cloud". Using Amazon's Web Services[6] the storage and database requirements could be separated from the processing. This meant that the high performance Elastic Block Storage (EBS) storage volume, a very fast striped and mirrored raid partition, could be connected to a number of different specification compute facilities depending on the requirements of each computation. Amazon offers a number of tiers of compute facilities ranging from "micro-instances" which are less powerful than most peoples phones but practically free, to "extra-extra-extra large" compute clusters which provide 10's of cores of processors and 100's of Gigabytes of memory. For the purposes of running the Co-Ordinator over the full Citebase dataset it was found that a large instance with four processor cores and 8Gb of memory was suitable for the processing phase; during the evaluation stage this could be downsized to a small instance, thus saving money and energy. Figure 6.5 provides a simple overview of the deployment of the Co-Ordinator system in "the cloud" showing the constituent parts, including the ability to be able to back up the high performance volume as a "snapshot".

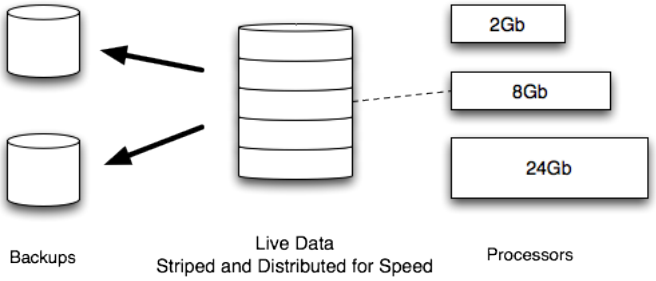

FIGURE 6.5: Overview of deployment of Co-Ordinator in the cloud

The storage volume not only contains the database snapshots, but also the metrics processor and analysis tools. Upon connecting this volume to a processing machine a single script is executed which prepares the processor by installing all the pre-requisite software and re-establishing the database connections. This also means that the storage is agnostic to the platform it is connected too, as shown by Figure 6.5. Here the processors can be of different specifications and kept up to date separately from the storage volume.

The main benefit of out sourcing the processing to "the cloud" is this lack of requirement needed to maintain a powerful local environment to process, store and backup the results. Additionally as hardware and storage becomes cheaper, so the amount of processing that can be done with the same amount of money increases. Thus as this work progressed, the amount of memory available in the processing machines and power of each has increased a number of times, making results quicker to obtain.

[6] Amazon Web Services - http://aws.amazon.com

6.4.1 Processing Results

Applying the metrics processor results in every publication obtaining a rank score relating to each algorithm. Although it is possible to analyse all 200,000+ publications, this does not help track sets of publication through their life cycle. The aim is not only to find a metric which still performs well when compared to Citation Count, but to evaluate if any metric can additionally be used to provide an early indication of subsequent impact.

In Figure 6.1, in can be observed that the majority of the 200,000 publications in Citebase will have very few citations, meaning earlier and subseqent impact are likely to be very similar. In order to properly evaluate the family of applied metrics, publications are required which gain subsequent impact. To alleviate this problem it was decided that a set of 100 publications, all of the same age, should be tracked over a three year life cycle. To ensure that this 100 does not consist of the 60,000 which only ever receive a single citation, the top 100 by Citation Count at the end of a three year life cycle were selected. Even after careful selection, there is still a chance of selecting 100 publications which never get ranked highly by CoRank, however when considering Citation Count as the target metric (due to adoption) than the others should be capable of reflecting this standing. If these 100 are ranked highly, earlier in the publication life cycle, then CoRank can be said to be an early indication metric.

To perform this comparison, the selected set of 100 publications, ordered by Citation Count, become the target set of publications which all algorithms are looking to match at some point in their life cycle. With the target set of publications selected, these must then be located in all of the snapshots for each algorithm, the rank position recorded and finally correlated against the target list.

Again it was decided to make the system used for locating and processing results as generic as possible and add this as a series of Web services on top of the Co-Ordinator. In all three primary services were added on top of the Co-Ordinator, listed as follows:

- **Correlation Calculator - month by month** - Provides early indication metric data pertaining to the order of the sample set of publications
- **Rank Comparator - month by month** - Tracks the overall positions of the sample set of publications in the whole dataset.
- **Publication Ages (Summariser)** - Provides a breakdown of the age of publications contained in many snapshots.

Figure 6.6 shows an overview of the system and the data required for each of the three tests. This diagram shows the Citebase snapshots along the top, with each algorithm

represented by a star. Thus each document within the table represents the set of results generated by the metrics processor. Figure 6.6 additionally shows the two basic sets of data required for each of the tests and which test requires which data set.

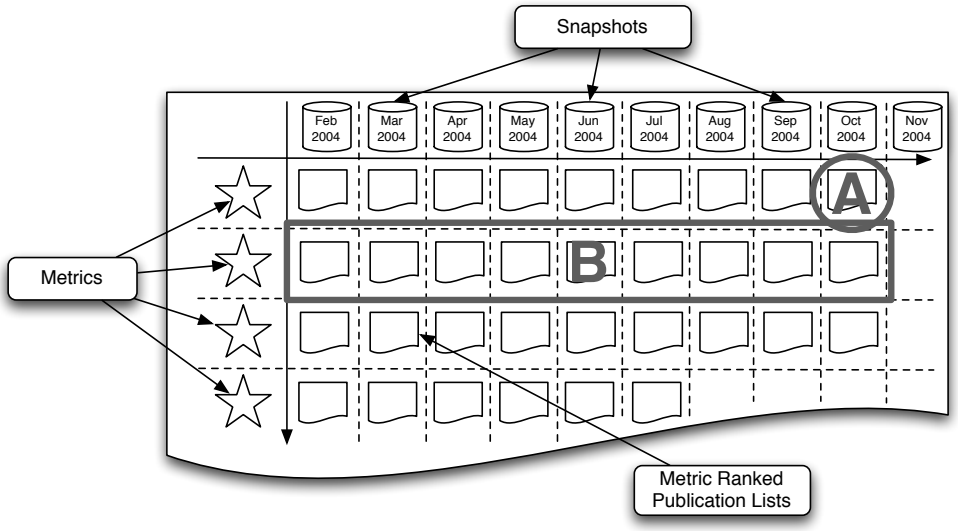

FIGURE 6.6: Overview of the Co-Ordinator's Result Processor

The following section introduces each of the three tests in more detail. Each test requires a number of different combinations of base and processed data, in order to not only evaluate each metric against each other, but also to do this in a temporal manner.

The first test is the Correlation Calculator, perhaps the most complex. This test requires both datasets and compares the rank of the target set of publications (shown as A in Figure 6.6) to each snapshot (one of the set B in Figure 6.6) generated by a each algorithm. To pick these datasets a number of variables can be defined, including which metric dictates where the target (A) is, and which metric to take the snapshot results from (B). The full list of parameters which can be provided to the Correlation Calculator is as follows:

- **Target Metric** - The metric used to obtain the target publications (A).
- **Target Snapshot** - The date of the snapshot which is regarded as the target (A).
- **First Snapshot** - Usually represents the snapshot three years prior to the target one for the start of snapshot (B).
- **Trial Metric** - The metric being trialled and about which all the results should be selected (B).
- **No. of Publications** - The number of target publications (if different from 100) (A and B).

Each set of publications selected as part of B consists of only the 100 selected during the sampling of A. Each set of 100 is returned in rank order where the rank position in the total dataset has been discarded; thus they are now ranked between 1 and 100. Both sets (A and B) are then processed by the correlation calculator resulting in a correlation coefficient being returned, representing the similarity between one of the sample sets in B and the target set A. Once a correlation coefficient has been calculated for all 36 snapshots in the set B, this can then be graphed over time to show how the correlation changes.

The second test involves the Rank Comparator and the same sets of information as the Correlation Calculator except this time, both the **A** and **B** datasets contain the actual rank position of the set of n publications in amongst all of the other publications. This is then used to assess the relative standing of the publications within the entire dataset.

It is important to examine that as well as the publications being ranked in a similar order to the target algorithm, they are also listed at a similar point in the overall standing when considering all publications. The output from this test will be an average rank value for the position of the set of publications and a value for the deviation to indicate how distributed they are from this result. This average rank can then be compared to the target rank average (from **A**).

Finally, **the third test** examines the age of the top ranked papers in each snapshot. This will provide a good indication of the behaviour of each algorithm on a real dataset, as well as help explain the expected variations from our target dataset. This final test, only requires the data pertaining to the algorithm being trialled. There is no target set of publications, rather the top $n\%$, as defined by the input to the test, are required to calculate the average age of this $n\%$. Thus this test can take the following inputs:

- **Trial Metric** - The metric being trialled and about which all the results should be selected.
- **No. of Publications** - The number of publications to be considered (as an alternative to percentage).
- **% of Publications** - The percentage of overall publications to be selected.

In the case of the results presented in this work, this test will be examining only the top 5% of all publications in rank order (as per the trial metric), not the previously selected 100. The age of the top 5% of publications will be recorded and each publication will be grouped into one of four age brackets: less than a year old, one to two years old, two to three years old and older than three years. The results will be presented in the form of a percentage breakdown of publications which are present in each category.

Previous research has found that a publication's impact can only be judged accurately after around three years (Moed 2005). Logically this would imply that the majority of high impact publications by Citation Count would be three years of age or older. Although too early to tell, it would be expected that the distribution in age of publications would look similar to that shown in Figure 6.7. In the top 5% it is anticipated there should be less publications younger than one year than aged between one and two years old, with the same applying to the subsequent age brackets. Even a metric revealing a good number of more recent publications should still maintain the high rank of older, highly prestigious material, fitting the same age pattern as shown in Figure 6.7.

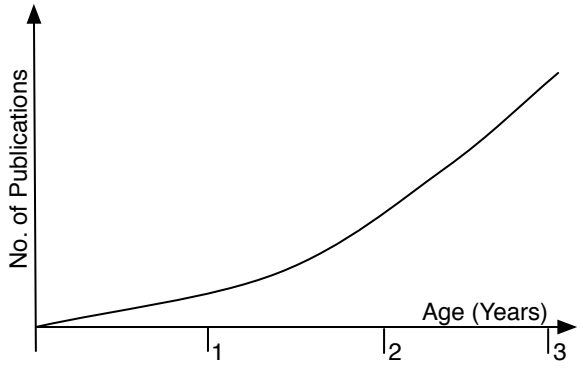

FIGURE 6.7: Expected distribution of publications by age (top 5%)

Table 6.2 gives a summary of the various tests outlined in this section which are designed to reflect those which were performed on the theoretical network.

Test Name	Description
Rank Reveal	Correlation between rank order of publications in current snapshot to target snapshot.
Mean Rank	Examine the mean rank and distribution of the publications in the snapshot.
Publication Age	What is the average age of a publication in the top 5% of each snapshot

TABLE 6.2: Summary of Test Strategy

6.5 Examining metrics on the Citebase Dataset

This section provides a metric-by-metric introduction to the tests carried out against a single target set of 100 publications. In each metric subsection the results of each test are graphed independently and a brief examination is given, outlining how the results compare to those observed against the artificial network in Chapter 5. In Section 6.6 the datasets being tested are expanded to include several sets of 100 target publications;

covering more of Citebase whilst also ratifying the results. Additionally in this section, the results are collated and plotted as a single result for each test, allowing closer comparison between the different metrics.

Throughout all tests, the target set of publications is selected from the set of results generated by the Citation Count metric. Additionally this metric is also the first one considered in this section. Although it may seem illogical to compare Citation Count to itself, it is always helpful to have a scientific control whenever possible.

6.5.1 Citation Count

With Citation Count acting as the control, it is possible to ensure that the metrics processor and results analyser parts of the Co-Ordinator are performing correctly. The correlation of Citation Count to itself must exhibit a perfect positive correlation after 36 months; the final snapshots from A and B should represent exactly the same data.

Beyond the correlation analysis, this test also allows the obtaining of a figure for where these 100 publications are ranked (by Citation Count) within all the other publications. This percentile figure will give an indication of the average rank of these 100 compared to the others in the entire dataset throughout their life cycle. This is important as it will demonstrate the overall standing of these publications amongst all the other publications.

Additionally it is useful to see if early Citation Count is in fact just an early indication of itself, i.e. will a publication, which begins its life more highly cited, remain so.

Figure 6.8 shows the first correlation graph where the Spearman Correlation (to the target list) is plotted against the age of the publication. At three years old, the correlation is a perfect 1 as expected. After 12 months it is already possible to observe a strong 0.6 correlation, while after 24 months this grows further to correlation greater than 0.8.

Figure 6.8 demonstrates that Citation Count is a good early indicator for itself when tested on a large dataset of peer reviewed publications. After 12 months there is a 0.65 correlation between the rank order, of the selected 100 and the target 100. This could simply mean that the publications are starting to obtain their rank order but are still of low rank compared to the other publications in the dataset, suggesting these publications are still unlikely to be found in a search.

To analyse the average rank for this set of 100 publications within the entire dataset requires the actual rank positions of each of the publications to be found. For example, publication #1253 might have a rank 2,146 and publication #1287 a rank position of 3,200, giving an average rank position of 2,673. From this average position, the percentile position within the entire dataset can also be calculated and plotted. On such a plot, the lower the percentile the higher the rank. Figure 6.9 shows the average rank percentile,

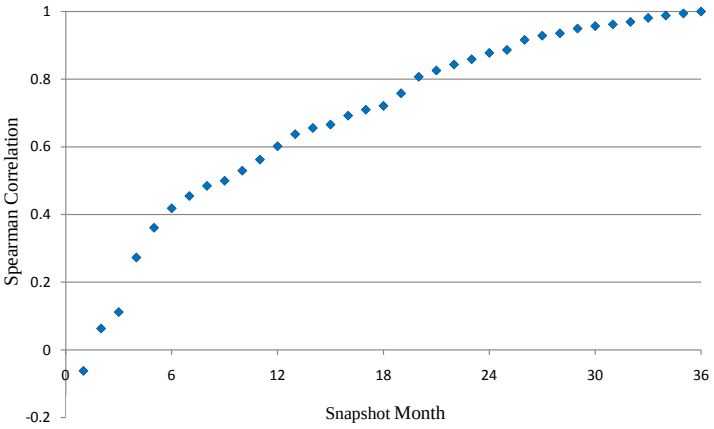

FIGURE 6.8: Citation Count correlation results per snapshot to target publications

by Citation Count, for the target set of publications. The average after three years places these 100 publications in the top 2.5% of all publications listed in Citebase.

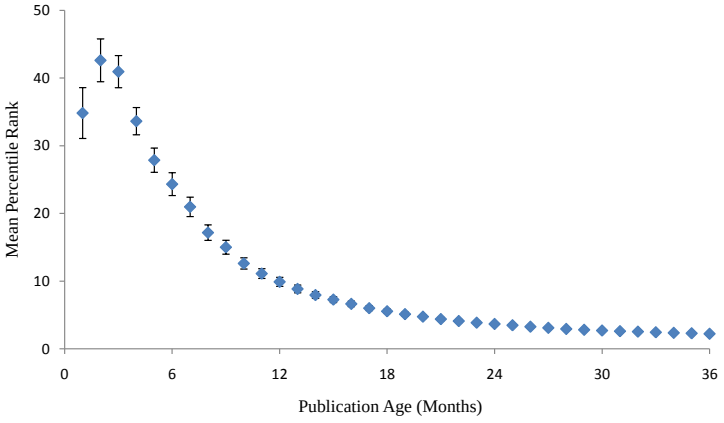

FIGURE 6.9: Percentile mean rank of target publications in each snapshot (Citation Count)

Figure 6.9 shows the average percentile (and deviation) for the target set of publications within each snapshot. After three years, the average ranking for the target set lies in the top 2.5% of all publications. This is a high aspiration for any other algorithms, although overall this still represents the 7,845th publication. The average deviation each month is a fraction of the percent (377 rank positions) with a high rank position of 42 and a low of 14,140.

Further evidence that Citation Count is a good early indicator comes from the fact that after 12 months, the mean publications rank lies in the 10.6th percentile — the average rank here is 24,578, and the deviation just over 1% (2,688 rank positions).

Lastly, the age of the top 5% of publications in each dataset is required to examine overall behaviour of the Citation Count ranking. As Citation Count is not based on a weighted algorithm it is logical to say that the most highly cited papers are likely to be the oldest, as these have had the most time to gain citations. Anomalies to this are only likely to exist in the case of a remarkable new publication.

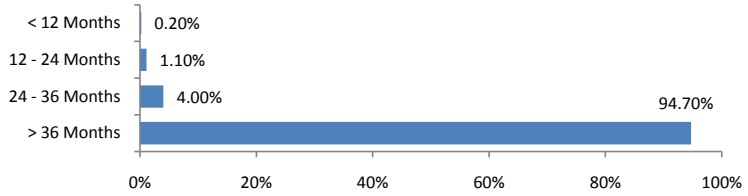

FIGURE 6.10: Average age of top 5% of publications in each snapshot (Citation Count)

Figure 6.10 shows the percentage breakdown of age for the top 5% of all publications contained in each snapshot. This data represents a much broader selection of publications than just the 100 used in the other comparisons. As expected, this result shows that almost 95% of the publications are over three years old. Out of the 430,000 (approximate) publications in the top 5%, only 860 are less than 12 months old. To be in the top 5% only 12 months after publication is a surprising result, but may be due to inaccurate input data or delayed publication where citations have already been accrued. This situation could be explained by a pre-print or other Open Access version being released prior to official publication and index in Citebase, this theory would also fit well with the findings of Eysenbach (2006).

Citation Count can be said to be an accurate indicator of itself, correlating in a near linear fashion over the 36 months to the target rank. What is surprising is that after only six months, there is a 0.5 correlation in rank order and average rank in the 25th percentile. Although this sounds like it breaks the typical publication life cycle model, which states that typically it takes nine months to obtain the first citation, this study is only looking at a set of 100 high ranking publications. This average would have been heavily influenced by the long tail of publications which only ever achieve one or two, or even zero, citations.

With articles typically reaching peak citation rate around three years after publication according to Moed (2005). This hypothesis certainly seems to hold in this study of Citation Count with the average age of a high ranking paper being over three years.

6.5.2 HITS - Authorities

In this section, the papers of most interest are those with a high Authority score. In the artificial network the HITS algorithm (Section 5.5.2) identified a few of the newer, possible summary publications that contributed to a greater ranking of publications which

ranked lower by Citation Count. If this pattern is to emerge on the Citebase dataset then a less positive correlation is expected, however the mean rank and publication age should remain strong for each set of publications.

As outlined in Section 6.3, 40 iterations were performed to calculate both the Hub and Authority scores. At the end of each iteration (Hub or Authority) the results were normalised such that the sum of all publication scores was one. As with all the other tests a precision of 31 decimal points was used to record the resultant decimal scores (32 bit decimal with 1 digit before the decimal place).

Figure 6.11 demonstrates a clear lack of correlation between the target dataset and any of the snapshots. The highest correlation value here is 0.16, obtained in the 14th month. While this figure is surprisingly low it may go some way to explaining the overall performance in the artificial test network earlier in Figure 5.10, where the behaviour was similar to the CoRank algorithm.

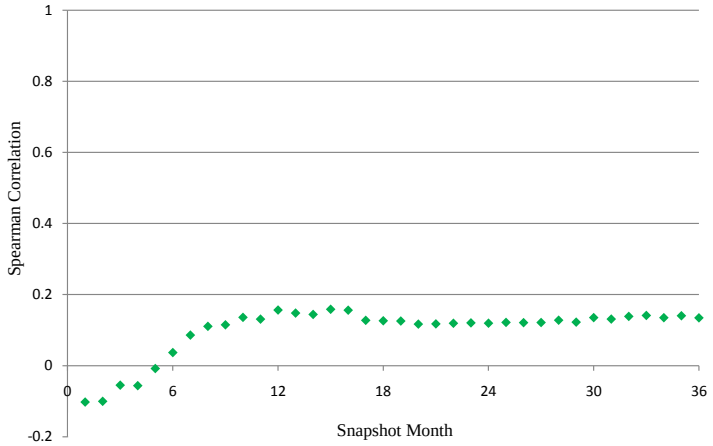

FIGURE 6.11: Authority Score correlation results per snapshot to target publications

In the other two tests, percentile rank (Figure 6.12) and publication age (Figure 6.13) things are slightly more predictable in behaviour. While the rank order correlation does not appear very close to the desired target, the mean rank of the publications still climbs quickly and breaks the top 10% after 21 months. After 36 months the mean rank lies in the 8th percentile with a deviation of just over 1%, giving a high rank of 30 and a low of 120,563.

A similarly predictable result can also be seen by publication age, shown in Figure 6.13. Here nearly 88% of the publications are over three years old and while the Authority score does reveal some newer publication in the top 5%, this change is not vastly different from Citation Count.

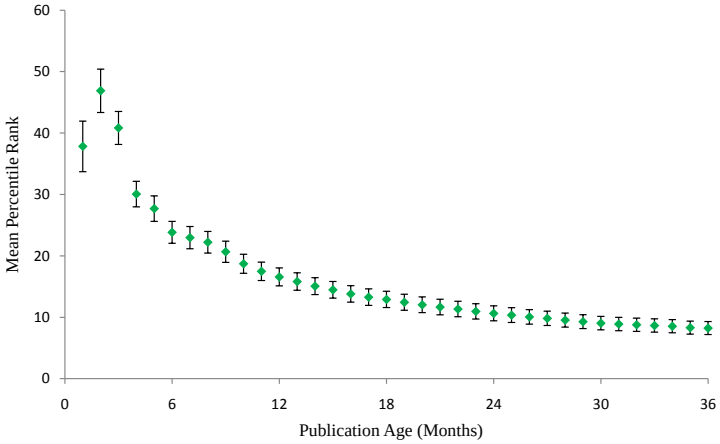

FIGURE 6.12: Percentile mean rank of target publications in each snapshot (Authority Score)

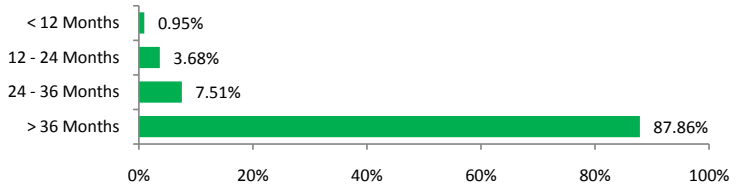

FIGURE 6.13: Average age of top 5% of publications in each snapshot (Authority Score)

6.5.3 PageRank

PageRank (Brin et al. 1998) is the first algorithm designed specifically to handle the masses of data on the Web and utilises the weighted citation idea. When used alongside text matching techniques, PageRank is able to effectively locate and rank the top websites which are relevant to a search term. Since this study is applying PageRank to a closed set of peer review literature, it is likely to behave differently to when applied to a Web graph. Additionally it is possible to cover all publications during each iteration.

Figure 6.14 demonstrates that PageRank, like Citation Count, obtains a 0.5 correlation to the target set of publications within 12 months. Unlike Citation Count, this correlation remains fairly constant and does not get improve much over time.

As with the HITS algorithm, 40 full iterations over every publication in the dataset were performed. As these were full iterations the damping factor has little effect on the result. Translating this to the number of publications and resources on the Web, would lead to the random surfer paradigm having a greater overall effect, as PageRank wouldn't follow every link between Web pages.

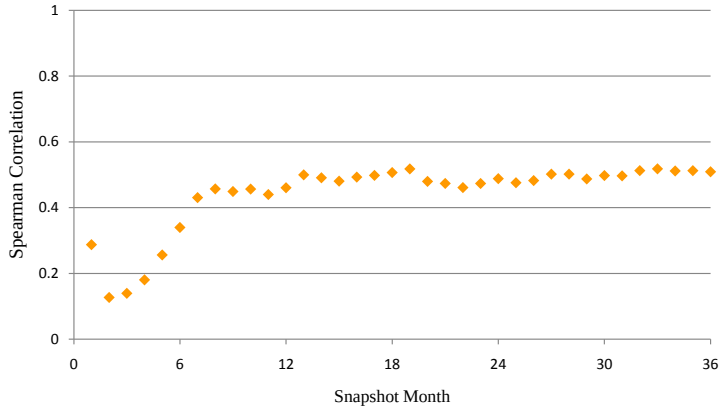

FIGURE 6.14: PageRank correlation results per snapshot to target publications

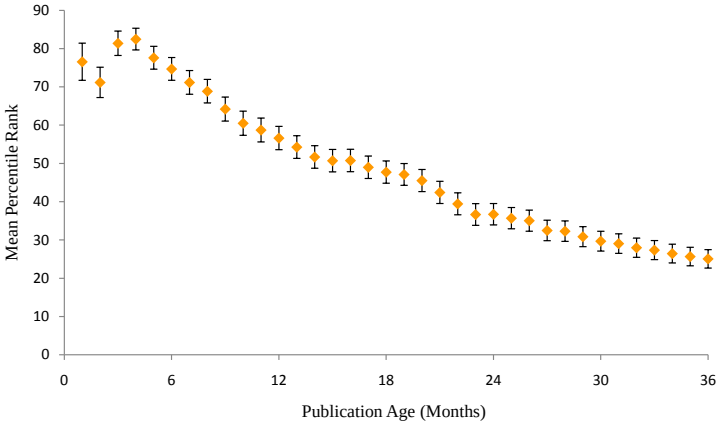

FIGURE 6.15: Percentile mean rank of target publications in each snapshot (PageRank)

Figure 6.15 shows a much steadier increase in mean rank for the target publications over the 36 months. With the highest rank percentile (24.9%) occurring after the full 36 months and a deviation of 2.4% (7,350 rank positions), this makes PageRank the worst overall algorithm so far, even though it is closer in rank order than HITS. This result means that PageRank is unlikely to rank these publications on the first page of results if the entire network of publications was indexed.

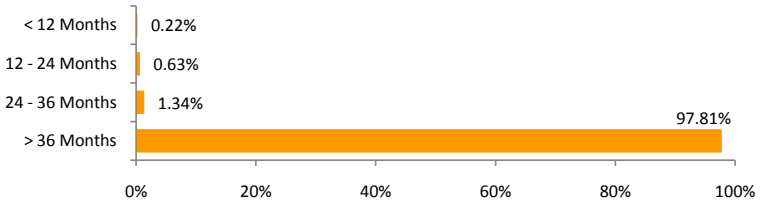

FIGURE 6.16: Average age of top 5% of publications in each snapshot (PageRank)

Looking at the average age of publications (Figure 6.16), shows a similar result to that by Citation Count; the majority of publications in the top 5% of each dataset older than three years. This result, combined with the earlier mean rank graph could imply that PageRank might in fact take longer than Citation Count to give publications their highest impact score.

With PageRank being so highly respected on the World Wide Web as a ranking algorithm, it is insightful to see how it actually performs in a network of fully peer reviewed publications. Surprisingly the correlation to Citation Count is low peaking at a 0.4 correlation level. On the Web this would not be surprising due to the ease at which links can be polluted with false citations, however in a fully peer reviewed network this should be less of the case.

The thing to remember is the impact of the weighting factor on the algorithm, perhaps the top one hundred by Citation Count are highly cited, but highly cited by a lot of low impact publications. Iteratively this is why PageRank has not highly regarded these publications. This would also be logical in a network with a long tail of publications which are only cited once or twice. On the Web it is likely that PageRank will not be indexing the newly established websites, due to either not discovering them yet or not following enough links to get to them due to the random surfer factor.

6.5.4 CoRank

In Chapter 5, it was found that PageRank held the closest relation to the behaviour of Citation Count, with HITS and CoRank showing a similar behaviour. Figure 6.17 confirms this similarity showing that CoRank holds little correlation at any point in the publication life cycle to the target dataset ordered by Citation Count.

During the application of CoRank on the artificial network (shown in Figure 5.10) a small gap was witnessed between the correlations of HITS and CoRank compared to the other algorithms. Figure 6.17 and 6.11 both suggest that this gap has widened, with both showing very small positive correlations compared to the other algorithms.

From the definitions of "better" given earlier in Chapter 5 (Section 5.1), it would appear from Figure 6.17 that CoRank is not going to be able to reveal the core community of publications at any point during this three year life cycle, in the same rank order. So far, PageRank has been the best alternative to Citation Count by Spearman Correlation.

Looking at the mean percentile in which the target set of publications resides, reveals a distinct difference to the result obtained via PageRank. Figure 6.18 shows that by using CoRank, the target set of publications are almost instantly all within the top 40% of all publications in the dataset.

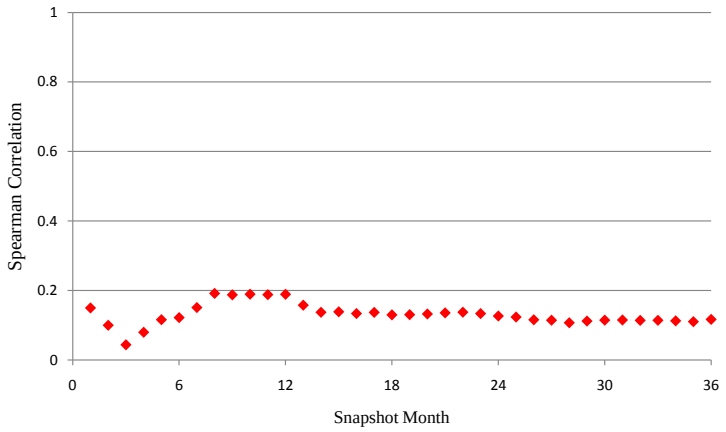

FIGURE 6.17: CoRank correlation results per snapshot to target publications

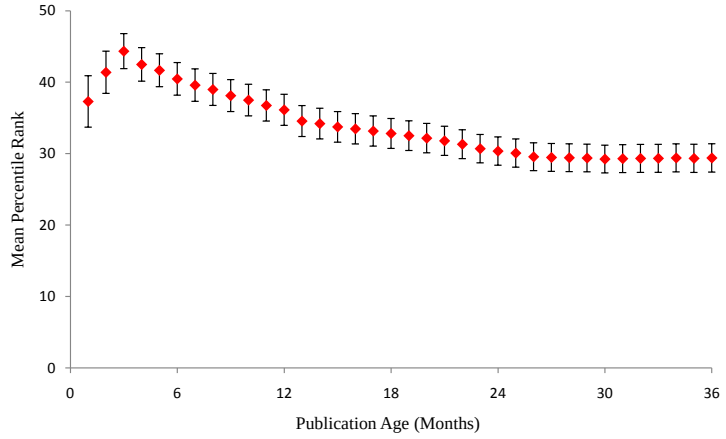

FIGURE 6.18: Percentile mean rank of target publications in each snapshot (CoRank)

After just 16 months, the mean percentile rank for the target publications dips under 30% and remains there. While this is consistent with the other algorithms, CoRank shows a slightly different behaviour which cannot be seen as clearly in Figure 6.18. In all other algorithms the mean rank (not percentile rank) lowers with each month, while in CoRank the lowest set of mean ranks are achieved between months 12 and 24 (Figure 6.19).

Remembering that the mean rank percentile is also related to the number of publications in the system, which is increasing in a roughly linear fashion, CoRank maintains a high ranking for our target set regardless of this fact. In the next section, each algorithm is analysed to find how it performs across a much broader sample set and hopefully show how this pattern is consistent no matter which set of publications are chosen to trace.

Lastly in this section, Figure 6.20 shows that almost 20% of the top 100 papers by CoRank are newer than three years old. Interestingly these publications appear to be

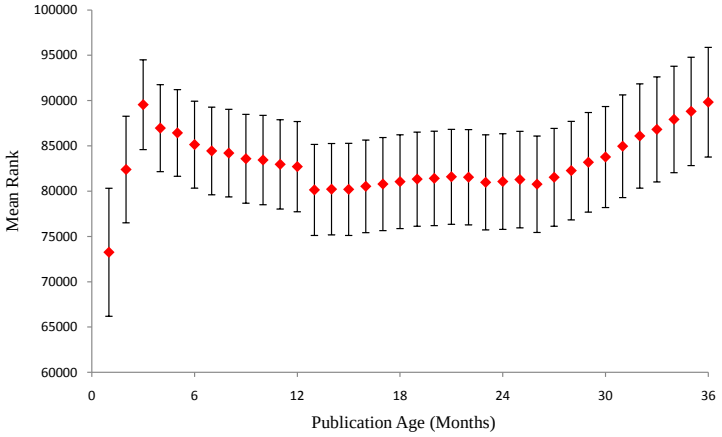

FIGURE 6.19: Actual mean rank of target publications in each snapshot (CoRank)

distributed evenly between each of the three groups. This does not tie in very well at this stage with the findings for the target 100 as shown in Figure 6.19 where the peak percentile is reached between 12 and 24 months in age.

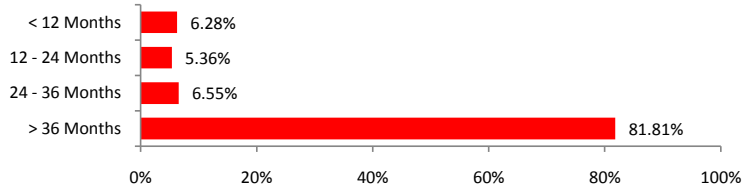

FIGURE 6.20: Average age of top 5% of publications in each snapshot (CoRank)

6.6 Comparing Results

With real results obtained from the Citebase dataset, it is positive to see that a number of the patterns that emerged during examination on the artificial network also emerge when examining a genuine publication network. The clearest example of this is shown by CoRank, including 6% newer publications than its nearest rival (HITS) and over 12% more than both Citation Count and PageRank.

While HITS seems to perform similarly to CoRank, the mean percentile rank for the target dataset is lower after three years, showing similarities to Citation Count.

Inversely, while the order of publications in the dataset holds a better correlation to Citation Count, the mean rank percentile takes over two years to fall below the 40th percentile compared to CoRank which achieves this in half the time.

At this point it would appear that CoRank sits between Citation Count, HITS and PageRank in terms of performance. This is the case for all aspects except publication age. Here CoRank is including a greater number of newer publications in its top 5% than all other algorithms, which is encouraging and follows the pattern observed when applying CoRank to the artificial network earlier in Chapter 5.

With initial testing complete, it is necessary to confirm the results by studying a much larger set of publications. As only a single set of 100 publications has been sampled up to this point, sourced from a single target month, this represented a very small view on the entire Citebase dataset. Expanding the input data sample can either be achieved by selecting a different set of 100 publications from the same snapshot and same time period, or by creating a greater number of snapshots so that different periods of three years can be studied. It was the latter option which was chosen here, thus a further 12 months of snapshots were generated, taking the dataset coverage back an additional year in time. Previously, where the target snapshot was taken in March 2007, the three years would start in March 2004, now with the target snapshot taken in January 2007, the correlation readings would be taken between January 2004 and January 2007.

In this section it is only necessary to repeat the tests relating to rank correlation and rank position as those pertaining to publication age in the previous section were sampled across the entire dataset.

Figure 6.21 shows the average rank correlations for the 12 datasets obtained by each algorithm. Here sets of 100 publications are still tracked through their three year life cycle and Figure 6.21 shows all 12 sets of results plotted as averages. Thus the age remains represented in months from a now abstracted start date.

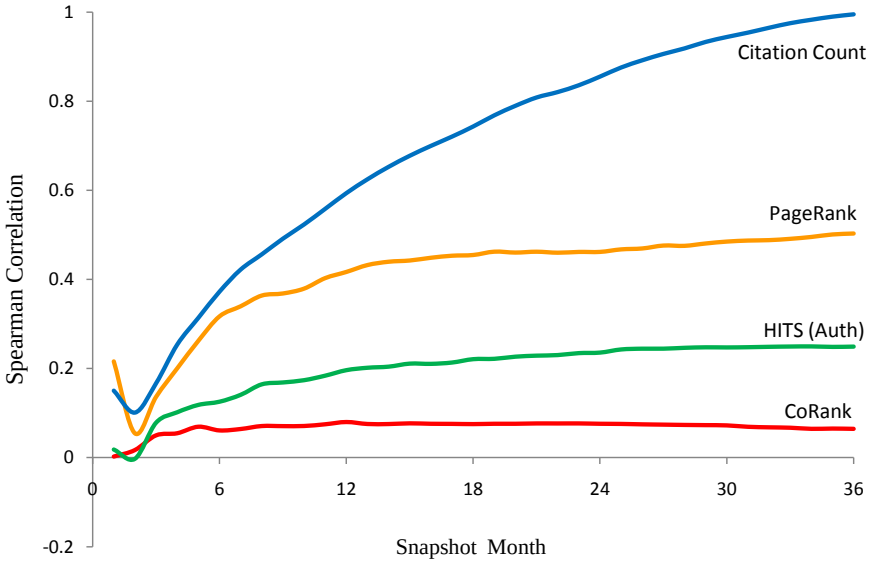

FIGURE 6.21: Combined correlation for 12 sets of results plotted against publication age

Figure 6.21 shows the same correlation trends as from the single set of 100 publications, affirming that a single dataset gave accurate results. This is especially the case with PageRank, where the plot line and average is almost exactly the same. The deviation of results for each individual algorithm from the average also tended to be very small, showing the consistent performance on the dataset.

Correlating this with the average rank results for CoRank, PageRank and Citation Count shows the key observation that, on average, a publication reaches its peak rank position between 12 and 24 months. PageRank and Citation Count both show that a publications rank is still climbing after this three year period, not surprising for Citation Count, but with PageRank being a weighted algorithm, this implies that many of the citations received are similar in contributing score. This could be easily explained by the Power Law of citations observed in Figure 6.1. With the Citebase data only containing publication data about peer reviewed publications, this observation would hold.

Figure 6.22 looks at the average percentile rank of the 100 publications across the 12 result sets. This data is almost exactly the same as given earlier by a single dataset.

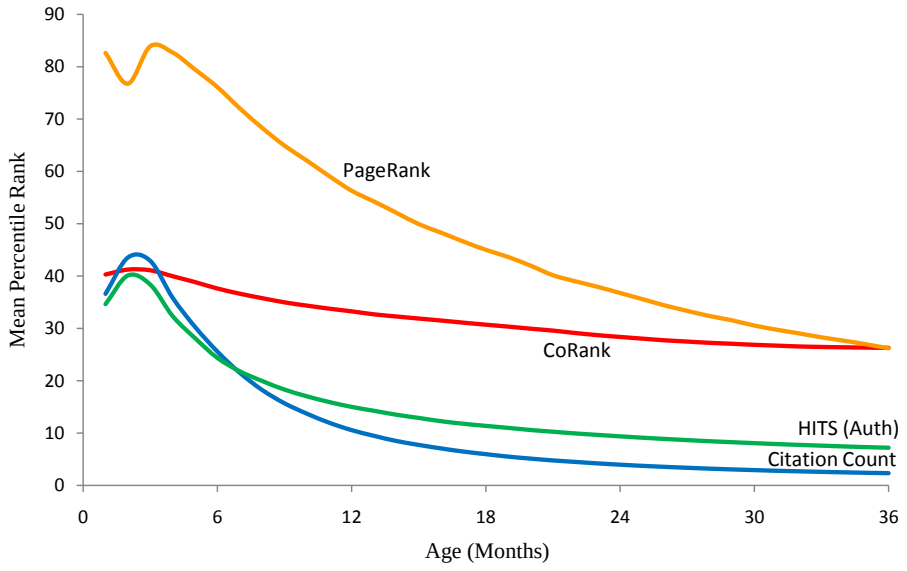

FIGURE 6.22: Percentile mean rank in each dataset against publication age

Figure 6.22 shows clearly how the percentile rank in which the publications reside by PageRank is not as stable as the other algorithms even after 36 months. CoRank almost does the opposite and stabilises very quickly, which the remaining two algorithms trend together towards very low percentile rank. Thus by this test, HITS and Citation Count perform best, with PageRank proving itself to not be a very good early indicator when the whole dataset of publications is considered.

To reinforce a trend noticed earlier in Section 6.5.4, Figure 6.23 shows the actual average rank figure for all 100 publications across the 12 snapshots. Note that deviations have

been disregarded from this graph as they remained approximately the same as plotted previously. In the mean rank test, CoRank is the only algorithm which ranks publications at their highest point during the three year period, rather than at the end of this same itself. This high ranking is achieved when the publications are between 12 and 24 months old and may suggest that CoRank gives its best indication of impact between these points, this is also partly reflected by the correlation result from Figure 6.21.

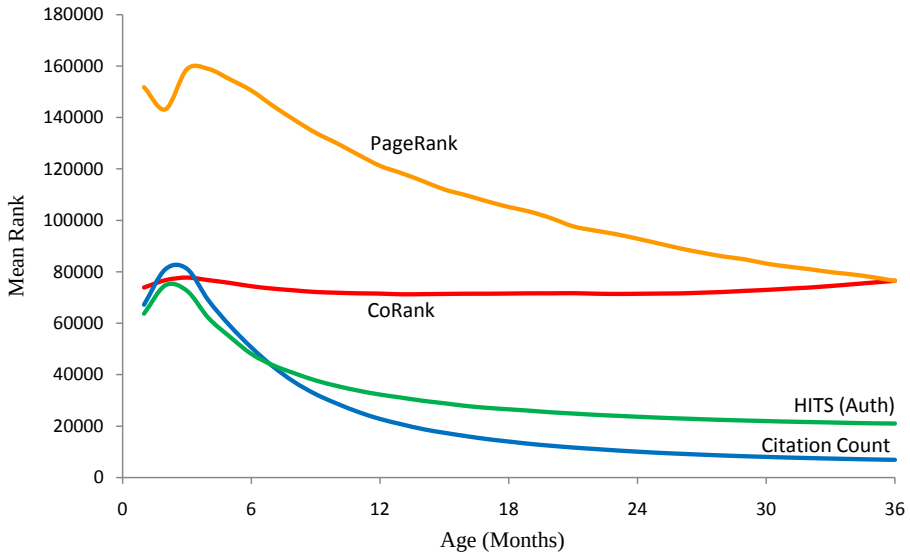

FIGURE 6.23: Actual rank for 12 sets of results by CoRank plotted against publication life cycle

6.7 Summary

This chapter has shown a mixture of positive and negative results. While CoRank has not been shown to be the ideal algorithm to replace Citation Count, the characteristics compared to PageRank and HITS lead to many interesting conclusions. In the test network it was demonstrated that CoRank might perform relatively similarly to Citation Count. The same similarity also existed for PageRank and the HITS algorithm. The correlations were so close between the algorithms that there was not much basis on which to choose between them. The same applied to the publication ages, although CoRank did show signs at this stage of revealing more recent publications existing within the network.

When applied to a real dataset, the small differences between each of the algorithms became clear. This is particularly the case with the comparison of the correlation between the rank order of the publications in question. Even after the full three years, the closest correlation was PageRank at 0.5. At this point if CoRank had been the only

algorithm to be distant in correlation, then the only possible conclusion would have been that CoRank in its current form is not highly useful when compared to Citation Count.

There are some conclusions which can be drawn here which help to explain this performance. The clearest comes from the values for the percentile rank of the publications within the entire dataset (Figure 6.22), here it was observed that CoRank appears to stabilise quickest, while PageRank cannot be said to be anywhere near stabilisation after three years. To analyse this further requires additional examination of the characteristics of the top 100 publications, namely the citation and co-Citation Count, and compare these to the typical distribution of citations within Citebase.

Figure 6.24 shows the citation growth on average for 12 sets of 100 publications over a three year period. As is evident, with the average number of citations after one month being just over four, this already puts all of the publications within the top 40% of all publications (from Figure 6.1) by Citation Count. This fact is shown by Figure 6.22 for all algorithms except PageRank, where citations are evidently not coming from high enough impact publications to cause this pattern to emerge.

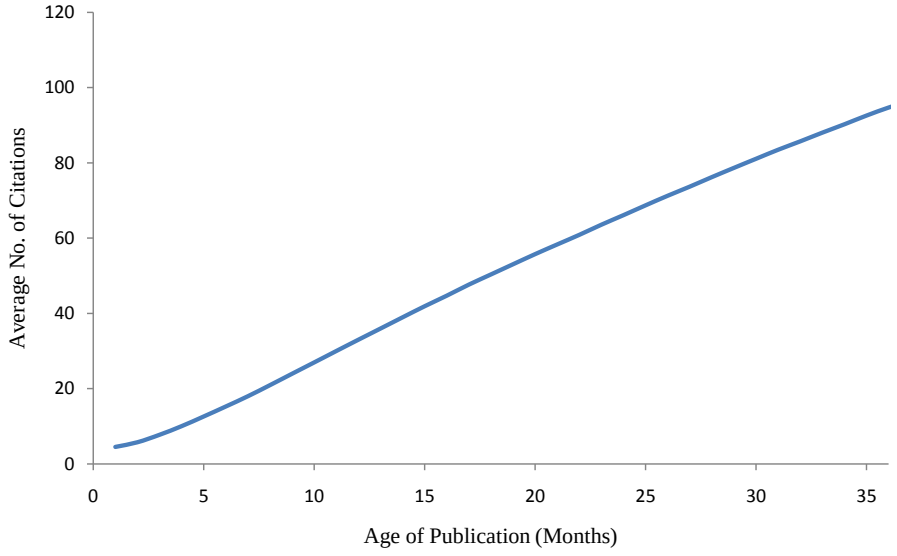

FIGURE 6.24: Average Citation Count for 12 sets of top 100 Publications over three years

When looking at the distribution of co-citations over the entire dataset, it would be logical to expect to see the same Power Law distribution seen earlier in Figure 6.1. As well as showing that the co-citation distribution maps to the Power Law, it is also possible to observe the growth in network size resulting from looking at co-citations rather than citations. Figure 6.25 shows that there is now a much broader distribution, where previously approximately $1/5th$ of the publications in Citebase have only one citation, $1/5th$ of the dataset now covers between 1 and 10 co-citations. Additionally, more publications are co-cited with two other items than with a single other item.

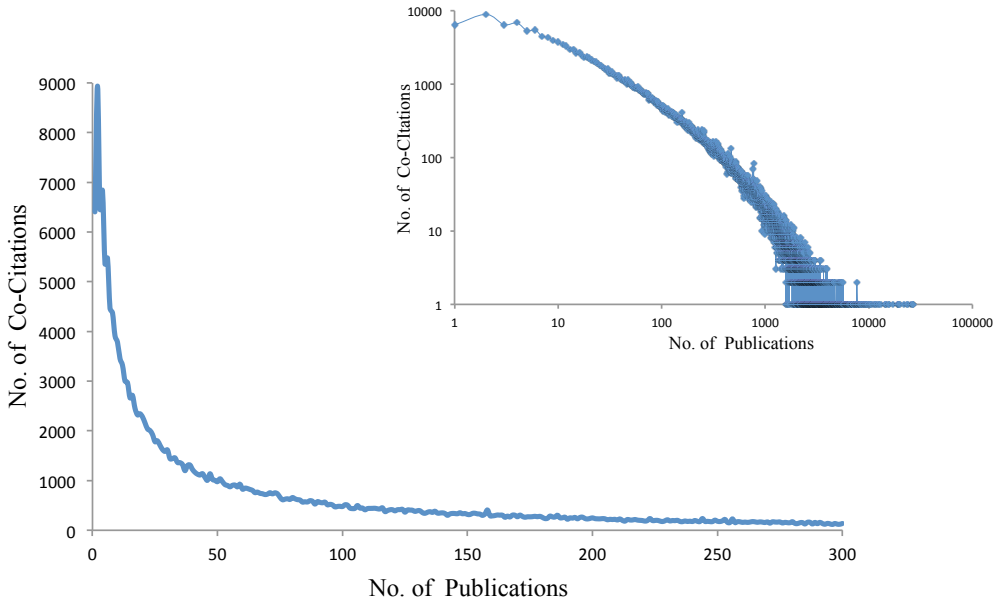

FIGURE 6.25: Distribution of Co-Citations in Citebase

The range of values has become much broader as the network of co-citations is much larger, thus the top cited publication with 4,209 citations has 23,758 co-citations. Interestingly this is not the most number of co-citations. The largest number of co-citations are related with the 2nd most cited publication, in this case 3,631 citations produce a network of 26,733 co-citations. Comparing this figure to that from Figure 6.1 from earlier shows the clear linear relationship between citation distribution and co-citation distribution.

Likewise by looking at the sets of 100 publications, it is expected that the number of co-citations will grow at a rate which correlates with the number of citations. This is shown in Figure 6.26, where after three years, the publications are amounting around 1,400 co-citations. As with Citation Count this puts them in the top 2.5% of all publications by co-citation count.

So theoretically, if the number of co-citations was counted and used as a ranking mechanism, this should map almost exactly to the Citation Count results. However, since both PageRank and CoRank look at weighted algorithms the performance is, predictably, not the same. Since all publications cannot hold an equal weighting when using CoRank and PageRank, then some publications have to be scored poorly and the publications which are scored worst are going to be those which are un-cited, including those that have been added recently. Logically, a PageRank score for a publication will not stabilise until all of the PageRank scores for the citing publications also stabilise, doubling the time required for these publications to reach a high rank value.

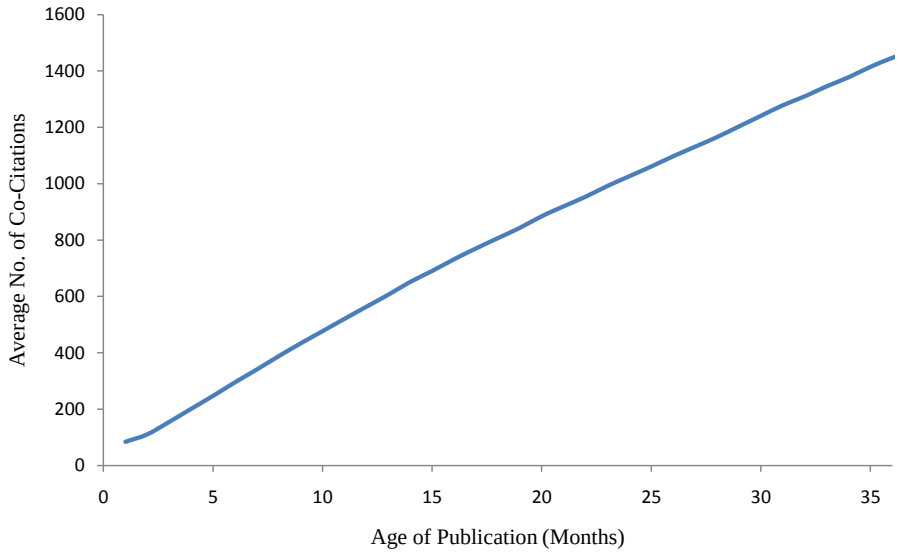

FIGURE 6.26: Average Co-Citation Count for 12 sets of top 100 Publications over three years

Co-citations map much closer to the Web paradigm, here each co-citation received can relate a single publication to a number of publications of a significantly older age than that of the directly citing publications. This does have the effect that the publications are almost instantly rated higher within the entire corpus of publications. However the correlation to Citation Count, although positive, does not reflect that a similar behaviour exists.

Further evidence that PageRank rates older publications highly comes from the observation that PageRank contains the highest amount of publications older than three years in the top 5%. Figure 6.27 shows the combined graph for the ages of the publication in the top 5% of all publications ranked by each algorithm.

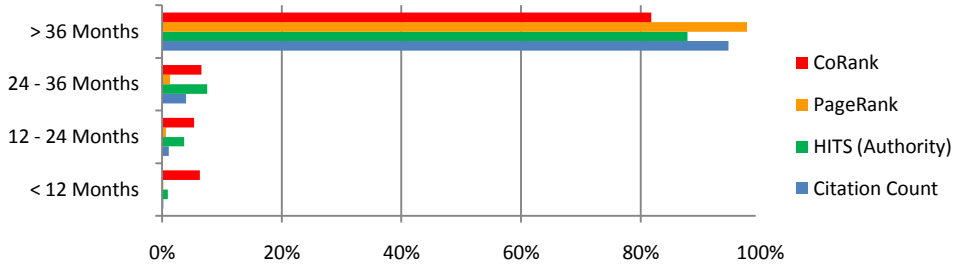

FIGURE 6.27: Combined age comparison of top 5% of publications by four metrics

Positively, CoRank does manage to reveal a number of newer publications, however when looking for a Power Law style curve to exist here where each younger category contains progressively less publications, then HITS (Authority) comes out best.

Looking a little further reveals that the top 5% of publications by CoRank only amass, on average, 335 co-relations, which by Co-Citation Count means that these top 5% of publications are in fact in the 70th percentile of all publications.

Chapter 7

Refining CoRank

The previous chapter introduced and examined a number of different metrics, demonstrating the variety of characteristics each utilises in order to output rank results. Due to a number of clear differences between these metrics, none have yet provided a balance of all the examination criteria set out by the Citation Count metric. These criteria specify the requirement for any new metric to exhibit high correlation to Citation Count, with the added ability of being capable of revealing more recent, potentially high impact publications.

This chapter introduces a number of further metrics which either refine or extend the CoRank algorithm. The aim remains the same, of finding an algorithm which is able to display a similar behaviour to Citation Count earlier in the publication life cycle. Each of the algorithms also maintains focus on the co-citation data, with a firm belief that an early indication metric can only be found as a result of having more data to process. The only substantial variation to CoRank is caused by decision to change the weighting factor, or remove it, in order to better relate CoRank to Citation Count, without removing the advantages provided by the co-citation network.

In Chapter 6 it was observed that the best correlation by a weighted algorithm to Citation Count was displayed by PageRank (0.5 Spearman Correlation Coefficient). Although this represented the best correlation out of all the algorithms, other than Citation Count itself, both PageRank and CoRank demonstrated that weighted algorithms also do not expose high ranking publications (by Citation Count) within the top percentiles of all publications at any stage. Logically this is due to each algorithm looking at different age groups of publications; PageRank ranks highly a good number of much older publications while CoRank does the opposite. Revealing a number of more recent publications represents one good aspect for CoRank, however these cannot be said to be high ranking publications at any point in their life cycle, so refinement is required.

A total of six different variations of the CoRank algorithm are outlined in this chapter, each designed to accommodate a number of factors and behavioural trends. As each

is more or less based upon CoRank, each is named CoRank-*Variation*, where CoRank implies it is based upon the Co-Citation data and the *Variation* outlines the adjustment. Staring with CoRank-LinkCount, a non-weighted variation of CoRank, the variations then build to the point where age of citations and co-citations is considered. Thus the most complex of the six metrics introduced examines if the most recent information pertaining to a publication is the most relevant.

As in the previous chapter, each algorithm is applied to a number of datasets before comparisons are made to Citation Count. With the Co-Ordinator in place, the application of a whole family of metrics becomes a simple and well managed process, with output formatted in a manner such that further analysis in spreadsheet applications is also easily possible.

With six algorithms outlined in this chapter, each is introduced briefly to explain why each variation has been chosen before the results of the percentile age test are presented. For clarity the comparison of Spearman Correlation and percentile rank is summarised after all six algorithms are introduced. This allows the results for all six algorithms to be summarised alongside the existing four from the previous chapter, making a total of 10 algorithms.

With ten algorithms being considered, Sections 7.2, 7.3 and 7.4 present the main body of this thesis, compare at how the 10 different metrics perform against the three hypothesis respectively:

- Co-Citation relations can be used to create an early indication metric for publication impact which correlates well with existing metrics.
- A metric based on co-citations will identify high impact publications sooner in their lifecycle.
- Applying co-citation metrics in search ranking will promote more recent publications.

Although a metric is not found which ideally fits all three hypotheses, one does show promise against the last two hypotheses without disgracing itself compared to the first. By removing the weighting factor, CoRank-LinkCount by design becomes much more similar to Citation Count than other metrics presented. By computing the Spearman Correlation between all of the algorithms and not just the correlation to Citation Count, allows initial analysis on which algorithms can be grouped by common characteristics, a theme which is carried throughout this chapter.

In order to ratify the significance of results, in addition to the thorough cross sampling applied throughout, a statistical significance test is required. This significance test should simply re-enforce that the methodology being used is sound, however it may also

reveal a non-sound algorithm. Usually a statistical significance test is used to ensure that the sampling of results is taken from a wide enough range, such that the result can be said to be accurate. Due to the variety of previously untested algorithms, a statistical significance test may reveal an algorithm which is unable to compute a stable result. Section 7.6.1 outlines the pros and cons of statistical significance testing further including why it is applied at this stage.

The final sections in this chapter examine the overall coverage of the algorithms, summarising findings from all tests in order to identify if any metric is proving itself as an early indication metric. All of these observations lead to the final section in this chapter which examines how the 10 algorithms can be divided into three groups based upon particular characteristics and their results. It is these groupings and coverages which are then looked at further in the final chapter of this thesis.

7.1 The Variations of CoRank

Following the successful application and interesting findings resulting from applying the CoRank algorithm, it is clear that there is opportunity for further refinement of the idea in order to perhaps find an algorithm which matches more closely the behaviour of Citation Count. Further, there is also opportunity to examine other factors which may lead to a new behaviour and potentially an early indication metric.

This section presents results obtained by adding six additional variations of the CoRank algorithm to the Co-Ordinator and performing all of the same tests as summarised in Table 6.2. Each algorithm has been refined in a number of ways, including one which no longer requires iterative processing due to the loss of the weighting factor.

Table 7.1 presents an overview of the six algorithms presented in this section, listing the primary factors in each algorithm and also the weighting factor. Due to all the weighting factors being dependant on other weighted factors in the system, this means that iteration is always required for these types of metrics.

Algorithm Name	Primary Factor	Weighting Factor
CoRank-LinkCount	Number of citations obtained by co-cited publications	None
CoRank-Divided	The CoRank algorithm	CoRank
CoRank-Scaled-LinkCount	Citation Count	CoRank
CoRank-Scaled-CoRank	CoRank	CoRank
CoRank-CiteTime	Age of the Citation	CoRank
CoRank-CoTime	Age of the Co-Cited Publications	CoRank

TABLE 7.1: Summary of the six variations of CoRank

As an example of how to interpret Table 7.1, take CoRank-CiteTime. This algorithm applies the normal CoRank algorithm and weights the contribution from each Co-Cited

publication by how old the direct citation is. Thus the age of direct citation is the main factor affecting the results (and constant in each sample), while the CoRank algorithm maintains the weighting. Each algorithm is explained in much greater detail during the following analysis sections.

7.1.1 CoRank-LinkCount

CoRank-LinkCount is an algorithm without a weighting factor, meaning that iterative calculation is not required. CoRank-LinkCount looks solely at the number of citations which are received, not by the publication in question, but by all the publications with which the subject paper is co-cited. This idea was introduced briefly at the end of Chapter 5 while examining the distribution of co-citations over the Citebase dataset, here it was found that the paper with the most co-citations was ranked 2nd by Citation Count.

Figure 7.1 shows how the sum of co-citing publications is worked out. Publication n represents the target publication and each p being the co-cited publication; each publication marked c is thus a directly citing publication to n. The total CoRank-LinkCount is the sum of all citations towards those papers (p) a publication (n) is co-cited with, including duplicates and c itself. In Figure 7.1 this is shown as all the papers $c(p)$ as well as the directly citing paper c which is also an instance of $c(p)$ as it also cites p as well as n. All together this gives the node (n) a CoRank-LinkCount score of 7 ($5c(p) + 2(c)$ as c cites 2 of the co-cited publications and is therefore counted twice).

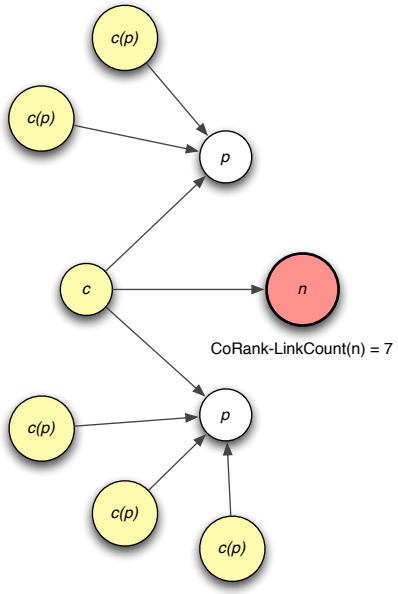

FIGURE 7.1: Network of Co-Cited publications from publication n

This algorithm is a simplification of the CoRank algorithm, removing the need for the iteration. As a consequence it also removes the weighting factor and ability to rule

out false positives. Shown by Equation 7.1, CoRank-LinkCount is an experiment to see if the total number of citations received by co-cited publications provide an indirect indication of impact. Equation 7.1 is summing up the LinkCount (LC) scores for all publications with which a paper n is co-cited, represented as p. With the relation to Citation Count re-established, it is expected that this algorithm will behave in a similar manner to Citation Count. The bigger network of co-citations is still utilised, thus there is some chance that an early indication of impact might be possible to judge.

$$CR_{LC}(n) = \sum_{p_j \in M(p_i)} LC\,(p_j) \tag{7.1}$$

Using the same datasets as in Chapter 5, Figure 7.2 shows that CoRank-LinkCount sits between Citation Count and HITS (Authority) in terms of publication age. While this is not as significant as CoRank it demonstrates that basing a publications impact on the citations of papers you are cited with is similar in nature to a plain citation count.

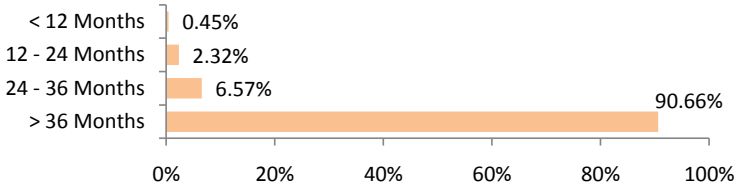

FIGURE 7.2: Average age of top 5% of publications in each snapshot (CoRank-LinkCount)

7.1.2 CoRank-Divided

This metric takes the CoRank score (given by the CoRank algorithm) for a publication and divides it by the number of co-cited papers that contributed their CoRank information. This would distribute rank information obtained from each co-cited paper both by the number of other co-citation relations (as per the original CoRank algorithm) and also by the number of co-cited papers that exist. Equation 7.2 shows this equation where the CoRank-Divided score for our publication (n) is calculated from the CoRank of this publication ($CR(n)$) divided by the number of publications with which n is co-cited ($CL(n)$).

$$CR_{Div}(n) = \frac{CR(n)}{CL(n)} \tag{7.2}$$

It should be fairly obvious to spot that this algorithm should create inaccurate results, as it goes against the methodology that impact starts low and steadily increases until

it stabilises. In the case of this algorithm an impact will start high as fewer co-citation Links ($CL(n)$) will result in a low denominator and thus a high impact score. This is reflected straight away in the graph of publication age (Figure 7.3), where there is a greater percentage of publications less than a year old than in the one to two and two to three year old periods.

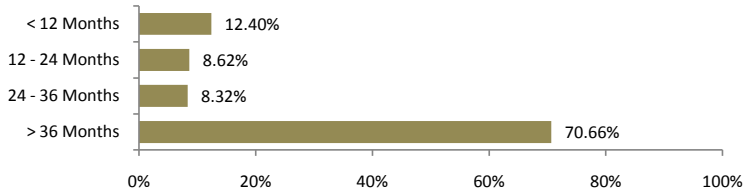

FIGURE 7.3: Average age of top 5% of publications in each snapshot (CoRank-Divided)

Although it may seem illogical to include this algorithm, as a control it should show better the relative performance of other algorithms in comparison to each other. It could also act as a differentiator between a badly performing algorithm and the ideal, so it will be useful to see how far each algorithm is adrift from CoRank-Divided when looking at the principal components of each in later sections.

7.1.3 CoRank-Scaled-LinkCount

In this variation of CoRank, shown by Equation 7.3, the CoRank score is multiplied at each stage by the number of citations towards the directly citing publication ($LC(c)$ where c is the directly citation publication as shown in Figure 7.1). The major factor in this algorithm becomes the Citation Count. By doing this, the more prestigious the directly citing paper (by Citation Count) the more weight the CoRank scores obtained via this publication hold. This full variation to the CoRank algorithm is shown in Equation 7.3 where the importance of inferring a directed network are very important. Here the LinkCount ($LC(c)$) is used to multiply the CoRank of the Co-Cited publications to n where the Co-Citation is established via the paper c. Thus the Co-Cited paper p in each case, has to be cited by the directly citing paper c.

$$CR_{SLC}(n) = \frac{1- \propto}{|V|} + \propto \sum_{c_i \in c} \left(LC\,(c) \times \frac{CR\,(p_c)}{CL\,(cp_c)} \right) \tag{7.3}$$

With this metric involving both data obtained from the co-citation network as well as prestige from the citation network, it is expected to behave in a similar manner to Citation Count. From Figure 7.4, it is encouraging to see that the relative age of publications in the top 5% is closer in nature to that obtained by the CoRank metric, than through Citation Count.

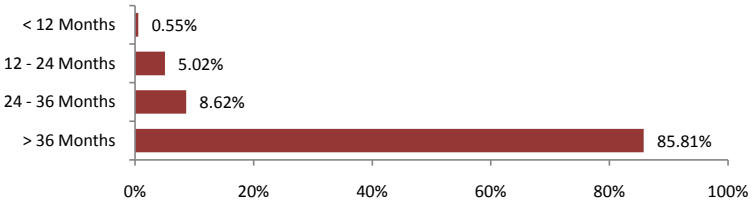

FIGURE 7.4: Average age of top 5% of publications in each snapshot (CoRank-Scaled-LinkCount)

7.1.4 CoRank-Scaled-CoRank

Much like CoRank-Scaled-LinkCount, this algorithm looks at the prestige of the directly citing paper. In this case the prestige score is given by the CoRank score of the citing paper ($CR(c)$). Instead of using the Citation Count a direct link is maintained to the CoRank algorithm across all parts of this algorithm.

Equation 7.4 shows this algorithm which, like CoRank-Scaled-LinkCount, is applied iteratively across the entire dataset.

$$CR_{SCR}(n) = \frac{1- \propto}{|V|} + \propto \sum_{c_i \in c} \left(CR\left(c\right) \times \frac{CR\left(p_c\right)}{CL\left(cp_c\right)} \right) \tag{7.4}$$

Due to the division by the CoRank score we would expect this algorithm to amplify the patterns seen in the CoRank algorithm, thus showing even more recent publications while at the same time performing worse against citation count. Figure 7.5 demonstrates that this first assumption holds.

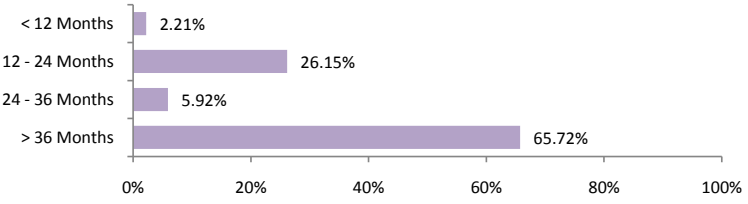

FIGURE 7.5: Average age of top 5% of publications in each snapshot (CoRank-Scaled-CoRank)

The reason to include this algorithm will become clearer when analysing its characteristics using Principal Component Analysis in Chapter 8.

7.1.5 CoRank-CiteTime

CoRank-CiteTime is the first of two algorithms which introduces a completely new type of weighting to the CoRank algorithm, one based upon temporal factors. In a related study, Maslov & Redner (2008) observe that citations could not be updated after publication, which makes the affects of ageing much more important in citation networks than on the Web. CoRank-CiteTime and CoRank-CoTime are designed to investigate the application of age factors into metrics designed to process citation data.

CoRank-CiteTime examines the most recent data, relating to the most recent citation, to see if this gives a better indication of impact. The theory is based upon a hypothesis of human behaviour, that of following by example. Here once a publication has established itself with a few reasonably high profile citations, it follows that others will find these citations and follow them in order to also cite the publication in question.

Thus the assumption is that the most recent citation network should reflect the publication's overall impact. It also follows that this may not have any immediate effect as it may take a while for a publication to find its impact niche, affecting early results significantly.

With the first iteration of this algorithm it is the age of **directly** citing article which is important (age_c). This age, taken in months, is then used as the divisor for the CoRank score obtained from the publication in question. Thus the primary CoRank score will come from the most recent publication which cites the target publication (controlled by a division operation), as shown in Equation 7.5.

$$CR_T(p) = \frac{1- \propto}{|V|} + \propto \sum_{c_j \in M(c_i)} \left(\frac{CR\,(p_j)}{CL\,(p_j)} / age_c + 1 \right) \tag{7.5}$$

With this equation fluctuating as more citations are gained, it is expected that the Spearman Correlation and percentile ranks will also fluctuate somewhat. As a direct result, this algorithm may reveal papers which are recently cited, whilst hopefully maintaining a large amount of older publications which are still highly cited within those which are ranked highly.

Figure 7.6, which depicts the age of the top 5% of all ranked publications, shows this to be broadly true. Frustratingly, the expected distribution of publications in the top 5% is reversed, with more younger publications (aged less than 12 months) being revealed than those which are established. If this was a graph depicting download metrics, then this may well be expected. Many consumers might look at a publication as soon as it is published, with many more downloading the publication again when it is cited.

Due to this algorithm considering the number of citing publications (and performing a division by this number), the distribution of publication ages can be explained by the

division by one, which will occur when a publication gains its first citation. This division by 1 will result in a high rank being obtained for that publication.

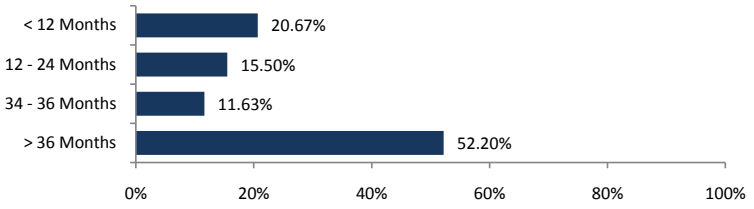

FIGURE 7.6: Average age of top 5% of publications in each snapshot (CoRank-CiteTime)

7.1.6 CoRank-CoTime

This second iteration of the CoRank-Time algorithm looks at the age of the publications the target publication is co-cited, with rather than the age of the citing publication. Thus the hypothesis is that a publications impact can be judged in more detail not only from other cited publications, as per CoRank, but also the age of these co-cited publications.

Unlike with the previous time based algorithm, where the most recent citations dictate the impact score, CoRank-CoTime looks to older publications to indicate rank. In Equation 7.6, this can be seen by the multiplication of the CoRank score by the Co-Cited publications age age_p. Logically the older the publication p the larger this multiplication of the normal CoRank score. This both eliminates the initial high impact and should limit the rate of change. Each publications CoRank-CoTime score obtained is a multiple of the age, in months, of the co-cited publications CoRank score as shown by Equation 7.6. As in CoRank-CiteTime, one is added to the age in order to eliminate errors caused through multiplication by zero.

$$CR_{CT}(n) = \frac{1- \propto}{|V|} + \propto \sum_{p_j \in M(p_i)} \left(\frac{CR\,(p_j)}{CL\,(p_j)} \times age_p + 1 \right) \tag{7.6}$$

Even with applying the time based factor in reverse however, the distribution of publication ages in the top 5% gives a very similar result to that given previous by CoRank-CiteTime, as shown by Figure 7.7.

7.2 Spearman Correlation - All Algorithms

The Spearman Rank Correlation, as already used earlier in Figure 6.8, computes correlations between the rank order lists generated by two metrics. As has been the case

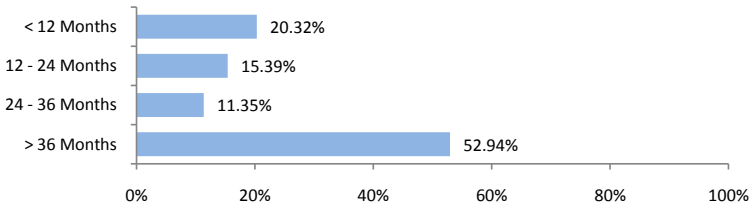

FIGURE 7.7: Average age of top 5% of publications in each snapshot (CoRank-CoTime)

throughout, Citation Count has been used as the target metric against which every other algorithm has been compared. In order to fulfil the first hypothesis — Co-Citation relations can be used to create an early indication metric for publication impact which correlates well with existing metrics — the ideal result would exhibit a high correlation at an earlier point in time.

As shown by Figure 7.8, Citation Count remains the clearest indicator of itself, with PageRank remaining the next best indicator. Unfortunately this means that the six algorithms introduced in this chapter are no better than PageRank in correlation to Citation Count at any point in the life cycle. This means that none fulfil the first hypothesis fully, but some algorithms do exhibit good positive correlation.

The majority of algorithms introduced in this chapter show a much improved correlation over the original CoRank algorithm. The exceptions to this are those which are based on a division by CoRank data (CoRank-Scaled-CoRank and CoRank-Divided), which show a similarly low level of correlation, something which was predicted during their introduction.

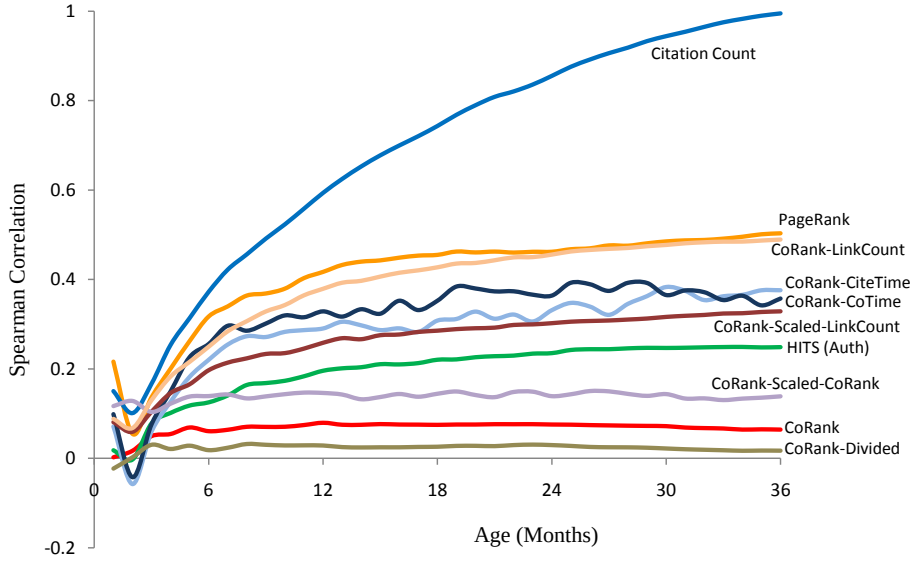

FIGURE 7.8: Comparison of Spearman Rank Correlation for all applied metrics

As expected the results from the time based metrics fluctuate as more data is added, however it is interesting to see these give a high correlation compared to many of the other metrics, given the percentage of newer publications in their top 5% (Figures 7.6 and 7.7). A more significant result would be if the sets of one hundred papers, measured in Figure 7.8, are in this top 5% of all papers by rank, something which is disproved.

The best performing of the six algorithms introduced in this chapter is CoRank-LinkCount. CoRank-LinkCount calculates impact by adding together the number of citations a publications' co-cited papers receive. PageRank remains the closest in correlation to Citation Count (other than Citation Count itself), explainable by both PageRank and CoRank-LinkCount being based upon the weighted rank of the directly citing publications. PageRank is based upon the PageRank (iteratively) of every other citing publication. Similarly CoRank-LinkCount takes the citation count rather than the PageRank. Neither consider a direct citation to hold a score of one.

In addition to just the correlation between each algorithm and Citation Count (as depicted by Figure 7.8), the correlation between the rank orders obtained by all algorithms can be used to help further demonstrate the similarities between a number of the algorithms. Table 7.2 shows the Spearman correlations between every algorithm after the full 36 months. Table 7.3 lists all the metrics in abbreviated forms with their relevant mappings.

	LC	H	PR	CR	CR_{LC}	CR_{Div}	CR_{SLC}	CR_{SCR}	CR_T	CR_{CT}
LC	1.00									
H	0.73	1.00								
PR	0.58	0.09	1.00							
CR	0.46	0.53	0.09	1.00						
CR_{LC}	**0.78**	**0.93**	0.12	0.57	1.00					
CR_{Div}	0.33	0.42	0.03	**0.98**	0.45	1.00				
CR_{SLC}	0.49	**0.92**	0.13	0.48	**0.95**	0.36	1.00			
CR_{SCR}	0.49	0.27	0.26	-0.01	0.43	-0.09	0.49	1.00		
CR_T	0.61	0.57	0.10	0.22	0.67	0.11	0.68	0.57	1.00	
CR_{CT}	0.60	0.55	0.10	0.29	0.66	0.19	0.65	0.53	**0.85**	1.00

TABLE 7.2: Spearman correlations between all algorithms after 36 months

Table 7.2 indicates the correlations between each pair of algorithms after a full 36 month life cycle. As such it is not possible to identify if a metric is a good early indicator of any other metric; only those which are similar in characteristics after 36 months can be identified here. A number of strongly related algorithms are visible in Table 7.2. Metrics where a direct or indirect correlation greater than 0.75 exists have been highlighted in bold (except when the two algorithms being compared are the same). Typically a strong correlation exists where the algorithms share similar characteristics, such as their input data and how they process this data. For example, a high correlation can be seen between the HITS, Citation Count and CoRank-LinkCount algorithms and similarly

Abbreviation	Algorithm
LC	Citation Count (Link Count)
H	HITS (Authority)
PR	PageRank
CR	CoRank
CR_{LC}	CoRank-LinkCount
CR_{Div}	CoRank-Divided
CR_{SLC}	CoRank-Scaled-LinkCount
CR_{SCR}	CoRank-Scaled-CoRank
CR_{T}	CoRank-CiteTime
CR_{CT}	CoRank-CoTime

TABLE 7.3: Algorithm abbreviations

between the two time based algorithms. With these correlations indicating similarities between metrics, a number of metric families can be created. Additionally these family groupings provide another means to ratify the results, as metrics based upon the same type of input data should logically exhibit similar characteristics. The data from Figure 7.2 can be used to mathematically identify these families of metrics using Principal Component Analysis, a technique which effectively plots the "distances" between each algorithm.

7.3 Rank Analysis

This section extends the work carried out in Section 6.6, summarised by Figure 6.22, looking at the percentile rank of the tracked sets of publications. This test is designed to examine if a metric is capable of revealing high impact publications, notably those known to be high impact by Citation Count, within the top n% of results. With Citation Count representing the accepted standard, a good metric should closely follow the same behaviour.

The hypothesis being examined in this section stated that by looking at the larger co-citation network, highly ranked publications should also be initially ranked higher in any given dataset.

Figure 7.9 shows the mean percentile rank of 12 sets of 100 target publications taken from subsequent snapshots between April 2003 and March 2004. Each of these 100 publications are then tracked over the following three years and the rank in each snapshot recorded. Finally these ranks are averaged to give the mean rank being translated into a percentile value. Figure 7.9 shows the percentile rank positions for the target set of publications for all algorithms over the full 36 months.

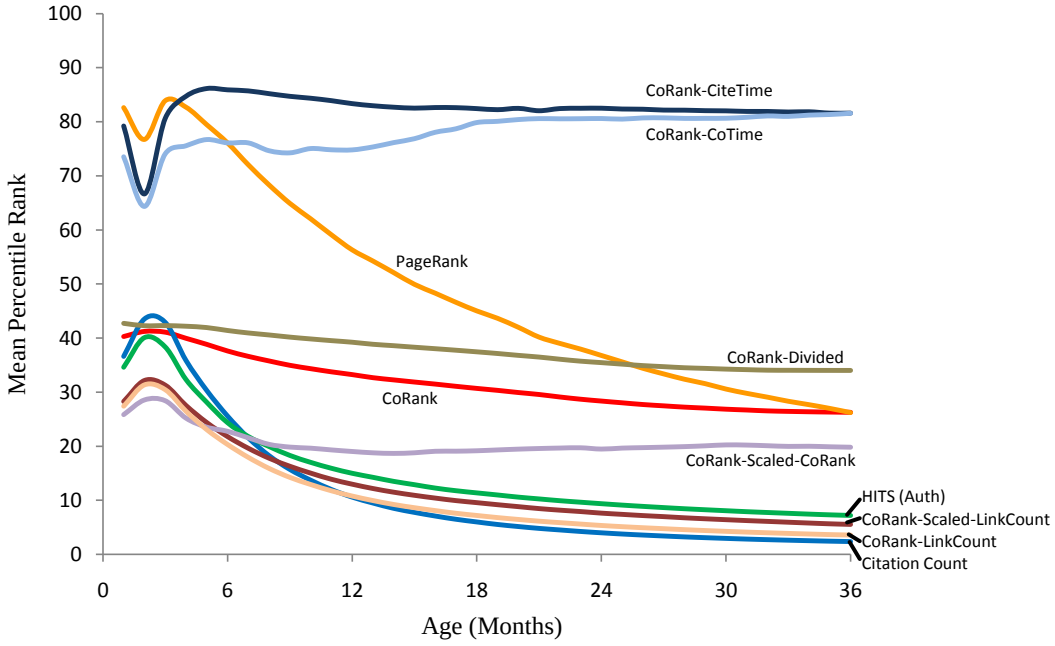

FIGURE 7.9: Average mean rank of top 100 publications (percentile measurement)

Citation Count sets a high target, however CoRank-LinkCount outperforms it over the first 12 months and then tracks closely with Citation Count over the next 24 months shown. A similar behaviour can be seen for two other algorithms (CoRank-Scaled-LinkCount and HITS) in Figure 7.9. Conversely though, over the first 12 months no metric manages to perform as well as CoRank-LinkCount. In this case CoRank-LinkCount does satisfy the hypothesis being tested.

Figure 7.9 is the first graph which shows clear groupings of algorithms beginning to emerge. The previous analysis by Spearman Rank Correlation (Figure 7.8) exhibited a distribution of algorithms which improve marginally over each other, however when looking at percentile ranks, there are some clear gaps between results.

The first of these groups includes the target algorithm (Citation Count) and can be seen as the four algorithms which trend very similarly towards a low mean rank (in Figure 7.9). The plotted results of these four metrics all follow each other in a very similar fashion over the three year period, with the majority of the variation being exhibited in the first 12 months. The second group can be seen at the other end of the scale, maintaining a very low rank position for the publications (high percentage). These time based algorithms seemingly never rank the target dataset highly, even though the Spearman Rank Correlation is better than many other metrics. The remaining four algorithms, which after 36 months can be said to be the mid-range performers, are the group which includes PageRank. All end up rating the target set of publications between the 20th and 34th percentile, but take a variety of different routes over time to get there.

7.4 Publication Age Analysis

As well as Spearman Rank Correlation and rank analysis, the average age of publications in the top 5% of results by each algorithm was also used to help indicate the behaviour of each metric. With one of the aims of a "better" algorithm being able to provide an earlier indication of later impact, publication age is important to show that an algorithm is boosting the visibility of younger publications. The third hypothesis predicted that this could be done by a metric which looks at the larger co-citation network.

Figure 7.10 shows the collated results for publication age, which throughout have shown that publications older than three years dominate the top 5%. PageRank, one of the most closely correlated to Citation Count, demonstrated this dominance most clearly with nearly 98% of the publications in its top 5% being older than three years of age. Conversely CoRank-LinkCount, which shows a similar correlation to PageRank, includes 7% more publications which are younger than three years. HITS (Auth) and CoRank-Scaled-LinkCount perform best in this analysis, revealing the highest amount of most recent publication whilst also maintaining the expected distribution shown by Figure 6.7 earlier.

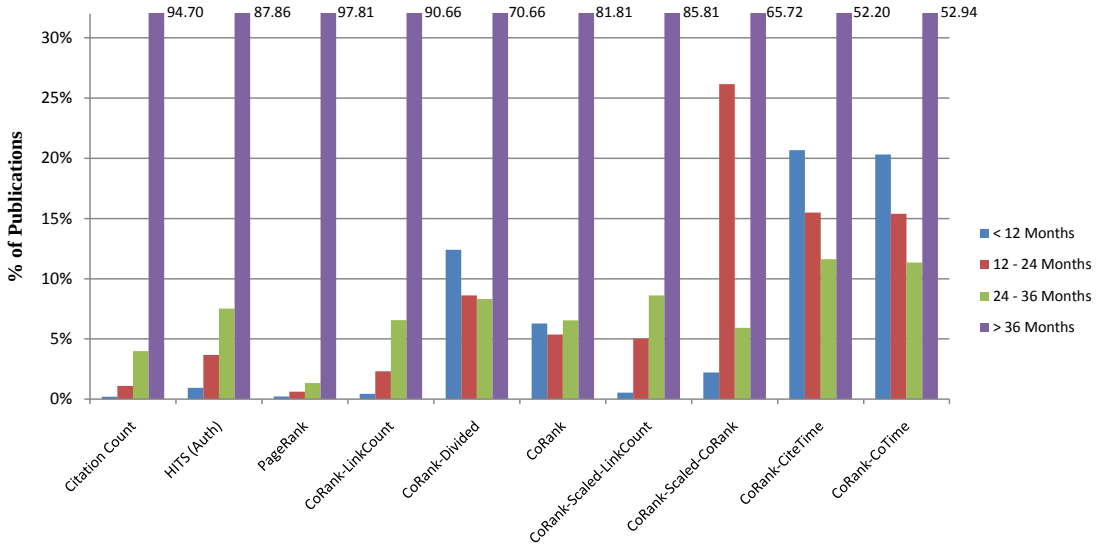

FIGURE 7.10: Average age of top 5% of publications, all metrics

Finally, from Figures 7.10 and 7.8, it would appear that CoRank-CiteTime to outperform many other algorithms by holding a 0.4 correlation with only 52.2% of publications in the top 5% being older than three years. However, with this not reflecting the expected distribution of ages (Figure 6.7), in addition to the poor percentile rank result (Figure 7.9), these algorithms don't reflect any logical behaviour.

7.5 Citation Distribution

Citation Count is often used as a surrogate indicator of impact; the higher the Citation Count the greater the perceived impact. Each time a search is performed, a set of publications will be returned with a varying number of citations. The maximum number of citations gathered by a single publication will be dependant on the search and average age of material; different subject and topics will also be at different points in their publication life cycles. As a result, a researcher will have to mentally assign their own boundaries to what constitutes a well cited publication.

In a similar fashion, this principal can be applied to the top 5% of publications as ranked by each algorithm, looking at the number of citations obtained by each publication. Logically, with the data ranked by Citation Count, the most highly cited publications are going to emerge top. It is how the other algorithms perform in comparison to this that is of interest.

As with the age tests, the top 5% are going to be examined, this time noting down the number of citations towards each publication before categorising publications into the following groups:

- $>$ 100 citations - Those highly cited publications with more than 100 citations.
- 10 - 100 citations - Those which have obtained between 10 and 100 citations.
- $<$ 10 citations - Publications which have obtained less than 10 citations.

Figure 7.11 shows the resultant percentages of publication in the top 5% which are in each citation categories. The results in the graph have been ordered from bottom to top to show which metrics reveal the greatest quantity of publications with a Citation Count above 100. Note that only 25% of publications have over 100 citations when ranked by Citation Count, representing the maximum possible figure in this category.

From this result, CoRank is maintaining a percentage (12%) of publications which achieve more than 100 citations, half of the figure obtained by Citation Count. Other than CoRank-LinkCount, all other algorithms based upon CoRank show disappointing levels of performance, revealing a very high number of minimally cited publications.

Looking at the results in this way reveals that two of the categories (Citation Count and CoRank-LinkCount) have no low impact results at all listed. PageRank reveals a significant number of lower impact publications, while CoRank is in fact worse based upon this metric, revealing a substantial number of low cited publications in the top 5% of all publications. HITS (authority) is the only algorithm to show a balanced set of results with 8% of publications being revealed from the low impact category which matches closely with the expected distribution shown by Figure 6.7.

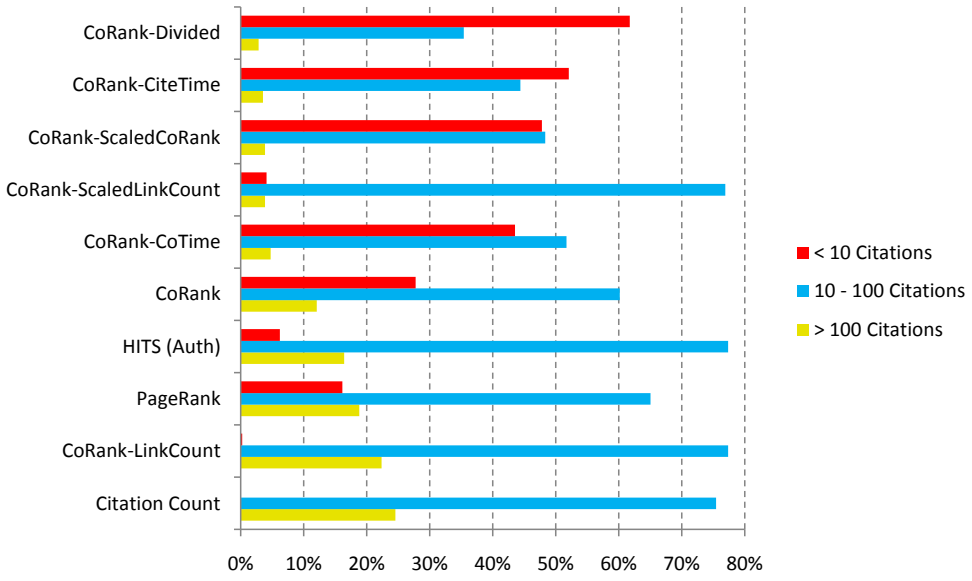

FIGURE 7.11: Percentage of publications (from top 5%) per Citation Count category

Citation Count has been used throughout this work as the target metric, again though it is important to emphasize that CoRank will, by design, exhibit a different behaviour. All the analysis performed in the last few sections (as well as in the previous chapter) is intended to both show the characteristics of CoRank as well as justify its potential usage.

7.6 Statistical Significance Testing

Before continuing analysis on which algorithms are showing promise and which are seemingly worthless it is necessary to ratify the results throughout this work by examining their statistical significance. Statistical significance is not about the results telling you anything important or meaningful; a result cannot be said to be significant just because it is statistically significant, but rather about how many times in 1000 you are likely to get the same result.

Statistical significance testing in publications has been mandated by many journals to ensure results are relevant and enough experimentation has been carried out to justify any conclusions drawn. However some argue that the technique is floored and abused and should be replaced. Carver (1993) argues a strong point against Statistical Significance testing, stating that is is often misunderstood to mean that the results are significant, but does not rule the method out as a potential way to locate misleading data. Mohr (1998) concurs with taking a cautious approach to significance tests stating: "One cannot be a slave to significance tests. But as a first approximation to what is

going on in a mass of data, it is difficult to beat the particular metric for communication and versatility". Others, including Johnson (1999) and Daniel (1998), argue the point to scrap statistical significance testing altogether in favour of other methods, stating that statistical significance testing can potentially "confuse the interpretation of data". Carver (1993) puts forward two points, the word "Statistical" should always appear in front of the word "Significance" so that readers (and authors) don't confuse statistically significant results with those which actually mean nothing. The second point is that all statistical significance testing should be carried out after the data has been analysed. Huberty (1987) perhaps sums it up best, "there is nothing wrong with statistical test themselves! When used as guides or indicators, as opposed to a means of arriving a definitive answers".

With the majority of the analysis already conducted, statistical significance is purely used here as an indicator to perhaps identify or confirm anomalies in the results.

Statistical significance is tied closely to the size of the sample set. For example if a die is rolled once, with the result being a 6, then it could be said the probability of rolling a 6 is 100%. On a fair die, this will never be the case, thus the result is not significant and repetition of the experiment would prove this.

Repetition and broad data selection is key to ensuring results are balanced and have more chance of being statistically significant. With the Citebase dataset containing data pertaining to over 170,000+ publications providing several million citations, there is no shortage of sample data. Each of the experiments run so far has selected a minimum of 100 publications per sample (with multiple samples used), in order to compute sets of results. This broad range of sampling and repetition of experiments should mean that the majority of results should be statistically significant.

There are two types of significance tests, one-tailed and two-tailed, where usage depends on the hypothesis. If the hypothesis states the direction of the difference, then this should be testing using a one-tailed probability e.g. Females will not score significantly higher than males on an IQ test. There is some debate (e.g. Eysenck 1960) on whether it is ever appropriate to use a one-tailed t-test; if you already know the direction of the difference, why bother doing any statistical tests?

The two-tailed test, which is simply double the value of the one-tailed test, is designed to test the null hypothesis that there will be no significant difference in the results of two experiments. In looking for a "better" metric, no preconception was made relating to the direction of difference or even if there will be one. Thus the null hypothesis can be used to predict no significant difference between the results, thus giving scope for a two-tailed test to be carried out.

Finally, there is a need to decide on a critical alpha level that is acceptable in order to say that the results are significant. In most cases 0.05 and 0.001 represent the two

levels of significance. Here anything above 0.05 is said to be likely down to chance, anything between the two values is significant and any results below the 0.001 threshold are strongly significant. Based upon the number of data items and samples that can be potentially obtained from the dataset, values below 0.001 should be looked for to indicate strong significance.

7.6.1 Results Significance

In order to test statistical significance between two sets of results a common starting point is required. In the case of this thesis, the results are the common starting point and the null hypothesis is examining if algorithm A is better then algorithm B at Job C.

Statistical significance represents that n times out of 1000 you will get the same result, however this concept cannot really work for individual papers ranked by the different algorithms. However, by taking the average results such as those used to plot Figure 6.22, the statistical significance relating to the differences displayed between each algorithm can be tested.

To test significance thoroughly it was decided to take 20 samples of 100 randomly selected publications which were published in Citebase x months ago, where x represents 6,12,18,24 and 36 months, thus covering the full range of data which was sampled throughout this work. This data covers 2000 publications in over 10 algorithms representing 20,000 publications in total. To make the data selection even more random, the time period over which the data was selected, was also random for each set of 100 publications, thus one set of 100 over a 6 month period could have been those added in January 2005 with the positioning rank data selected from the July 2005 dataset.

In order to select sets of 100 papers randomly from the raw data each publication was given a sequential numeric ID assigned in the order, by time, that the publications were added to Citebase. Using the *rand()* function in PHP[1] a number was chosen between 1 and the number of publications added that month until 100 unique publications had been chosen from the input set of publications, from which results data was gathered in the same manner as before.

Once selected, the positions for each publication in the sample set of 100 could be calculated using each of the 10 algorithms, with the average position fed back as the result. Having 20 sets of results for each of the 10 algorithms allows the calculation of statistical significance between each in a matrix format where all algorithms are compared to one another.

[1] PHP - http://www.php.net

	LC	PR	H	CR_{LC}	CR_{Div}	CR	CR_T	CR_{CT}	CR_{SLC}
PR	✓								
H	✓ ✗36	✓							
CR_{LC}	✓ ✗36	✓	✓						
CR_{Div}	✓	✓	✓	✓					
CR	✓	✓	✓	✓	✓ ✗6,12				
CR_T	✓	✓	✓	✓	✗✓36	✗✓36			
CR_{CT}	✓	✓	✓	✓	✗✓36	✗✓36	✓ ✗36		
CR_{SLC}	✓ ✗18,24	✓	✓	✓	✓	✓	✓	✓	
CR_{SCL}	✓ ✗6	✓	✓ ✗6	✓	✓	✓	✓	✓	✓

TABLE 7.4: Statistical Significance Testing all 10 metrics, ✓= Significant, ✗= Not Significant

The results of a two-tailed t-test are shown in Table 7.4, with algorithm names again abbreviated (see Figure 7.3). In Table 7.4, a tick represents statistical significance stronger than a 0.001 probability, while a cross represents the opposite. As this data is representative of all the time periods (6,12,18,24 and 36 months), the tick or cross represents the majority of the results, while any number listed represents the time periods for which the results passed or failed, depending on the symbol that the numbers are next to (so ✓✗18,24 represents a reading where all results were significant to the 0.001 level except for those from the 18 and 24 month snapshots).

By comparing the metrics using statistical significance tests, positive results should show that the algorithms are distinguishable from each other. While this is clearly true in most cases, there are a few instances where, for a few snapshots, the algorithms cannot be differentiated. CoRank-LinkCount and Citation Count (LC) provide one such example where the final reading was taken after 36 months. This is very likely due to the algorithms being similar in characteristics.

Helpfully the clearest results involves the time based algorithms (CR_T and CR_{CT}), which show no significance shown when comparing these algorithms against others in the CoRank family. Although the time base algorithms are significantly different from a number of other algorithms, this result along with results from the previous section does not offer any more encouragement to say that these metrics are useful in their current form. CoRank-Time and CoRank-CoTime are simply not consistent and distinct in their behaviour to be of any real value.

The bulk of the remaining results are statistically significant, the majority to a degree much smaller than 0.001 due to the sample size. This ratification of the methodology and applicability of the results implies that earlier observations can be said to be accurate, especially in the case of the applicability of the time based metrics.

Potential clustering of algorithms has also already been observed, and further supporting evidence is also offered by the statistical significance tests. Table 7.4 shows a number of

metrics which are statistically significant except in a few minor cases (e.g. ✓✗36). It is this minor cases which suggest a similar relationship between metrics such as Citation Count (LC) and CoRank-LinkCount (CR_{LC}).

7.7 Equation Groups

Throughout this thesis, three tests have been undertaken which analysed the characteristics of a number of citation metrics in the hope of finding a new early indicator of subsequent impact. While none of the metrics outperformed Citation Count in a controlled publication environment, the different characteristics exhibited do not necessarily mean that the metrics do not have usage in other situations.

It is these differing characteristics which groups the various metrics together. This is particularly clear in Figure 7.9 which led to the introduction of three groups of metrics in Section 7.3. This section looks at these groups in more detail, explaining the reason why each algorithm exists in the group and merits of each.

There are three groups of algorithms in total; one containing algorithms based upon direct citations, one based upon co-citations with the last containing the time based metrics that perform badly in all tests. While the majority of algorithms can be easily placed into one of these three categories, based both on their characteristics and test results, PageRank is an exception. PageRank is a citation based algorithm which includes a weighting factor, however from the results for rank position (shown in Figure 6.22) PageRank reflects the behaviour of the Co-Citation based metrics. Something not reflected in the other two test where PageRank behaves in a similar manner to classic citation based metrics it is based on. For this reason, PageRank remains in the category of algorithms based upon direct citations, but could be said to be a bridging algorithm.

7.7.1 Citation Based Algorithms

This group of metrics consists of the following:

- **Citation Count** - The best performer, predictably as all tests are carried out against itself. As the control, it was observed that Citation Count is the best early indicator of itself by rank correlation order. However high ranking publications take time to gain overall standing and very few new publications are ever revealed.
- **PageRank** - One of the highest Spearman Rank Correlations to Citation Count but the mean rank of the target set of publications takes some time to establish and is still not stable after 36 months, a direct result of which being that PageRank does not reveal any recent publications.

- **CoRank-LinkCount** - Overall the best performer, almost identical to PageRank by Spearman Correlation result, while the mean rank of the publications starts higher than Citation Count and maintains this standing. Lastly it also reveals a good number of newer publications along the way.
- **CoRank-Scaled-LinkCount** - Just behind CoRank-LinkCount by mean rank, slightly better at revealing newer publications but a significant distance from both CoRank-LinkCount and PageRank by Spearman Correlation.
- **HITS (Authority)** - A generally average performer. Excellent mean rank but other than that, average in all other categories.

CoRank-LinkCount, the most successful new algorithm in this group, reflects the behaviour that a highly respected paper is likely to get cited alongside other highly respected papers. This also follows that it is in human nature to simply copy citations verbatim. By examining the citation count of the publications with which a paper is co-cited, CoRank-LinkCount has the potential to be a good early indication metrics of subsequent Citation Count after three years.

In the early stages of the publication life cycle, CoRank-LinkCount outperforms Citation Count in two of the three tests, that of mean rank and publication age. During the first 12 months, on average, a typical publication which is later ranked highly by citation count, is already ranked highly when CoRank-LinkCount is applied. As a direct result the number of more recent publications in the top 5% is higher. The only result that lets CoRank-LinkCount down is the Spearman Correlation of the rank order between the two lists, meaning that a general search will return the same publications sooner in the life cycle, but never in the same order as by Citation Count.

CoRank-Scaled-LinkCount performs in a very similar fashion to CoRank-LinkCount except by Spearman Correlation, where this algorithm shows that it is based more upon the original CoRank algorithm. Again though, the mean rank of the publications shows similar promise to CoRank-LinkCount and it can be seen to be revealing a substantial number of more recent publications. Being based on the original CoRank algorithm will have the benefit that CoRank-Scaled-LinkCount is able apply a prestige factor to citations. This is due to the weighting factor being maintained, even if this is now not the primary factor in the metric.

PageRank, which also applies a prestige factor to citations, has established itself on the Web, however performance on a controlled network is clearly not as promising. This is not surprising due to the temporal, static nature of citation networks. PageRank is dependent on the PageRanks of all the directly citing publications to be stable before an accurate rank can be given. As a result this actually doubles the time it takes to stabilise the mean rank, rather than the opposite, meaning that the chances of revealing newer publications are almost none. Once the PageRank does begin to stabilise, it redeems

itself by holding a fairly good rank correlation to Citation Count, however this has been matched now by CoRank-LinkCount.

7.7.2 Time Based Metrics

This second group of metrics are a strange set, the time based factor does lead to newer publications being revealed in the top 5% fairly quickly. However, this 5% does not contain any of the publications which are rated highly by Citation Count. Additionally, while the sampled 100 publications are not ever listed near the top 5%, there is a good correlation between their rank order in the dataset and that obtained via Citation Count. However, not being statistically significant, combined with the bad mean rank and publication age data means any benefits exhibited should put down to pure chance. A good course for further investigation would be to find a means of adjusting these algorithms in a generic way such that the results from the publication age tests reflect the more idea result outlined in Figure 6.7.

The time based algorithms, were designed to take into account the age of citing or co-citing publication. CoRank-CiteTime was designed to see if the most recent citation is the most accurate reflection of impact, while CoRank-CoTime was designed to do the opposite with the publications with which you are co-cited. Neither of these reflect a typical human behaviour, logically a paper is just as likely to get cited by a low impact publication as by a high impact one, hence the fluctuations in results.

7.7.3 Co-Relation Based Metrics

Having accounted for seven of the ten algorithms, this leaves the three which are all based on the original CoRank algorithm; all making use of the co-citation network. This group of metrics consists of the following:

- **CoRank** - Not brilliant by Spearman, peaking at around 0.1 correlation to Citation Count. Not the best by age, while it reveals a spread of recent publications of all ages, this is not conforming to the distribution outlined in Figure 6.7. CoRank is similar to PageRank after 36 months by mean rank but starts much better (the only positive result).
- **CoRank-Divided** - As expected this algorithm performs worse than CoRank. All high ranked publications by CoRank-Divided turn out to be those with very few citations.
- **CoRank-Scaled-CoRank** - This is the best performer in this set of algorithms showing the a good correlation and best mean rank while also revealing an extremely high number of publications between 12 and 24 months old (26%).

Although it might be quite easy to discard this set of algorithms, the fact that CoRank performs better than PageRank in two of the three tests is significant. CoRank stabilises the rank of the publication in a much shorter amount of time and to a higher value, while also revealing a good number of more recent publications in its top 5%. The major problem is the distance between CoRank and the results of the citation based algorithms in both the rank position test and, more significantly, the Spearman Correlation test. As CoRank takes influences from PageRank, it is positive to see the potential benefit of CoRank over PageRank, however when compared to Citation Count neither perform well.

The obvious benefit for the CoRank and PageRank algorithms comes from the fact that they are both weighted and apply some level of prestige to high quality citations. Conversely this technique can also be used to rule out false positives, which are highly cited by papers of low prestige score. In an open publishing environment, such as on the Web where PageRank is used already, Citation Count may not be the best metric to use. Additionally, CoRank has been shown to hold some benefit in a scholarly network over PageRank.

7.8 Conclusion

This chapter has revealed that CoRank, and all of its variations, show behaviours different from metrics based on Citation Count, except in the case of CoRank-LinkCount. CoRank-LinkCount drops the weighting factor in favour of being based upon Citation Count. CoRank-LinkCount is calculated from the total number of citations accumulated by all the co-cited publications to the one in question.

By counting the number of citations towards the papers a publication is co-cited with, an immediate increase in the publication rank is witnessed (Figure 7.9). This increased rank is something not matched by Citation Count until the publications are older than a year. Additionally, CoRank-LinkCount is revealing a number of more recent publications aged between two and three years without being overly generous to those which are younger. This means that CoRank-LinkCount satisfies two out of the three hypotheses relating to using the co-citation network; it is revealing high impact publications sooner in their life cycle.

Although no metric matches or outperforms Citation Count when looking at correlation, CoRank-LinkCount shows a similar correlation to PageRank; a positive correlation around the 0.5 level. It also exhibits this similar performance to PageRank, while outperforming it significantly in the other two tests, making CoRank-LinkCount the best all round performer.

The problem with CoRank-LinkCount is that it still does not help when trying to account for false positives, even though it does apply some factor of prestige (that of a paper being co-cited with other highly cited publications). In order to increase the rank of an article, an author would need to publish a paper which cites the original article alongside many highly cited publications. This is where the original CoRank algorithm is beneficial. Being based upon PageRank means that it can handle false positives while taking into account a more established network of CoRank scores. CoRank outperforms PageRank in all but the correlation tests, proving its worth as a weighted algorithm, however when compared to a non weighted algorithm, the benefits are not as clear.

If a global need arises for a better metric than Citation Count, then this metric either needs to be widely adopted, or be very similar in characteristics to Citation Count. In Section 7.9 a number of groups of algorithms started to form based upon their performance. In the following chapter, these differing characteristics are examined in more detail, with potential benefits of application in different environments discussed.

Chapter 8

Navigation and Usage of the Metrics Landscape

The measurement of impact is one of many applications of bibliometrics. Common to all bibliometric research is the aim to study human behaviour and common pattern which apply to large groups of people. Examples include studies on how people use words, how networks are constructed between people and publications and how people judge impact. It is this last topic, the ranking of publications with the aim of reflecting the usage of materials which has been focused upon in this thesis. It was observed how the study of citations and a simple Citation Count is accepted as a good indicator of the popularity of a publication. Citation Count is a good surrogate indicator for how much a publication is being read as logically the greater the number of readers the higher the possible Citation Count.

Often an author follows citation links between publications and even copies these citations verbatim for inclusion in their own works. In a study by Simkin & Roychowdhury (2005) of citations in papers, it was found that up to 90% of citations were copied verbatim between reference sections. This shows that authors are either being lazy when compiling references but still reading the cited material, or just copying the citations verbatim from an existing publication in order to make the same point within their own work. This study was carried out by observing the propagation of mistakes and citation styles (e.g. Author order) between reference sections in different publications. Pop culture — the acceptance of something within the mainstream of a given culture — can be seen to be influential with citations once a publication achieves a certain number. At this point in the publication life cycle, general acceptance of the work will lead to an increased citation rate, thus making the work more popular. At this point the in depth study and review of the cited work may not be carried out as "everyone else is citing this publication".

Bibliometric techniques along with the various ranking algorithms have represented attempts to model social behaviour in order to quantify social patterns. Consequently though, these studies have also ended up influencing the behaviour of institutions and individuals when publishing work, something observed when talking about the publish-or-perish paradigm. Logically, if authors know what metrics are going to be used to judge their work, then they will endeavour to publish in a manner that achieves the maximum mark possible according to this metric. In turn the place of publication may go against the choice they would have otherwise made in order to share their work most effectively with the community.

Typically journals are recognised as the most important publication medium for scholarly communications, however the pressures relating to journal publishing and the time frames involved can often lead to problems. The explosion in Computer Science research during the 1990s led to a saturation in journal publishing where only 13% of all Computer Science publications made it into journals, compared to the 50% in other areas (Fry et al. 2009). The amount of publications and rate of scientific process was seen as too fast for journals to keep up with and thus conference proceedings became a key part of the field. Today in Computer Science there are many conferences covering different topic areas, each publishing their own proceedings, a move which has increased the number of publications in the area as well as the rate of publication. Each of these conferences often starts with a specialist community wanting to establish a new area of study and publication. Likewise, after a number of years many of these conferences may cease due to lack of popularity.

Computer Science is one example of an area which is breaking away from the journal publication paradigm. This indicates that a metric based purely upon journal data may not accurately portray the behaviour and impact in this research area. Subsequently, any resultant study which affects decisions relating to employment and funding needs to take into account the changing nature of the field itself. The proposed 2012 (now delayed) Research Excellence Framework (REF) study, is one such example which, at initial time of proposal, was intended to be entirely based upon bibliometric techniques. It has since been found that it would be unsuitable to carry out a study which uses a fixed set of bibliometric techniques to measure performance over all areas of study (Macpherson Barrett 2009).

This chapter begins by looking at the background work which led to the conclusion that a single bibliometric indicator should not be used to rate academic institutions in the United Kingdom. By further examining the changing nature in scholarly communications, it is possible to observe that many areas of study use mediums other than journals to disseminate their works, thus compounding the problem of choosing a journal based metric for subject evaluation. To cope with the many different mediums and journals, Institutional Repositories (IRs) are playing a key role in collecting records of these works

in order to collate a complete portfolio of institutional output. From this portfolio an institution can select and put forward a number of publications to any research assessment exercise.

IRs are not the only source of information about highly distributed publications. Services such as Google Scholar also provide valuable endpoints from which such information can be obtained. Early research suggests that information from such services could also be a valuable source of publication information which presents a balanced view of research areas. A number of these early studies are also outlined in this chapter.

The initial part of this chapter focuses on re-enforcing the changing nature of scholarly communications led by changing behaviours. It is these changes which have to be considered very carefully when deciding on any metric used to measure to value of institutions working in that area. A metric which does not accurately reflect the current behaviour of the research area, be that because of the metric characteristics or the input dataset, is not a very useful metric.

The main section of this chapter then looks at the "landscape" of metrics available and shows how these can be mapped in 2D space to show the clear differences between many families of metrics. Drawing on both knowledge about the ways in which different disciplines operate and as well as work carried out by the MESUR project, allows the visualisation of various different impact metrics, ranging from impact factor to citation and usage metrics. From this it is possible to place CoRank in amongst the ecosystem of metrics and observe the key role that CoRank based metrics may play as a new "species" in this environment.

8.1 Higher Education Metrics

In order to measure the excellence of an academic institutions' research (in the UK), a judging panel has been used to review a number of institutional outputs and indicators. Each institution submits what it regards to be some of its best academic outputs, including a set of publications, to the panel of judges who then analyse these contributions against those of other institutions to rate academic performance. Since the last study, there has been a query regarding the possibility of using only quantitative bibliometrics in place of panels to generate this ranking. As a result, a pilot study was commissioned to develop the bibliometrics element of the Research Excellence Framework (REF), with one of the conclusions addressing whether automated bibliometric techniques are robust enough to use in place of expert review (Macpherson Barrett 2009).

Central to this REF study, is the capability to gather and submit a collection of publications which are likely to score highly. In this process an Institutional Repository (IR)

can play a key role. Other than being a means to disseminate and preserve an institutions work, IRs can be a central place from which collections can be gathered, even if the records indexed in the IR do not include a full text copy. IRs can also monitor and collect citation data pertaining to publications making the gathering of those with high impact (if Citation Count represents this) easier.

The important role of the IR was particularly evident in the REF 2012 pilot study (Macpherson Barrett 2009). 22 institutions volunteered to take part in the pilot study with total output counts ranging between 500 and 38,000 publications including books, articles and conference proceedings. Of those who volunteered to partake, very few did not have an IR.

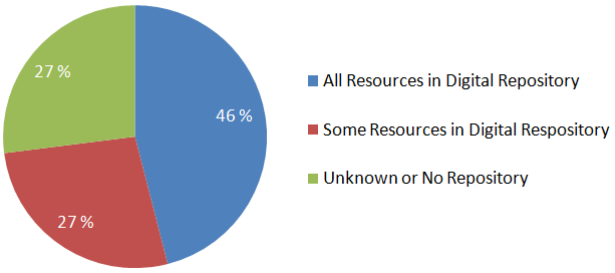

FIGURE 8.1: Coverage of IRs in REF pilot institutions

Figure 8.1 shows the results of a brief search for the presence of an institutional repository listed in the Repository of Open Access Repositories[1] (ROAR) or found on the Web. Only 27% of the institutions taking part in the pilot study did not have a repository. This demonstrates the widespread uptake in the UK and value of such resources. Perhaps key is the fact that nearly 50% of the institutions could have submitted all of their publications data directly from the repository. This is due to the repository containing more records currently than the number submitted to the survey.

These figures demonstrate how many institutions are now taking advantage of repositories as a central place to collate works. Consequently they are able to track citations and downloads to utilise these factors as early indication and surrogate metrics for impact in surveys including the REF. Repository systems, being a hub for information, could also be seen as a candidate to deploy and analyse newer metrics including CoRank.

Additionally (or conversely), IRs provide a repository of multi-disciplinary works, implying that there may been a requirement to apply different metrics which model the behaviour of each research area more effectively.

[1]Repository of Open Access Repositories (ROAR) - http://roar.eprints.org

8.2 Author and Subject Publishing Trends

Open publishing is one of many factors which has also led to a change in authorship and collaboration groups. With the Web and technologies, such as email, now being a part of everyday life the perceived distances between authors has shrunk significantly. In turn this leads to world wide collaboration between both authors in the same area and also between different subject areas. This growing trend in co-authorship is also changing the nature of publications in certain areas as authors have different publication preferences. Co-authorship among disciplines is also an important factor affecting eventual publication medium, often dictated by each authors desired audience. It is this multi-author, cross discipline work which represents another key area where co-citations can be studied.

Kyvik (2003) outlines the changing trends in co-authorship, which is occurring as a result of greater collaboration and enhanced communication channels, while Levitt & Thelwall (2008*a*) look at the benefits of multi-disciplinary research attempting to find if any resulting publications are more highly cited. While it is clear that co-authorship is on the increase, as yet no distinct difference has been found between citation rates of these cross-disciplinary articles compared to those more traditional publications.

Wuchty et al. (2007) observes that an increase in co-author counts is much more apparent in certain areas. Figure 8.2 (taken directly from this work) goes some way to proving the effect that factors, such as better communication channels, are having on publication trends, this is particularly the case during the 1990s with the uptake of the Web and email.

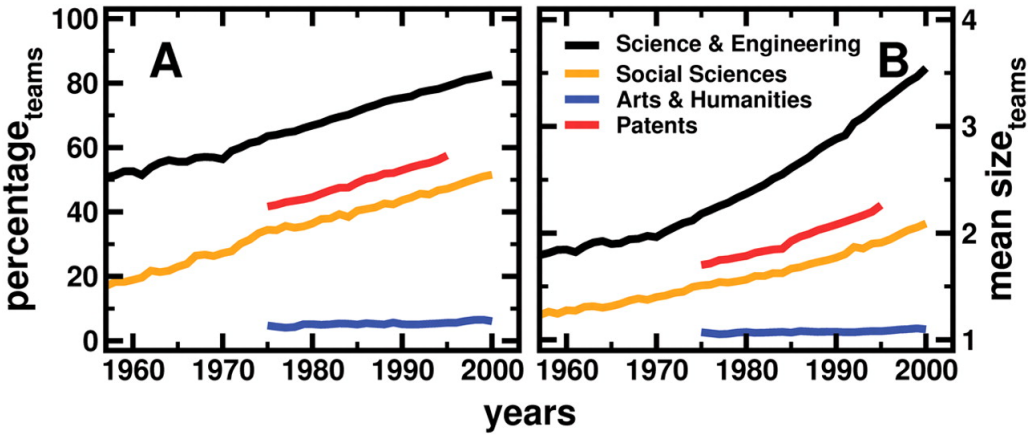

FIGURE 8.2: Growth in Author Teams by Subject Area (from Wuchty et al. 2007)

Distinctly different paradigms, such as co-author collaboration. is not the only means to differentiate between the various areas of scholarly study. Different areas also have preferred means of publication. Traditionally, the preferred method for publication has been the journal, led by the Journal Impact Factor and having well established techniques of publication. More recently other publication mediums, such as peer reviewed

conference proceedings, have been gaining traction in certain subject areas. Additionally, each medium has a different barrier to publishing. Drott (1995) observes that in many fields a conference publication goes on to become a journal publication, except in the area of information science (an earlier name for Computer Science).

In Computer Science many believe that a conference program committee are often greater domain experts than journal peer reviewers, thus conference proceedings could be said to have a higher quality (Drott 1995). With the majority of subject areas at that the time of Drott's study achieving a 50% translation from conference to journal article, Computer Science was moving at such a rate that only 13% of articles went on to form the basis of journal publications. This observation is most likely due to a number of factors including publication delay for journals and rate of scientific progress in the area during the 1990's through into the 21st century.

The question then becomes, if Computer Science articles are primarily in conference proceedings, can a metric which only measures outputs in Journals be used to accurately judge institutional excellence?

Fry et al. (2009) reports on a thorough investigation into the publishing practices of many disciplines in the UK, concluding similarly with findings regarding multi-authorship and publication mediums. By conducting a widespread survey, Fry was able to establish which publication types were viewed as most important in each research area. Seven different research areas were covered including Medicine, Engineering/Computing and Humanities. Experts in each area were asked to rate the importance of five different types of publication medium, including journals, monographs and conference proceedings. Figure 8.3 presents a summary of this survey showing the percentage of respondents who viewed each medium as being very important.

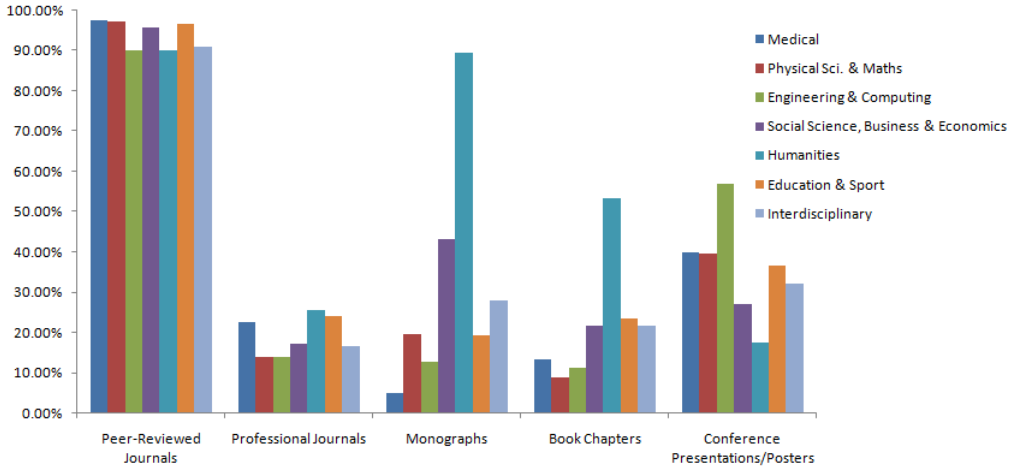

FIGURE 8.3: Perceived importance of publication types

Although there were a different number of respondents in each subject area, normalising the figures into percentages still shows clear trends in importance of each publication

type. Respondents were asked to rate each form of publication in a five point scale, from very important to not applicable. Both Figures 8.3 and 8.4 show the percentages relating only to the highest category, very important. From these figures, journal publications are still observed as important in all subject areas.

Re-plotting the figures by subject area rather than by publication type, shows that other publications mediums (particularly monographs in the area of Humanities) are also regarded as very important. Also clear is the relationship between Journal and Conference proceedings in Engineering and Computing Sciences.

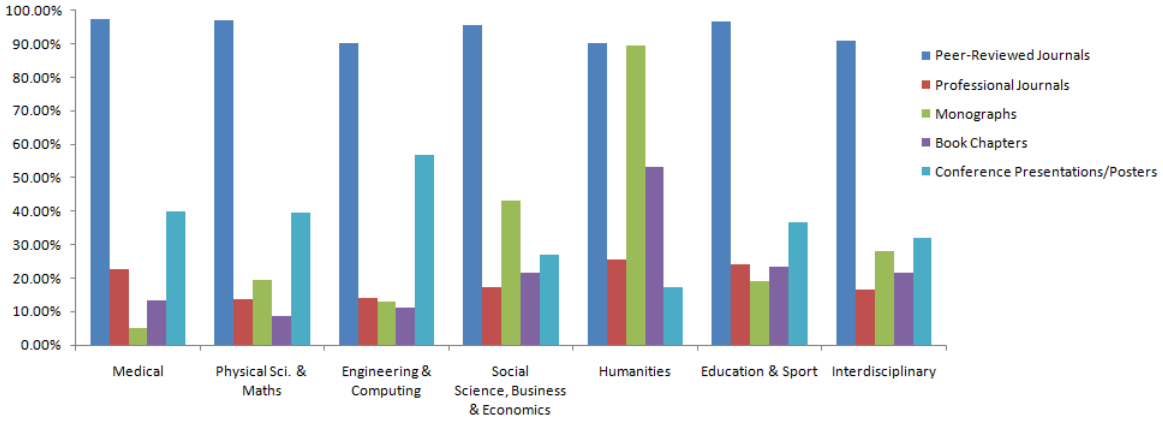

FIGURE 8.4: Perceived importance of publication types, by subject area

Fry et al. (2009) discovers that although there are other important publication mediums which could be considered when rating an institution or subject area, journal publications are perceived as equally important in all areas. So while Journals may not present a complete picture of research in each area (if only 13% of articles are covered), they still remain the best general source of bibliometric data for all subjects.

Charles Oppenheim, a specialist researcher in the area of bibliometrics, demonstrates the statistically significant relation between impact as judged by the RAE (the previous name for the REF) and journal citation rates in two papers (Oppenheim 1995, 1997). Each of these publications looks at a number of research areas including Archaeology (covered under the Humanities header by Fry et al. 2009) which is renowned for low Citation Count monograph publishing. In all areas a string correlation was found between the Citation Count and the RAE score. Oppenheim concludes, that if the correct literature is included in the study, then Citation Count should be the primary, but not the only means of calculating RAE scores.

On a Web scale, Li et al. (2003) look at the correlation between links to the various Computer Science departments in the UK and their RAE score. Like Oppenheim, Li finds a strong correlation suggesting that the proliferation of Computer Science publications on the Web can also be used to help judge impact of not just those publications but also the institutions which they represent.

Such direct correlations have driven the idea of a quantitative only study of institutional excellence. However caution is advised in areas which are still evolving, as demonstrated by results from Fry et al. (2009). With information pertaining to relevant publication now being spread among conference proceedings, monographs and journals (if only considering the top 3 mediums), there is a clear need to build services which can harvest and search all of these mediums. Institutional repositories represent one such technique, however on a larger scale Web services such as Citeseer and Google Scholar could play a critical role.

Goodrum et al. (2001) look at Citeseer as a potential alternative source of citation data. Citeseer contains a considerably greater amount of articles from conference proceedings than ISI (the former name for Web of Science) and Goodrum et al. (2001) find that these articles are also more prominent in the top 500 most cited. By only considering the top 500 most cited publications indexed in Citeseer and ISI, Goodrum finds that 15% of the top 500 in Citeseer are conference proceedings, 37% are journal papers and 42% books. Looking at the top 500 by ISI reveals that while the percentage of journal papers in the top 500 remains about the same, books increase to 56%, decreasing conference publications to only 3%. This represents a significant difference between the two indexes, which may end up affecting any citation study significantly.

Noruzi (2005) proposes Google Scholar as one of a new generation of citation indexes while Meho & Yang (2007), Pauly & Stergiou (2005) and Rahm (2008) all look at how such free services compare to those provided by Web of Science (or ISI dependant on the time of study).

Meho & Yang (2007) finds about 40% of articles in the field of Computer Science indexed by Google Scholar are journal publications and this is roughly matched by conference proceedings, leaving the remaining 20% to thesis and other technical reports.

Pauly & Stergiou (2005) offer an interesting insight and research comparable with that of Brody & Harnad (2004) and Antelman (2004). They show that conferences and articles indexed by Google Scholar tend to be more highly cited that those in journals indexed by ISI. This can simply be attributed to discoverability of the article; the bigger the audience for a publication, the greater the potential number of citations (Brody & Harnad 2004).

Finally, Rahm (2008) asks “While it is logical that articles indexed by Google may achieve a greater number of citations (according to Google Scholar), does this map to data obtainable via ISI?” By taking a selection of articles indexed by both services, Rahm (2008) concludes that there is a significant positive trend between data obtainable from Google Scholar and ISI. However, because of the preference for conference proceedings in Computer Science, Rahm finds that the impact factors for each conference are significantly higher when calculated from Google Scholar data rather than the data obtainable from ISI.

The consensus is clearly one of support for services such as Google Scholar. Such services are no longer constrained to focusing on only journal articles, due to the processing capabilities now available to help normalise citations and disambiguate between authors. By opening up to all forms of publication, whilst still focusing on the citation as a means to rank each publication, should allow the more accurate tracking of human behaviour in many disparate areas of research.

Counting citations is not the only bibliometric technique which can be used to indicate scholarly significance. As outlined in the course of this thesis, there are many other potential metrics which can be used for judging the impact of different things (e.g. Publications, Authors and Institutions), at different stages of the publication life cycle.

In addition to carefully considering the metric to be used, the sources of data are significant to ensure a fair representation of the chosen community. In this thesis, rather than addressing if areas of research are modelled accurately, it was chosen to compare the performance of a number of algorithms on a single dataset. Equally valid is the extension to considering a wider data source and the consequent changes in each algorithms performance, and accuracy, in modelling behaviour in the area of scholarly communications.

With bibliometrics being the study of human behaviour it may also be the case that each group of people, area of study or type of publication could also be judged better using different metrics. This is still an open question and may become more important if automated quantified studies of research excellence, gain major traction.

8.3 The MESUR Project

Throughout this thesis, a number of different metrics have been introduced which can all process the same input data (e.g. citation data) or be used to process different types of data. In addition, the source of input data may also affect coverage of an area of study. This aspect is particularly pertinent when looking at research data which is published in many locations and not just as journal papers.

Figure 4.1, first introduced in Chapter 4, introduced the four main factors — F (Frequency), R (Readers), S (Structure) and A (Authors) — that combine in pairs as bibliometric indicators. FA (Frequency-Author), of which the Web of Science (WoS) Impact Factor is the perfect example, represents the mapping between authors and popularity of journals measured via the counting of citations. SA (Structure-Author) contains all citation metrics which look at the links between papers including the PageRank algorithm. RF (Reader-Frequency) covers aspects including download statistics (Section 4.1), linking a reader to publications they have downloaded and potentially read.

Lastly, RS (Reader-Structure), relates to the paths that readers take between materials, by studying the pathways a reader follows rather than the citations an author gives, a behavioural study can be performed on actual usage data. It is this usage data that is the core focus of the MESUR project[2]. By taking usage data, the MESUR project looks at the potential of using this new data source as an indicator of scholarly impact.

In order to process such data, the MESUR project applies an approach similar to that outlined in this thesis. A number of different metrics (new and existing) are applied to this new data source in order to access their potential to indicate scholarly impact. With each algorithm able to accept a number of subtle changes and re-designs in order to be customised for the new data source, a significant number of new metrics can be conceived. Bollen et al. (2008) share this realisation as part of the MESUR projects study of the differences between download and reader pathway metrics in the field of bibliometrics and a total of 47 metrics are applied to usage data as part of the project.

During the course of the study, the MESUR project collected together one billion article usage events spanning five years from 2002 to 2007 pertaining to over 100,000 serials, 10,000 journals and 2,000 institutions. This data, collected from six publishers as well as consortia and aggregators aims to represent a comprehensive overall view of scholarly communications.

In order to examine this data, a database similar to that outlined in this thesis was constructed, upon which 47 bibliometric algorithms[3] were applied. These algorithms represent two sets of 23 where one set is fed citation based information and the other usage based information (an example of how each algorithm can exist in multiple categories). On top of these two sets of results is added Impact Factor, making 47 algorithms in total.

A correlation study between these algorithms, similar in nature to the first test criteria in this thesis, was then carried out in order to find the similarity between all of the metrics. This resulted in a 47 x 47 matrix of Spearman Rank Correlations. Using Principal Component Analysis, this correlation matrix can then be mapped onto a 2D plot showing the similarities and difference between the various metrics (shown in Figure 8.5).

Principal Component Analysis (PCA) (explained further in Appendix A) is used to obtain a set of eigenvectors relating to the correlation (or co-variance) matrix of which n (in this case two) can be chosen as the principal eigenvectors. A principal eigenvector is classified as being significant due to having a high eigenvalue. The correlation matrix is then mapped into the space spanned by the two principal eigenvectors to give a 2-dimensional (x,y) map of the matrix correlations as dictated by PCA.

[2]The MESUR Project - http://www.mesur.org/MESUR.html
[3]MESUR metrics: counts and networks - http://www.mesur.org/Metrics.html (May 2010)

Figure 8.5 shows the similarity of the different metrics as applied to both citation and usage statistics. Each algorithm has first been plotted on the graph before colour has been added to indicate clustering of algorithms, shown by a dark colour representing a greater concentration of algorithms.

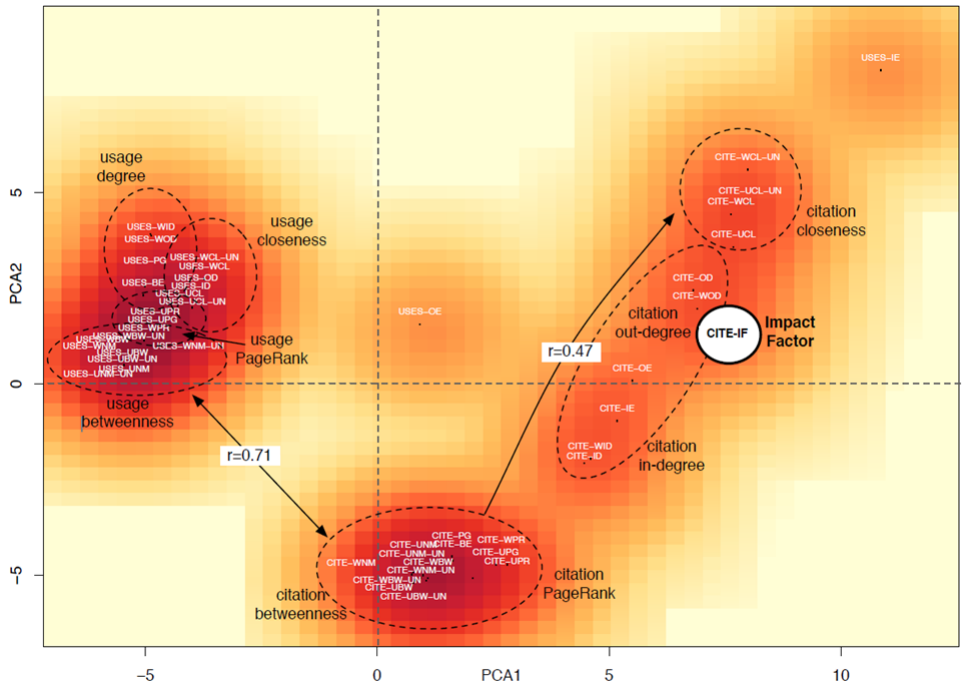

FIGURE 8.5: Principal component analysis of Spearman Rank Correlation between 47 preliminary MESUR metrics (from Bollen et al. 2008)

Figure 8.5 shows the 1st principal component and represents the largest amount of variance between the metrics. Unsurprisingly, a clear separation between the usage and citation based metrics is demonstrated, with Impact Factor being clearly located amongst the citation based metrics. Bollen observed something interesting when mapping algorithms in this way. Some usage based metrics approximate certain citation-based metrics better than some citation based metrics approximate one another. This is shown on Figure 8.5 by the two examples given in r, which represent the correlation coefficient relationship for citation-based betweenness centrality and usage-based betweenness centrality (0.71) and citation-based closeness centrality (0.47).

Bollen et al. (2008) observe that PCA2 (the vertical difference on Figure 8.5) represents a more subtle variation between closeness centrality and degree centrality (positive to negative respectively).

Also of note is the considerably higher level of agreement between results obtained from the usage data than the citation data. This results in Impact Factor seeming to be in a sparsely populated area far from the usage based metrics which it is implied it can infer! This observation alone questions the applicability of using Impact Factor as an indicator for institutions when purchasing journal subscriptions. While all the work of

Bollen et al. (2008) may not help in identifying the best bibliometric algorithm, it does help model what the current human behavioural trends are in terms of usage.

Another valid question is why is there such a perceived difference in usage based metrics to actual citations? There is a 0.71 correlation between citation betweenness/PageRank and the usage statistics and only a 0.47 correlation between these same two and the citation based metrics. Interestingly the correlation between citation metrics and usage metrics is not shown by Bollen et al. (2008).

The result obtained by the MESUR project demonstrates how algorithms behave in different ways based on both the type of the algorithm and what data is used (usage or citation data in this case). Bollen et al. (2008) show that by reducing a complex correlation matrix to a simple 2D plot using PCA, enables distance calculations between groups of algorithms to be performed easily.

In the rest of this chapter, PCA is applied to the correlation matrix obtained as a result of analysing the 10 algorithms outlined in this thesis to Citebase. Unfortunately, without the full original correlation matrix and dataset used by Bollen, the 10 algorithms outlined in this thesis cannot be mapped onto Figure 8.5. This is due to the principal eigenvectors being different for each set of algorithms. However it is possible to plot a similar heat map showing the similarities and differences of each of the 10 algorithms covered in this work.

A full guide to using Principal Component Analysis in the context of this thesis is presented in Appendix A.

8.4 Principal Component Analysis of CoRank and Variations

In order to perform PCA on the data in this thesis, the correlation matrix showing all of the relations between each pair of algorithms in Table 7.2 from Chapter 7 is required. In Section 7.2 it was stated that this matrix could be used to depict the distances and relations between the various algorithms. Principal Component Analysis (PCA) represents the best technique by which this can be done. Additionally, such analysis aligns with that undertaken by Bollen et al. (2008) as part of the MESUR project allowing some conclusions to be drawn between the two.

By reflecting all of the values in Table 7.2, a 10x10 co-variance matrix can be calculated. From this the eigenvalues and eigenvectors can be calculated using a program capable of solving the complex equations. This program (detailed in Appendix B), written in python, uses the NumPy library (a scientific calculation library) to perform all of the

4.83	3.43	1.28	2.86×10^{-1}	1.32ϵ−1	3.37ϵ-02	7.09ϵ-03	2.77ϵ-03	6.01ϵ-05	1.36ϵ-16
0.30	0.08	0.62	-0.25	-0.36	0.09	-0.41	0.12	0.26	0.26
-0.20	0.32	0.58	-0.24	0.01	-0.03	0.50	-0.10	-0.29	-0.34
0.36	-0.28	0.22	0.18	0.34	-0.05	-0.41	0.12	-0.44	-0.47
0.40	-0.24	0.13	0.17	0.03	-0.09	0.22	-0.80	0.19	0.03
-0.13	-0.51	-0.04	-0.13	-0.40	0.06	0.12	0.13	0.40	-0.59
-0.08	-0.52	0.00	-0.14	-0.40	0.05	0.12	-0.04	-0.62	0.37
0.40	0.14	-0.28	-0.39	0.05	0.74	0.11	-0.04	-0.10	-0.11
0.39	0.11	-0.29	-0.55	-0.08	-0.65	0.06	0.07	-0.06	-0.07
0.43	-0.15	0.13	0.29	0.12	-0.01	0.58	0.54	0.13	0.20
0.23	0.41	-0.16	0.50	-0.64	-0.02	-0.01	-0.01	-0.21	-0.22

TABLE 8.1: Eigenvalues (first row) and Eigenvectors (column data) corresponding to the ten metrics applied in this thesis

calculations. The program was tested using existing examples with known answers to ensure produced results are precise.

Eigenvectors are "inate" properties of a dataset, the German translation for the word eigen is "own", thus they can be used to identify the principal feature vector in a set of results. An eigen value dictates the weight of the eigenvector, like a multiplier in a quadratic equation, an eigenvalue can be used in place of dividing every eigenvector to have an eigenvalue of 1. With the eigenvectors calculated, the two with the highest eigenvalue are thus said to be the two principal components by which the data can be plotted.

With the python program also calculating the co-variance matrix, the Spearman Correlation matrix can be fed directly into the program with the results produced representing the 10 corresponding eigenvalues and eigenvectors. Table 8.1 shows the 10 resultant eigenvalues sorted in order with their 10 corresponding eigenvectors, the two left most columns of this table show the principal eigenvectors.

In Table 8.1, each value listed in the top row represents the eigenvalue, with the values below this representing the corresponding eigenvectors. With the eigenvalues sorted from left to right in decreasing orders of importance, the two principal eigenvectors are those with eigenvalues 4.83 and 3.43 respectively.

Multiplying the two primary eigenvectors by each of the co-variance values produces a series of values for PCA1 and PCA2 shown in Table 8.2. Here each subsequent row represents the new plot points for each algorithm as dictated by the principal components.

The final stage is then to plot the graph of these points, which translates all of the correlations defined in the 10x10 matrix into a 2-dimensional plot where the two axis represent the principal components of the data. This plot, shown by Figure 8.6, represents the same type of output achieved by Bollen et al. (2008), except that here it is

Algorithm	PCA1	PCA2
LC	0.23973624	0.02992122
PR	-0.0919008	0.1015423
H	0.22092888	0.06965757
CR_{LC}	0.19296469	-0.01664116
CR_D	-0.4489163	0.04598265
CR	-0.29661728	0.00383153
CR_T	-0.06051326	0.0625485
CR_{CR}	-0.03296975	-0.01759318
CR_{SLC}	0.24106223	-0.02978862
CR_{SCR}	0.03622535	-0.2494608

TABLE 8.2: Normalised PCA co-ordinates calculated from the principal Eigenvectors

only possible to plot the points corresponding to the 10 algorithms covered in this work, where full source data is available.

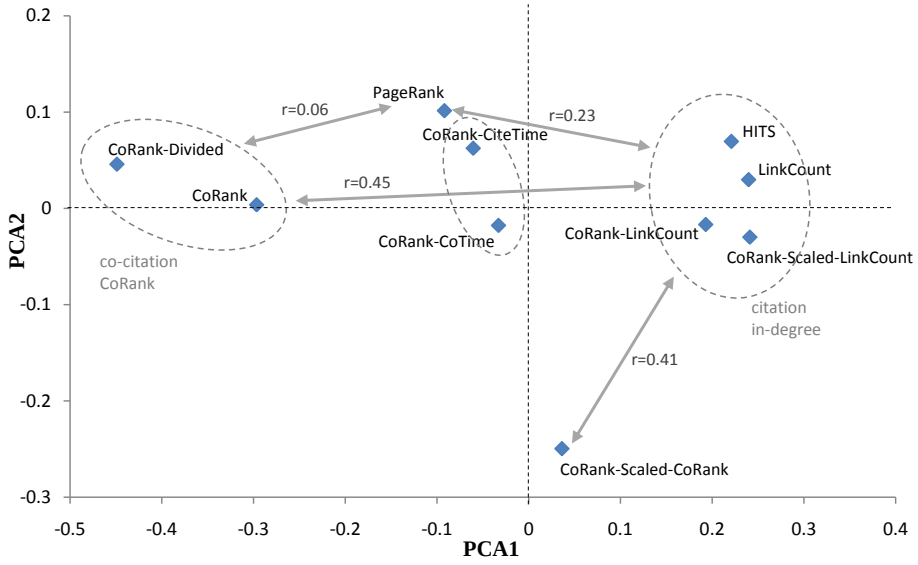

FIGURE 8.6: PCA plot of the 10 metrics detailed in this work

Much like the plot by Bollen et al. (2008), Figure 8.6 includes groupings of algorithms which show a strong correlation with one another. With the graph plotting the various metrics based upon their principal components (or inate properties), the space between each group can be said to indicate significant differences between the algorithms. Conversely any clustering of algorithms helps identify those whose principal characteristics are similar.

PCA is designed to represent the key characteristics, not the correlation between groups, a value which has been added in Figure 8.6 as it was in Figure 8.5. An example of why this is important can be seen in Figure 8.6 as the difference between PageRank and the Co-citation CoRank algorithms and the Citation In-Degree group. In this figure, the correlation differences between PageRank and the two groups (0.05 and 0.23) does not

approximate the overall difference between the two groups themselves (0.45). The main difference is that PageRank is very weakly related to the co-citation CoRank group.

Figure 8.6 shows groupings of algorithms similar to those observed in the conclusion of Chapter 7. These groups represent the citation based metrics which achieve high impact (including Citation Count and CoRank-LinkCount), co-relation based metrics and the time-based metrics. When previously classifying algorithms into groups (Section 7.7), some challenges were faced when classifying PageRank into one of these groups, and this can be observed again in Figure 8.6 where PageRank sits alone.

Up to this point the time based algorithms (shown to not be statistically significant in Section 7.6) have been left in place, in order to demonstrate their effect on the various metric plots. Choosing to remove the time based metrics, may result in the principal components changing, thus to remove these metrics, the whole PCA calculation process has to begin again from the start.

Skipping over all of the complex calculation stages to the result gives a pair of principal eigenvalues equalling 3.85 and 3.16 respectively (the next closest eigenvector has an eigenvalue of 0.848). These two primary eigenvectors can be said to be substantially primary due to the distance from the next closest value. Re-plotting the results according to their principal eigenvectors gives the plot depicted in Figure 8.7.

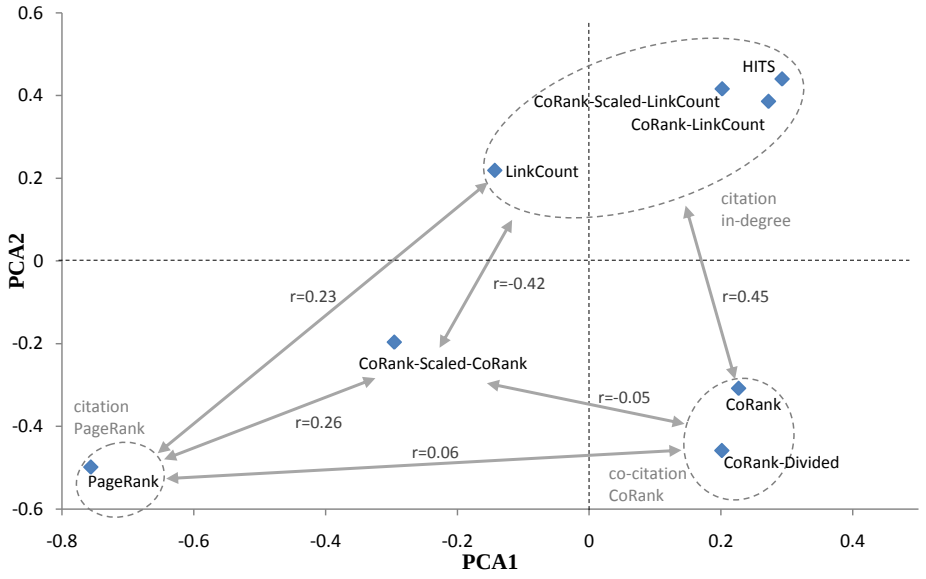

FIGURE 8.7: PCA plot excluding the time based algorithms

In Figure 8.7 the same groupings of metrics can still be observed as previously. The exception here is LinkCount, which has moved very slightly away from the other three algorithms in this group, due to the fact that LinkCount is actually fairly weakly correlated to CoRankScaled-LinkCount (again not 100% clear in the plot), but still shares many of the behavioural characteristics. Figure 8.7 also shows clearly that PageRank

stands very much alone in this network. This can be attributed to the fact that PageRank was the only algorithm of its type (using weighted direct citations) applied during this thesis. In the work by Bollen et al. (2008), a number of algorithms are utilised similar in nature to PageRank, thus there is good basis on which to justify giving PageRank its own group.

Overall, PCA1 and PCA2 are not as clearly different as discovered by Bollen et al. (2008). With the graph showing two distinct eigenvalues, 3.85 and 3.16, these are not substantially far enough apart to state that one principal component is dominant over the other in Figure 8.7. Including the time based algorithms, ironically gives a clearer result and PCA1 can be stated to represent the largest variation in correlation between the algorithms. In Figure 8.6, PCA2 represents a much more subtle variation with algorithms based upon Co-Citations holding a lesser PCA2 value.

In both Figures (8.6 and 8.7) the principal components of the CoRank based algorithms leave them some way apart from both the citation in-degree algorithms and the citation PageRank algorithms, further confirming the conclusions and algorithm groupings outlined in Chapter 7. Comparing these figures with Bollens PCA plot in Figure 8.5, it is likely also that the CoRank family of metrics would form a new group some way from any usage based metrics as well, unfortunately without having any source data on which to try out this hypothesis, it has to remain that.

This thesis has thus provided another means to re-affirm the conclusions of Chapter 7, showing that the CoRank algorithms are in fact a new family of algorithms and do not map to the characteristics of algorithms based on Citation Count. This means that through the definition of "better" outlined in this work, CoRank was never going to satisfy all conditions, however CoRank-LinkCount proved to be a real contender in the category of citation based metrics. Figure 8.7 shows how CoRank-LinkCount clusters more closely with the traditional Citation Count and HITS metrics rather than the CoRank based algorithms. Along with the early correlation, high mean impact score and benefit in revealing a number of more recent publications, CoRank-LinkCount does show good performance in all testing categories.

8.5 Conclusion

Measuring impact, although simple sounding, is much like many other statistical techniques. A series of complex observations are made in an attempt to quantify human behaviour, at the end of which, all the data is distilled down into a number or a simple graph. The challenge is to use a calculation which accurately depicts these trends and gives the "correct result".

Throughout this thesis, the changing nature of scholarly communications has been challenged to ascertain whether the current practices are ideally suited to the study, or simply just too embedded to allow consideration of other techniques. For example, it was observed how different subject areas have preference over different publishing mediums. Fortunately, studies comparing the journal citation metrics have found strong correlation to applying these same metrics to non-journal publications. Changes in publishing trends should be observed carefully along with the potential to utilise different metrics in order to quantify this changing behaviour.

Through the study of ten different metrics, mainly based around the principal of using co-citation data as input, this thesis has addressed the potential for new types of metric to be used to study changing behaviour. The idea behind using co-citation data is based on the observation that the speed of research has been increasing as scholarly dissemination techniques improve (e.g. the Web). Additionally and equally significant is the observation that the co-citation network grows faster and contains more established publications than the citation network.

By taking the Citation Count as the current accepted standard used for judging the impact of individual articles, the idea was to examine the suitability of algorithms based upon co-citation to provide a surrogate measure for later impact (an early indication metric). The result of this study concluded that at no point during the life cycle of the publication could the original CoRank algorithm be said to provide an accurate surrogate measure. However, by studying several derivatives of the algorithm, CoRank-LinkCount, a simplification of CoRank, was found to perform in a manner similar to Citation Count and provide some early suggestion of impact.

CoRank-LinkCount clearly shows the biggest potential when compared to Citation Count (LinkCount) even though the difference is not itself huge. Due to the removal of the weighting factor from CoRank, this algorithm is thus based more closely on Citation Count than weighted algorithms such as PageRank. CoRank-LinkCount does maintain the benefits provided by the Co-Citation network and therefore is able to provide an early indication of subsequent impact from very little citation data. In both this and the previous chapter, CoRank-LinkCount associates itself very strongly with neighbouring algorithms including Citation Count, HITS and CoRank-Scaled-LinkCount.

Conversely PageRank, the Webs most prolific algorithm, never ends up being grouped together with any other algorithms. Removing the weighting factor from CoRank had such a profound effect on the results, that leaving it in, as PageRank does demonstrates the significance of this characteristic in the algorithm. It turns out that one of the most important characteristics of PageRank, designed to improve accuracy of results, does not map as well as expected to a network of peer-reviewed publications. In a similar study, Thelwall (2003) analyses the relationship between link counts to University websites and their PageRank and finds very little correlation between the two. He concludes that

PageRank is not very useful as a stand alone metric when used for measuring the value of such information and must be combined with other factors in order to obtain the quality of results provided by services such as Google. This thesis confirms that much the same is true of a network of publications and their citation links.

CoRank and those algorithms based upon CoRank also clearly stand apart in their own family. Showing low levels of positive correlation while maintaining a good number of more recent publications in the top 5% shows them to be significantly different again to both Citation Count and PageRank algorithms. The key differentiator from PageRank lies in the fact that the CoRank algorithms stabilise the average rank of publications much faster than PageRank. With PageRank relying on the citation data, which takes substantial time to establish itself, it is no surprise that the PageRank of any single publication is going to take even longer to establish itself.

It is unfortunate that the set of CoRank algorithms cannot be mapped back into the study by Bollen et al. (2008), as this may have revealed their overall position in their ecosystem of 47 other metrics. It would be interesting to see if CoRank maintains a separation from other metrics, and which sets it relates more closely too. Out of all the algorithms, it is CoRank-LinkCount which provides the best potential to became a surrogate measure for later impact by Citation Count. Throughout this thesis, only those publications which were considered to be high impact by Citation Count have been considered, and in these tests CoRank-LinkCount performed well, giving a high correlation and mean rank to these sets of publications throughout their life cycle. CoRank-LinkCount also manages to reveal a marginal number of newer publications, which have gained a good number of citations, suggesting its applicability as an early indication metric.

Chapter 9

Concluding Remarks and Future Directions

The study of bibliometrics provides an important mechanism for the identification and classification of resources. The early pioneering work of Garfield (1973) in the area of citation metrics enabled the scholarly community to quickly access the popularity of others work in any field of research. In subsequent years, measures of popularity including Citation Count and Impact Factor have also been used directly as measures of prestige (Garfield 2005, Moed 2009). Pinski & Narin (1976) were the first to realise a key difference between popularity and prestige in the area of scholarly communications; an important factor when performing critical assessment exercises.

In modern society, critical assessment can play an important role in funding opportunities and job security. Projects and people are assessed based upon their success, and bibliometrics provide a large number of techniques to aid in this area. For early career researchers, factors such as Citation Count and h-index which both take some time to establish, should not be used as performance indicators. Similarly, research projects are not able to be judged from the impact of their outputs. Open Access (OA), a paradigm which has been enabled by the Web, has been shown to reduce the time frame between publication and first citation (Eysenbach 2006). Download counts, which can be applied to both OA and non-OA materials, have also shown promise as good early indication metrics (Brody & Harnad 2006). It is these early indication metrics which provide the only technique for aiding the assessment of recent research.

Garfield realised the importance of peer review in the area of scholarly communications and suggested that citation analysis could be used to help judge impact (Garfield 1955). Peer review provides credibility to published works and improves research performance (Goel & Faria 2007). A citation demonstrates an author's intellectual honesty in their own work and is often used as a method to reference background work or assist in backing up any points the author is making. It is the combination of these statements which

make peer review and citation better to use when measuring prestige, than factors such as download count, which can only measure popularity. When designing a new early impact metric in order to measure prestige, this thesis examined the potential to utilise a publication co-citation network, something built directly via the citation network.

Each time a publication is cited, it is also cited alongside a good number of others; thus the co-citation network of relations between papers builds much quicker than the citation network, and relates together a greater number of publications. Additionally, unlike on the Web, a citation can only be created by a newer publication. A co-citation is able to exist between the publication in question and a number of much more established publications. It is these two factors which make the co-citation network a good candidate to use as the source of information when ranking publications earlier in their life cycle.

Pinski & Narin (1976) formed a theory that a popular publication is highly cited, while a prestigious publication is cited by other prestigious publications. On the Web, Hubs and Authorities (Kleinberg 1999) and PageRank (Brin et al. 1998), make the same realisation and apply it to the Web graph where a citation to a new item can be made from a high profile existing website. By looking at the co-citation network, this thesis investigated whether similar principals can be applied to make a new early impact indicator for publications.

Taking influences from existing metrics, including Citation Count and PageRank a number of metrics were introduced and tested against the Citebase dataset. Citebase, designed to be the "Google for the refereed research literature" (Hitchcock, Brody, Gutteridge, Carr, Hall, Harnad, Bergmark & Lagoze 2002), was intended to help demonstrate the benefits which come from not only open access publishing, but also open access citation data. Indexing over 3.3 million citations at the time of study, made Citebase the ideal candidate for application of a number of new and existing metrics. With 3.3 million citations amounting to 46 million co-citations, the differences in amount of available links in the two networks is clear. With this amount of data and the snapshots required such that early indication potential could be examined, a system needed to be designed, capable of applying multiple metrics to large graphs in a timely fashion.

The Co-Ordinator system represents a novel approach to distributed processing of data in an environment where services can be provisioned and combined to provide different levels of storage, processing and speed. Through separation of these parts, application of the metrics themselves could be carried out using a high processor module plugged to a fast database system. Likewise when processing results, the demand on the processor module is less, while the storage and database modules remain in demand. Not only does this dynamic approach make more effective use of resources, it also ensures the longevity of the system as the processor module can be updated independently of the rest of the system. This made the Co-Ordinator the ideal system to apply any number

of metrics to any co-related dataset. With minor changes the Co-Ordinator could be used to apply many other metrics to other types of dataset in the same manner.

This thesis examined three hypotheses relating to co-citation data and early indication metrics:

- Co-Citation relations can be used to create an early indication metric for publication impact which correlates well with existing metrics.
- A metric based on co-citations will identify high impact publications sooner in their lifecycle.
- Applying co-citation metrics in search ranking will promote more recent publications.

All three of these hypotheses could be examined using the Co-Ordinator system. All 10 of the metrics, both new and existing, were applied to each Citebase snapshot by the Co-Ordinator, which then summarised the results in a way such that the three hypotheses could be examined. This process was cross-checked using a broad series of data selected by the target metric, Citation Count. In each target snapshot the top papers, rated by Citation Count, were selected and rank order and position recorded. The results of applying each algorithm to all snapshots was examined to find any strong correlation to this target data, thus fulfilling the need of the first hypothesis. To test the second hypothesis, the rank positions from each snapshot were recorded and compared to those from future snapshots by Citation Count. Finally, the age of the top 5% of publications in each snapshot was examined to determine the average age being revealed in this sample by each algorithm. The expected result was that the average age of a good early indication metric should be lower.

In order to look at prestige and not popularity, both PageRank (applied to the citation network) and the variant introduced in this work CoRank (applied to the co-citation network), were examined on the Citebase data. Both metrics apply a weighting factor on the source of citation or co-citation respectively, giving CoRank a distinct advantage over PageRank in citation networks. As links (or citations) in the scholarly communications network can only exist between newer and older material, the prestige of the newer publication will take some time to establish, so too will a publication's PageRank score; a fact reflected clearly in the results. By applying the same principal, but to a publication's co-citation network, CoRank will be able to rank a publication more highly, earlier in its publication life cycle. When comparing CoRank to PageRank this hypothesis holds true and CoRank satisfies all tests, except for the demand to be rank order correlated to Citation Count.

Even though CoRank was able to exhibit some benefit over PageRank, it was still a long way from the results exhibited by Citation Count. Subsequently a number of

refinements and extensions to the CoRank algorithm were applied in order to further examine features of the co-citation network, in the hope of finding an early indication metric based upon it. By changing the main features of the algorithms, a number of families of metrics were created, all with their own leading characteristics, including some which take publication age into account (later found not to be statistically significant in results). By removing the weighting factor, a closer relation with Citation Count was re-established and CoRank-LinkCount became the algorithm which performed best when compared to Citation Count.

CoRank-LinkCount represents a simplification of the CoRank algorithm and works by looking at the number of citations towards publications with which the publication in question is co-cited and not the number of publications towards the publication itself. This metric exhibited a good positive correlation with Citation Count rank order, but more significantly was able to rate publications much higher than any other algorithm, sooner in the publication life cycle. Over the first 12 months after publication, CoRank-LinkCount is consistently able to identify subsequent high impact publications. After 12 months Citation Count has also identified this same set and both metrics show a very similar rank, from this point forward with CoRank-LinkCount showing slightly lower rank as it is already revealing the next set of highly ranked publications. So while CoRank-LinkCount did not completely satisfy the first hypothesis, it was the best all round performer and satisfied the requirement to reveal subsequent high impact publications sooner in the publication life cycle.

CoRank and the simplified CoRank-LinkCount are two examples of completely different metrics, based upon their principal factors of one being weighted and the other not. This has a major effect on the results produced by each, as does the application of PageRank. By applying Principal Component Analysis (PCA) to the correlations between all the metrics trialled in this thesis, these distinct families of metrics, first identified when looking at publication rank order results, became much clearer. PCA, a method applied to similar usage based metrics by Bollen et al. (2008), revealed three groups of metrics (when ignoring the non statistically significant time based metrics). The first group contained both the target metric Citation Count and the best relative performer CoRank-LinkCount, showing the similarities in these algorithms even when based upon different networks of data. The second group contained all the weighted CoRank algorithms, those taking into account some form of prestige of each co-cited publication. What is surprising is that while CoRank-LinkCount and Citation Count are closely related, CoRank and PageRank are not by this evaluation; PageRank sits on its own, disconnected from the other two groups even though the correlation distance from CoRank is small. This significant difference is reflected in the results where, due to the characteristics of PageRank it is never likely to reveal any recently added material. Being based on the co-citation network, CoRank is able to perform better than PageRank in these areas thus distinguishing it when used on a citation network.

In related studies, some have found PageRank to be highly correlated to currently accepted impact measurements in certain fields (Chen et al. 2007, Ma et al. 2008). With CoRank performing better than PageRank, there is still opportunity to apply the original CoRank algorithm in these same studies and examine if similar benefit can be gained in these areas as found in this thesis.

With bibliometrics studying human behaviour, what is clear is that people in different fields of study have different practices and publishing methods. Any evaluative study and metric applied should be able to take account of localised behaviour, meaning that one metric is unlikely to be able to measure the same behaviour in all areas. Additionally data sources are changing and the Web has had a profound effect on how people find and consume information. Having a vast array of metrics and data sources is providing huge opportunity in the area of bibliometric study and this thesis has covered only a small number of the possible families of metrics.

9.1 Possible Future Directions

In this thesis, the originally proposed CoRank metric evaluated poorly against the initial criteria, set with the aim of attempting to better the widely used Citation Count standard. Likewise, it was also found that PageRank performs poorly, something which is logical when examining a newer citation network. By relaxing the stringent correlation criteria from the first hypothesis, whilst introducing other factors, may see significant benefits revealed that Citation Count does not hold. These could also help identify other potential scenarios where each family of metrics is perhaps best applied.

One way in which the correlation could be adjusted was briefly looked at in Section 7.5. Here it was observed that readers identify acceptable boundaries between groups of publications based upon Citation Count. Exact order is not as important when publications are classified as highly, average or lowly cited. Utilising groups rather than correlations still resulted in CoRank-LinkCount performing best of the new metrics, while the more general CoRank family were still not as positive in performance, suggesting the correct result was discovered either way.

Taking the idea of grouping articles dependent on perceived impact could lead to an adaptation of the author h-index being created to cover individual articles in a collection, the "a-index". Here a set of publications, possibly all in the same journal, could be evaluated to find the number of citations they each obtain, a plot or distribution of the number of citations could then reveal the critical "a-index" value. This value could then be used to give a figure for what constitutes the most influential articles rather than the best authors.

Another approach would be to continue developing or refining the applied metrics, investigating the effects of different factors. Yan & Ding (2011) and Dellavalle et al. (2007) applied a weighted version of PageRank to author networks and dermatology and show very positive results when compared against their chosen set of criteria. Although application of temporal factors produced low quality, non statistically significant results, Maslov & Redner (2008) discusses the importance of such factors in the area of scholarly communications showing there is still potential here for further investigation.

The importance of not using a factor such as Citation Count as the sole indicator of prestige is clear from many studies (Garfield 2005, Moed 2009). This is also true online for services such as Google, who combine metrics with text mining and other bibliometric techniques in order to generate results. Such hybrid measures represent another potential way to automate the process of discovering prestigious material in the field of scholarly communications. Such measures could also be customised and weighted for each field of research if such faceted services and capabilities were available. Applying faceted techniques could also help identify citation boundaries in the different fields and help process data from publication sources other than journals.

Closer to the work carried out in this thesis would be the extension to evaluate whole other families of metrics on the same data using the flexible framework provided by the Co-Ordinator. A much closer and direct comparison could then potentially be drawn with the work of Bollen et al. (2008), outlined in Chapter 8, if similar metrics are chosen.

Conversely, input data could be changed to cover more than just the data provided by Citebase. Evaluating modern publishing trends revealed that scholarly communications is opening up to new forms of publication, not necessarily harvested by services such as Citebase. Evaluating the different metrics against these areas, with possibly relaxed criteria, may result in more effective application of the CoRank algorithm or confirm the effectiveness of the CoRank-LinkCount algorithm further.

Extending outside the area of publication data, the Co-Ordinator system and the family of metrics could also be used to evaluate different co-relation based datasets. A logical extension would be to apply CoRank to the links which exist between Web pages. On the Web, PageRank represents the most widely used indicator, thus the evaluation criteria would be different. In this situation it would certainly be interesting to see how CoRank-LinkCount performs, or if the ability to weight citation links is so essential, that the original CoRank algorithm will perform much better in this scenario.

In addition to the Web, co-relation networks can be observed between people. Taking the social network of links between people as stabilised on Twitter, for example, the concept of a re-tweet could be mapped closely with that of a citation. Currently there are no widely adopted measures for judging influence of people in such networks and the Co-Ordinator system and methodology outlined in this thesis could easily be applied in these areas. Newman (2003) provides a review of work carried out to understand and

predict behaviour in such networks, showing the different influences of strong and weak links between nodes in a network and how this affects clustering and separation.

9.2 Final Remarks

The main novel contribution of this thesis is an investigation of the effectiveness of introducing co-citations in citation based ranking metrics. In order to achieve this, a new family of metrics were proposed, which focused on utilising the benefits provided from the much larger and well established network of co-citations. The Co-Ordinator system was built to process this network of co-citations and allow application of any number of metrics. Outputted results could then be evaluated against a set of test criteria.

The overall aim of this thesis was to try and improve citation metrics through early identification of subsequent high impact publications. A wide variety of both new and existing metrics were evaluated against a large body of real literature. The results presented revealed the importance of different characteristics in metrics which define their success against different criteria. When looking for a co-citation based metric to be an early indicator of citation impact, a clear link was required between any new algorithm and Citation Count in order to make this possible. Similarly for weighted metrics, this work discovered that benefits over existing techniques can also be gained via sourcing input data from a co-citation network.

The work for this thesis has provided a mechanism to evaluate competing metrics in a context where the measurement of academic impact has become increasingly important. By creating an environment in which novel techniques and refinements there of can easily be investigated, the outputs of this work can support the development of an evidence base to satisfy the requirements of the academic and further research community.

Appendix A

A Brief Guide to Principal Component Analysis

Principal Component Analysis (PCA for short) requires a dataset containing at least two sets of results. By calculating the co-variance of these results it is then possible to calculate the eigenvalues and eigenvectors. Once calculated, the principal eigenvectors, dictated by the eigenvalues, can be used to transform the original data to enable it to be re-plotted according to its principal components. Due to the fact that this is quite a complex process on any matrix larger than 3×3 this section gives a worked example based upon the excellent tutorial given by Smith (2002).

Principal Component Analysis is about identifying multi-dimensional patterns and reducing these to a simple 2 dimensional space. Essentially it is another method for finding the correlation between datasets, such that direct, indirect and zero relations can be identified visually. PCA's major benefit becomes clear when trying to reduce multivariate data, with corresponding results given in matrix form (bigger than 2×2) down into something which can be plotted on a 2D graph where each axis represents one of the principal factors affecting the results.

In an example however, working through anything larger than a 2×2 matrix will require stages to be skipped, which are calculated using algorithms on a computer. For this reason, the input data used here consists of some hypothetical heights and weights, which are correlated slightly to show that as people get taller they also get heavier. This sample data is shown in Table A.1.

The first stage of PCA is to adjust the data such that the mean of each column is 0. This can simply be done by subtracting the mean from each column, thus 170 from the Height column and 10.9 from the Weight column. This gives the "DataAdjust" table also shown in Figure A.1.

Data =

Height (X)	Weight (Y)
184	13
163	10
172	11
154	9
166	9
180	10
160	10
177	14
176	12
168	11

DataAdjust =

$X - \bar{X}$	$Y - \bar{Y}$
14	2.1
-7	-0.9
2	0.1
-16	-1.9
-4	-1.9
10	-0.9
-10	-0.9
7	3.1
6	1.1
-2	0.1

FIGURE A.1: Sample input data for PCA calculation

The next stage is to calculate the co-variance matrix for this data. Co-variance is a derivative of the variance calculation, which is in turn a derivative of the Standard Deviation function. Standard Deviation measures the spread of data in a single dataset. This is calculated using Equation A.1, resulting in the standard deviation for the height data being 9.49cm and 1.66 stone for weight.

$$s = \sqrt{\frac{\sum_{i=1}^{N} \left(X_i - \bar{X}\right)^2}{(n-1)}} \tag{A.1}$$

Variance is another measure for the spread of data which is calculated as simply the Standard Deviation squared and is represented by the symbol s^2:

$$s^2 = \frac{\sum_{i=1}^{N} \left(X_i - \bar{X}\right)^2}{(n-1)} \tag{A.2}$$

Variance is still at this point a 1-dimensional equation; it can only operate over height or weight in the case of our data. Co-variance is the simple extension to variance which provides a technique through which one can find out how much the results vary from the mean with respect to each other. Co-variance is always measured between two dimensions, so if a 3-dimensional was added the the input dataset (representing age for example), then a 3×3 matrix of co-variance results would be obtained rather than a 2×2 one. Adapting the variance equation into the co-variance equation is achieved through expanding the $\left(X_i - \bar{X}\right)^2$ brackets and changing the second set of $\left(X_i - \bar{X}\right)$ to $\left(Y_i - \bar{Y}\right)$ as shown by Equation A.3.

$$cov(X,Y) = \frac{\sum_{i=1}^{N} \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{(n-1)} \tag{A.3}$$

It is also worth noting that if X and Y are the same dataset then the co-variance can be calculated using the variance algorithm as X and Y will be the same.

By re-labelling the height and weight columns as X and Y respectively and working out $\left(X - \bar{X}\right)$ and $\left(Y - \bar{Y}\right)$, the products $\bar{X}\bar{Y}$, $\bar{X}\bar{X}$ and $\bar{Y}\bar{Y}$ can be calculated, as shown in Table A.1. This table also shows the sums of these three product columns and the respective co-variance, which is calculated by dividing this sum by $n-1$ (9 in this case).

X	Y	$X - \bar{X}$	$Y - \bar{X}$	$\bar{X}\bar{Y}$	$\bar{X}\bar{X}$	$\bar{Y}\bar{Y}$
184	13	14	2.1	29.4	196	4.41
163	10	-7	-0.9	6.3	49	0.81
172	11	2	0.1	0.2	4	0.01
154	9	-16	-1.9	30.4	256	3.61
166	9	-4	-1.9	7.6	16	3.61
180	10	10	-0.9	-9	100	0.81
160	10	-10	-0.9	9	100	0.81
177	14	7	3.1	21.7	49	9.61
176	12	6	1.1	6.6	36	1.21
168	11	-2	0.1	-0.2	4	0.01
			Sum	102	810	24.9
			cov	$11.\dot{3}\dot{3}$	90	$2.7\dot{6}\dot{6}$

TABLE A.1: Co-Variance calculations for sample PCA data

These covariances can thus be represented in a 2×2 matrix as shown by Equation A.4.

$$cov = \begin{pmatrix} 90 & 11.\dot{3}\dot{3} \\ 11.\dot{3}\dot{3} & 2.7\dot{6}\dot{6} \end{pmatrix} \tag{A.4}$$

With both the diagonal elements in this covariance matrix positive, it should expected that both X and Y increase together; showing a direct correlation.

Having established the covariance matrix, the final stage is to calculate the eigenvalues and eigenvectors.

Eigenvectors are "inate" properties of a dataset, the German translation for the word eigen is "own", thus they can be used to identify the principal feature vector in a set of results. An eigen value dictates the weight of the eigenvector, like a multiplier in a quadratic equation, an eigenvalue can be used in place of dividing every eigenvector to have an eigenvalue of 1.

Since the data chosen produces quite a complex co-variance matrix with values containing many decimal places, it will be easier to demonstrate the calculation of eigenvectors and eigenvalues using a simplified matrix. The following example comes with thanks to

Dr. E. Garcia's excellent example[1], which as shown in Equations A.5 and A.6, has been simplified and then extended to include all the factors important in Principal Component Analysis.

$$A = \begin{pmatrix} 13 & 5 \\ 2 & 4 \end{pmatrix} \tag{A.5}$$

$$|A| = 13 \times 4 - 2 \times 5 = 42 \tag{A.6}$$

In Equation A.5 the values 13 and 4 represent $\bar{X}\bar{X}$ and $\bar{Y}\bar{Y}$ thus by subtracting the scalar matrix from matrix A, a quadratic equation can be formed as shown in Equation A.6.

$$cl = \begin{pmatrix} c & 0 \\ 0 & c \end{pmatrix} \tag{A.7}$$

The Scalar Matrix (Equation A.7), can then be subtracted from the matrix **A** (Equation A.5) to give the result shown below (Equation A.8).

$$A - cl = \begin{pmatrix} 13 - c & 5 \\ 2 & 4 - c \end{pmatrix} \tag{A.8}$$

From this matrix the determinate can be found (Equation A.9) which can then be re-arranged to give a solvable quadratic equation (Equation A.10).

$$|A - cl| = (13 - c) \times (4 - c) - 5 \times 2 = 0 \tag{A.9}$$

$$c^2 - 17c + 42 = 0 \tag{A.10}$$

Finally solving the quadratic equation gives the eigenvalues, here 3 and 14. Thus the higher eigenvalue dictates the principal eigenvector which can be calculated by simply substituting c=14 back into the matrix **A - cl** in Equation A.11.

$$A - cl = \begin{pmatrix} 13 - 14 & 5 \\ 2 & 4 - 14 \end{pmatrix} = \begin{pmatrix} -1 & 5 \\ 2 & -10 \end{pmatrix} \tag{A.11}$$

[1] Matrix Tutorial 3: Eigenvalues and Eigenvectors - http://www.miislita.com/information-retrieval-tutorial/matrix-tutorial-3-eigenvalues-eigenvectors.html

Using p and q to represent $\begin{pmatrix} p \\ q \end{pmatrix}$, the possible eigenvectors can be calculated possible eigenvectors pq as follows in Equation A.12.

$$\begin{aligned} -1p + 4q &= 0 \\ 2p - 10q &= 0 \end{aligned} \tag{A.12}$$

Which solves to give the following results:

p	5	10	15	...
q	1	2	3	...

Thus $\begin{pmatrix} 5 \\ 1 \end{pmatrix}$ is the principal eigenvector, when the eigenvalue is 14.

Following the steps outlined above for the dataset of heights and weights in a spreadsheet (where the calculation can be carried out with huge amounts of decimal places) gives eigenvalues of 91.45 and 1.32 (both to 2 d.p.) with eigenvectors of $\begin{pmatrix} -0.13 \\ 0.99 \end{pmatrix}$ and $\begin{pmatrix} 0.99 \\ 0.13 \end{pmatrix}$ respectively. The fact that these eigenvectors are very similar is purely an effect of them being rounded to 2 decimal places. However it is not until 12 decimal places are considered that the values differ from the +/-0.99 and +/-0.13 values.

The final stage of Principal Component Analysis is to transform the source data so that it can be represented it in terms of its two principal eigenvectors. This can be achieved by multiplying each data point by the corresponding eigenvector as per Equation A.13.

$$TransformedData = EigenVector \times DataAdjust \tag{A.13}$$

Applying this to all of the pairs of values in the height/weight dataset gives the values shown in Table A.2).

As an example of a full calculation (shown in A.14), the first pairs of PCA values (14.15 and 0.31) in Table A.2 are calculated as follows from the (14,2.1) coordinate in the DataAdjust table (Figure A.1):

$$\begin{aligned} PCA1 &= (14 \times 0.99) + (2.1 \times 0.13) = 14.15 \\ PCA2 &= (14 \times -0.13) + (2.1 \times 0.99) = 0.31 \end{aligned} \tag{A.14}$$

Even for this small dataset, it can be observed that PCA1 and the original $X - \bar{X}$ values are not that much different, where as $Y - \bar{Y}$ has been reduced substantially. This will

PCA1	PCA2
14.15	0.31
-7.06	-0.01
2	-0.15
-16.11	0.14
-4.21	-1.38
9.81	-2.16
-10.03	0.37
7.34	2.19
6.09	0.33
-1.97	0.35

TABLE A.2: Translated PCA output data

be due to PCA identifying that the X values are more significant and represent the principal trend in this dataset.

Obviously a 2-dimensional dataset is only going to have two eigenvalues and eigenvectors, thus the two principal vectors are the only two. Thus when plotting the data back onto a 2D plot (as it could be in the first place), the original results have simply been translated (or rotated) around 0,0 to a direction dictated by the two principal components. Figure A.2 shows a plot of both the original data with the translated plot shown alongside.

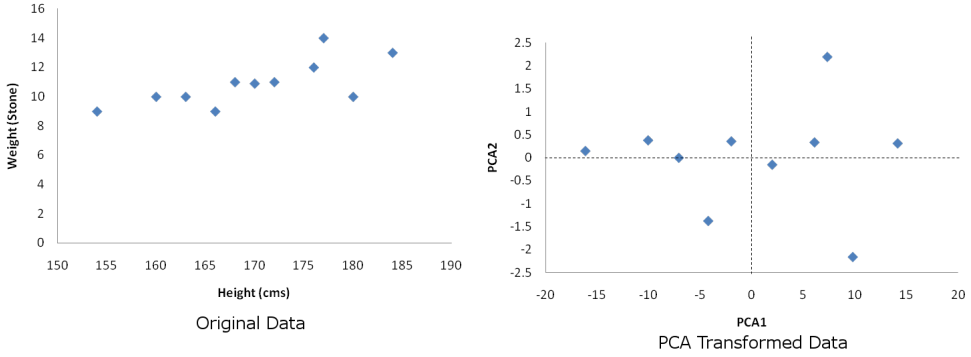

FIGURE A.2: PCA transformation example plots of both original data and PCA transformed output

In a dataset where the data consists of multivariate data, PCA can be used to plot the data according to the two principal factors (or components). These principal components do not always represent the two factors which most effect the results. As with the example in Figure A.2, a trend can be seen where the results vary widely in PCA1 but very little by PCA2, thus it could be said that PCA1 is the most significant. The fluctuation from 0 by PCA2 is showing the outliers are much clearer than in the original data plot. These points can be seen clearest when PCA1 is 10. As with all 2-dimensional PCA plots, each point corresponds to a single point from the original results.

Applying PCA to the results of the 10 algorithms outlined in this thesis will help to both group the algorithms and measure the distances between each group. PCA1 and

PCA2 will, as a result, help define the major factors influencing the algorithms studied in this thesis and help align the result, with the work of others.

Appendix B

Python PCA Eigenvectors and Eigenvalues Program

The following piece of code takes the 10×10 correlation matrix (from a file on disk) and performs the Principal Component Analysis as outlined in Appendix A. The resultant output consists firstly of the co-variance matrix followed by the eigenvalues and related eigenvectors. Thanks go to my friend and colleague James Morse for helping me put this code library together.

```
#!/usr/bin/python
import os
import sys
import numpy as np
from numpy import linalg as LA
from numpy import *

if len(sys.argv) < 2:
        print "Usage:",sys.argv[0], "<csv-file>"
        sys.exit(1)

inputFile = sys.argv[1]
if not os.path.exists(inputFile):
        print "No such file or directory:", inputFile
        sys.exit(1)

# import the data
raw_results = np.loadtxt(inputFile, delimiter=',', usecols=(0, 1, 2, 3, 4,
              5, 6, 7, 8, 9), skiprows=1)

# build two arrays of all the raw values
raw_results_rot = raw_results.T
raw_p = raw_results_rot[0]
raw_q = raw_results_rot[1]
raw_r = raw_results_rot[2]
raw_s = raw_results_rot[3]
```

```
raw_t = raw_results_rot[4]
raw_u = raw_results_rot[5]
raw_v = raw_results_rot[6]
raw_w = raw_results_rot[7]
raw_x = raw_results_rot[8]
raw_y = raw_results_rot[9]

# normalise the two arrays
means = ((raw_p.mean(),raw_q.mean(),raw_r.mean(),raw_s.mean(),raw_t.mean(),
        raw_u.mean(),raw_v.mean(),raw_w.mean(),raw_x.mean(), raw_y.mean()))
data_p = raw_p - raw_p.mean()
data_q = raw_q - raw_q.mean()
data_r = raw_r - raw_r.mean()
data_s = raw_s - raw_s.mean()
data_t = raw_t - raw_t.mean()
data_u = raw_u - raw_u.mean()
data_v = raw_v - raw_v.mean()
data_w = raw_w - raw_w.mean()
data_x = raw_x - raw_x.mean()
data_y = raw_y - raw_y.mean()

# Calculate the covariance matrix
covar_matrix = np.corrcoef((data_p,data_q,data_r,data_s,data_t,data_u,
                            data_v,data_w,data_x,data_y))

print "COVARIENCE\n"
print covar_matrix,"\n"

# Calculate the eigenvalues and eigenvectors
eigenvals,eigenvects = LA.eig(covar_matrix)

# sort the eigenvalues and eigenvectors
ind = argsort(-eigenvals)
eigenvals = eigenvals[ind]
eigenvects = (eigenvects[:, ind]).T

print "eigenvals\n"
print eigenvals,"\n"
print "eigenvects\n"
print eigenvects,"\n"
```

References

Aksnes, D. & Taxt, R. (2004), Peer Reviews and Bibliometric Indicators: A Comparative Study at a Norwegian University, *Research Evaluation* 13(1), 33–41.

Alpert, J. & Hajaj, N. (2008), We knew the web was big..., *The Official Google Blog*. Available at `http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html`

Antelman, K. (2004), Do Open-Access articles have a greater research impact?, *College and Research Libraries* 65(5), 372–383.

Bakkalbasi, N., Bauer, K., Glover, J. & Wang, L. (2006), Three options for citation tracking: Google Scholar, Scopus and Web of Science, *Biomedical Digital Libraries* 3(1), 7.

Bergman, M. (2001), The deep web: Surfacing hidden value, *Journal of Electronic Publishing* 7(1).

Bergstrom, C. (2007), Eigenfactor: Measuring the value and prestige of scholarly journals, *College and Research Libraries News* 68(5), 314–316.

Bergstrom, C., West, J. & Wiseman, M. (2008), The Eigenfactor Metrics, *Journal of Neuroscience* 28(45), 11433–11434.

Bergstrom, T. & Bergstrom, C. (2004), Will Open Access compete away monopoly profits in journal publishing, *Working Paper*.
Available at `http://octavia.zoology.washington.edu`

Berners-Lee, T. (1989), HyperText and CERN, *A information management system proposal at CERN*.
Available at `http://www.w3.org/Administration/HTandCERN.txt`

Bjorneborn, L. & Ingwersen, P. (2004), Toward a basic framework for webometrics, *Journal of the American Society for Information Science and Technology* 55(14), 1216–1227.

Blood, R. (2004), How blogging software reshapes the online community, *Communications of the ACM* 47(12), 53–55.

Bollen, J., Van de Sompel, H. & Rodriguez, M. (2008), Towards usage-based impact metrics: first results from the MESUR project., *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries* pp. 231–240.

Bollen, J., Van de Sompel, H., Smith, J. & Luce, R. (2005), Toward alternative metrics of journal impact: A comparison of download and citation data, *Information Processing and Management* 41(6), 1419–1440.

Bookstein, A. (1990), Informetric distributions, part 1: Unified overview, *Journal of the American Society for Information Science* 41(5), 368–375.

Botafogo, R., Rivlin, E. & Shneiderman, B. (1992), Structural analysis of hypertexts: Identifying hierarchies and useful metrics, *ACM Transactions on Information Systems (TOIS)* 10(2), 142–180.

Bradford, S. (1934), Sources of Information on Specific Subjects, *Journal of Information Science* 10, 176–180.

Brin, S. & Page, L. (1998), The Anatomy of a Large-Scale Hypertextual Web Search Engine, *Computer networks and ISDN systems* 30, 107–117.

Brin, S., Page, L., Motwani, R. & Winograd, T. (1998), The PageRank Citation Ranking: Bringing Order to the Web, Technical report, (SIDL-WP-1999-0120), Stanford Digital Library Technologies Project.
Available at `http://ilpubs.stanford.edu:8090/422/`

Brody, T. (2006), Evaluating research impact through open access to scholarly communication, Master's thesis, University of Southampton.

Brody, T. & Harnad, S. (2004), Comparing the impact of Open Access (OA) vs. non-OA articles in the same journals, *D-lib Magazine* 10(6).
Available at `http://www.dlib.org/dlib/june04/harnad/06harnad.html`

Brody, T. & Harnad, S. (2006), Earlier web usage statistics as predictors of later citation impact, *Journal of the American Society for Information Science and Technology* 57(8), 1060–1072.

Cailliau, R. (1995), A little history of the World Wide Web, *w3c*.
Available at `http://www.w3.org/History.html`

Carr, L. & Harnad, S. (2005), Keystroke economy: a study of the time and effort involved in self-archiving, *Working Paper*.
Available at `http://eprints.ecs.soton.ac.uk/10688/`

Carr, L. & MacColl, J. (2005), IRRA (Institutional Repositories and Research Assessment) RAE software for institutional repositories, *White Paper, IRRA Project*.
Available at `http://irra.eprints.org/white`

Carrière, S. & Kazman, R. (1997), WebQuery: Searching and visualizing the Web through connectivity, *Computer Networks and ISDN Systems* 29(8-13), 1257–1267.

Carver, R. (1993), The case against statistical significance testing, revisited, *Journal of Experimental Education* 61(4), 287–292.

Castells, M. (2010), *End of Millennium: The Information Age: Economy, Society, and Culture*, Vol. 3, Wiley-Blackwell.

Chen, P., Xie, H., Maslov, S. & Redner, S. (2007), Finding Scientific Gems with Google's PageRank Algorithm, *Journal of Informetrics* 1(1), 8–15.

Chesney, T. (2006), An empirical examination of Wikipedias credibility, *First Monday* 11(11).
Available at `http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/viewArticle/1413/1331`

Claivaz, J., Le Meur, J. & Robinson, N. (2001), From fulltext documents to structured citations: CERNs automated solution, *HEP Libraries Webzine* 5(2).

Clapham, P. (2005), Publish or Perish, *BioScience* 55(5), 390–391.

Crow, R. (2006), The Case for Institutional Repositories: a SPARC Position Paper, *ARL Bimonthly Report* (223).
Available at `http://works.bepress.com/ir_research/7`

Daniel, L. (1998), Statistical significance testing: A historical overview of misuse and misinterpretation with implications for the editorial policies of educational journals, *Research in the Schools* 5(2), 23–32.

Davis, P. (2010), Does Open Access Lead to Increased Readership and Citation, *The Physiologist* 53, 197–201.

Day, M. (2004), Institutional repositories and research assessment, *Supporting Study* 4.

Dellavalle, R., Hester, E., Heilig, L., Drake, A., Kuntzman, J., Graber, M. & Schilling, L. (2003), Going, going, gone: Lost Internet references, *Science* 302(5646), 787–788.

Dellavalle, R., Schilling, L., Rodriguez, M., Van de Sompel, H. & Bollen, J. (2007), Refining Dermatology Journal Impact Factors using PageRank, *Journal of the American Academy of Dermatology* 57(1), 116–119.

Dingley, B. (2006), US Periodical Prices–2005, *US Periodical Price Index 2005, American Library Association* pp. 1–16.

Drott, M. (1995), Reexamining the role of conference papers in scholarly communication, *Journal of the American Society for Information Science* 46(4), 299–305.

Egghe, L. (2005), Relations between the continuous and the discrete Lotka power function, *Journal of the American Society for Information Science and Technology* 56(7), 664–668.

Eysenbach, G. (2006), Citation Advantage of Open Access Articles, *PLoS Biology* 4(5).

Eysenck, H. (1960), The concept of statistical significance and the controversy about one-tailed tests, *Psychological Review* 67(4), 269–271.

Falagas, M., Pitsouni, E., Malietzis, G. & Pappas, G. (2008), Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses, *The FASEB Journal* 22, 338–342.

Fassoulaki, A., Paraskeva, A., Papilas, K. & Karabinis, G. (2000), Self-citations in six anaesthesia journals and their significance in determining the impact factor, *British Journal of Anaesthesia* 84(2), 266–269.

Fielding, R., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P. & Berners-Lee, T. (1999), HyperText Transfer Protocol - HTTP/1.1, Technical report, RFC 2616.
Available at `http://www.ietf.org/rfc/rfc2616.txt`

Fry, J., Oppenheim, C., Creaser, C., Johnson, W., Summers, M., White, S., Butters, G., Craven, J. & Hartley, R. (2009), Communicating knowledge: how and why researchers publish and disseminate their findings, *Report commissioned by the Research Information Network (RIN) and Joint Information Systems Committee (JISC).*

Gandal, N. (2001), The dynamics of competition in the internet search engine market, *International Journal of Industrial Organization* 19(7), 1103–1117.

Garfield, E. (1955), Citation indexes for sciences: A new dimension in documentation through association of ideas, *Science* 122(3159), 108–111.

Garfield, E. (1972), Citation Analysis as a Tool in Journal Evaluation Journals, *Science* 178(4060), 471–479.

Garfield, E. (1973), Citation Frequency as a Measure of Research Activity and Performance, *Essays of an Information Scientist* 1, 406–408.

Garfield, E. (2003), The meaning of the Impact Factor, *International Journal of Clinical and Health Psychology* 3(2), 363–369.

Garfield, E. (2005), The Agony and the Ecstasy: the History and the Meaning of the Journal Impact Factor (2005), *in* Fifth International Congress on Peer Review in Biomedical Publication, Chicago, USA.

Gaule, P. & Maystre, N. (2008), Getting Cited: Does Open Access Help?, *CEMI working paper 2008007.*
Available at `http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1427763`

Ginsparg, P. (1994*a*), Electronic publishing in science, *Computers in Physics* 8(4), 390–396.

Ginsparg, P. (1994*b*), First steps toward electronic research communication, *Gateways to Knowledge: The role of academic libraries in teaching, learning and research.*

Goel, R. (2003), A market mechanism for scientific communication: a comment, *Kyklos* 56(3), 395–400.

Goel, R. & Faria, J. (2007), Proliferation of academic journals: effects on research quantity and quality, *Metroeconomica* 58(4), 536–549.

Goodrum, A., McCain, K., Lawrence, S. & Lee Giles, C. (2001), Scholarly publishing in the Internet age: a citation analysis of Computer Science literature, *Information Processing & Management* 37(5), 661–675.

Gray, M. (1996), Web growth summary, *Massachusetts Institute of Technology.* Available at `http://stuff.mit.edu/people/mkgray/net/web-growth-summary.html`

Hardy, I. (1996), The Evolution of ARPANET email, *History Thesis paper. Berkeley.* Available at `http://www.ifla.org/documents/internet/hari1.txt`

Harnad, S. (1995), The PostGutenberg galaxy: How to get there from here, *Information Society* 11(4), 285–292.

Harnad, S. (2006*a*), Maximizing research impact through institutional and national open-access self-archiving mandates, *Proceedings of CRIS2006. Current Research Information Systems: Open Access Institutional Repositories Bergen, Norway.* Available at `http://cogprints.org/4787/`

Harnad, S. (2006*b*), Publish or Perish. Self-Archive to Flourish: The Green Route to Open Access, *ERCIM News* 64.

Harnad, S., Brody, T., Vallieres, F., Carr, L., Hitchcock, S., Gingras, Y., Oppenheim, C., Hajjem, C. & Hilf, E. (2008), The access/impact problem and the green and gold roads to open access: An update, *Serials review* 34(1), 36–40. Available at `http://eprints.ecs.soton.ac.uk/15852/`

Hecht, F., Hecht, B. & Sandberg, A. (1998), The Journal Impact Factor: A Misnamed, Misleading, Misused Measure, *Cancer Genetics and Cytogenetics* 104(2), 77–81.

Hick, S., Halpin, E. & Hoskins, E. (2000), *Human rights and the Internet*, Palgrave Macmillan.

Hirsch, J. (2005), An index to quantify an individuals scientific research output, *Proceedings of the National Academy of Sciences of the United States of America* 102(46), 16569–16572.

Hitchcock, S., Brody, T., Gutteridge, C., Carr, L., Hall, W., Harnad, S., Bergmark, D. & Lagoze, C. (2002), Open citation linking: The way forward, *D-Lib Magazine* 8(10). Available at `http://www.dlib.org/dlib/october02/hitchcock/10hitchcock.html`

Hitchcock, S., Woukeu, A., Brody, T., Carr, L., Hall, W. & Harnad, S. (2002), Evaluating Citebase, an open access Web-based citation-ranked search and impact discovery service, Technical Report ECSTR-IAM03-005, University of Southampton. Available at `http://eprints.ecs.soton.ac.uk/8204/`

Huberty, C. (1987), On statistical testing, *Educational Researcher* 16(8), 4–9.

Hulme, E. (1923), *Statistical bibliography in relation to the growth of modern civilization: two lectures delivered in the University of Cambridge in May, 1922*, Printed for the author by Butler & Tanner; Grafton & Co.

Johnson, D. (1999), The insignificance of statistical significance testing, *The Journal of Wildlife Management* 63(3), 763–772.

Jones, R., Andrew, T. & MacColl, J. (2006), *The institutional repository*, Chandos Pub.

Kessler, M. (1963), Bibliographic coupling between scientific papers, *American documentation* 14(1), 10–25.

Kleinberg, J. (1999), Authoritative sources in a hyperlinked environment, *Journal of the ACM* 46(5), 604–632.

Kyvik, S. (2003), Changing trends in publishing behaviour among university faculty, 1980-2000, *Scientometrics* 58(1), 35–48.

Langridge, S. & Hickson, I. (2002), Pingback 1.0 Specification, Technical report. Available at `http://www.hixie.ch/specs/pingback/pingback`

Lawrence, S. (2001*a*), Free online availability substantially increases a paper's impact, *Nature* 411(6837), 521.

Lawrence, S. (2001*b*), Online or Invisible, *Nature* 411(6837), 521.

Lawrence, S. & Giles, C. (1999), Accessibility of information on the web, *Nature* 400(6740), 107–109.

Lawrence, S., Giles, C. & Bollacker, K. (1999), Digital libraries and autonomous citation indexing, *IEEE Computer* 32(6), 67–71.

Levitt, J. & Thelwall, M. (2008*a*), Is multidisciplinary research more highly cited? A macrolevel study, *Journal of the American Society for Information Science and Technology* 59(12), 1973–1984.

Levitt, J. & Thelwall, M. (2008*b*), Patterns of annual citation of highly cited articles and the prediction of their citation ranking: A comparison across subjects, *Scientometrics* 77(1), 41–60.

Levitt, J. & Thelwall, M. (2011), A combined bibliometric indicator to predict article impact, *Information Processing & Management* 47(2), 300–308.

Li, X., Thelwall, M., Musgrove, P. & Wilkinson, D. (2003), The relationship between the links/Web Impact Factors of Computer Science departments in UK and their RAE (Research Assessment Exercise) ranking in 2001, *Scientometrics* 57(2), 239–255.

López-Munoz, F., Alamo, C., Rubio, G., García-García, P., Martín-Agueda, B. & Cuenca, E. (2003), Bibliometric analysis of biomedical publications on SSRI during 1980-2000, *Depression and anxiety* 18(2), 95–103.

Lotka, A. (1926), The frequency distribution of scientific productivity, *Journal of Washington Academy of Sciences* 16, 317–323.

Lynch, C. (2003), Institutional repositories: essential infrastructure for scholarship in the digital age, *Portal-Libraries and the Academy* 3(2), 327–336.

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

Macpherson Barrett, P. (2009), Report on the pilot exercise to develop bibliometric indicators for the Research Excellence Framework, *HEFCE Issues Paper*.
Available at `www.hefce.ac.uk/pubs/hefce/2009/09_39/`

Maslov, S. & Redner, S. (2008), Promise and pitfalls of extending Google's PageRank algorithm to citation networks, *The Journal of Neuroscience* 28(44), 11103–11105.

Massie, D., Campbell, K. & Williams, A. (1995), Traffic Accident involvement rates by driver age and gender, *Accident Analysis & Prevention* 27(1), 73–87.

Matthews, B., Duncan, A., Jones, C., Neylon, C., Borkum, M., Coles, S. & Hunter, P. (2009), A protocol for exchanging scientific citations, *in* 2009 Fifth IEEE International Conference on e-Science, IEEE, pp. 171–177.

McVeigh, M. (2005), Open Access Journals in the ISI Citation Databases: Analysis of Impact Factors and Citation Patterns A citation study from Thomson Scientific, *Thomson Scientific*.
Available at `http://scientific.thomson.com/media/presentrep/essayspdf/openaccesscitations2.pdf`

Meho, L. & Sonnenwald, D. (2000), Citation ranking versus peer evaluation of senior faculty research performance: A case study of Kurdish scholarship, *Journal of the American Society for Information Science* 51(2), 123–138.

Meho, L. & Yang, K. (2007), Impact of data sources on citation counts and rankings of LIS faculty: Web of Science vs. Scopus and Google Scholar, *Journal of the American Society for Information Science and Technology* 58(13), 2105–2125.

Mihalcea, R. (2004), Graph-based ranking algorithms for sentence extraction, applied to text summarization, *in* 42nd annual meeting of the association for computational linguistics (ACL 2004), Association for Computational Linguistics.

Mitzenmacher, M. (2004), A Brief History of Generative Models for Power Law and Lognormal Distributions, *Internet Mathematics* 1(2), 226–251.

Moed, H. (2005), Citation analysis of scientific journals and journal impact measures, *Current Science* 89(12), 1990–1996.

Moed, H. (2007), The Effect of Open Access on Citation Impact: An Analysis of ArXiv's Condensed Matter Section, *Journal of the American Society for Information Science and Technology* 58(13), 2047–2054.

Moed, H. (2009), New developments in the use of citation analysis in research evaluation, *Archivum immunologiae et therapiae experimentalis* 57(1), 13–18.

Moed, H. & Van Leeuwen, T. (1995), Improving the accuracy of Institute for Scientific Information's Journal Impact Factors, *Journal of the American Society for Information Science* 46(6), 461–467.

Mohr, L. (1998), *Understanding significance testing*, Sage.

Moore, G. (1998), Cramming more components onto integrated circuits, *Proceedings of the IEEE* 86(1), 82–85.

Newman, M. (2003), The structure and function of complex networks, *SIAM review* 45, 167–256.

Noruzi, A. (2005), Google Scholar: the new generation of citation indexes, *Libri* 55(4), 170–180.

Odlyzko, A. (1995), Tragic loss or good riddance? The impending demise of traditional scholarly journals, *Notices of the American Mathematics Society* 42, 49–53.

Odlyzko, A. (2002), The rapid evolution of scholarly communication, *Learned Publishing* 15(1), 7–20.

O'Neill, E., Lavoie, B. & Bennett, R. (2003), Trends in the evolution of the public web, *D-lib Magazine* 9(4), 1082–9873.
Available at `http://www.dlib.org/dlib/april03/lavoie/04lavoie.html`

Oppenheim, C. (1995), The correlation between citation counts and the 1992 Research Assessment Exercise Ratings for British Library and information science university departments, *Journal of Documentation* 51, 18–18.

Oppenheim, C. (1997), The correlation between citation counts and the 1992 Research Assessment Exercise ratings for British research in genetics, anatomy and archaeology, *Journal of documentation* 53(5), 477–487.

Panitch, J. & Michalak, S. (2005), The Serials Crisis: A White Paper for the UNC-Chapel Hill Scholarly Communications Convocation, *Chapel Hill, NC: University of North Carolina*.
Available at `http://www.unc.edu/scholcomdig/whitepapers/panitch-michalak.html`

Pauly, D. & Stergiou, K. (2005), Equivalence of results from two citation analyses: Thomson ISI's Citation Index and Google's Scholar service, *Ethics in Science and Environmental Politics* 5, 33–35.

Perra, N. (2008), Spectral centrality measures in complex networks, *Physical Review E* 78(3), 036107.

Peter, I. (2010), The history of email, *NetHistory*.
Available at `http://www.nethistory.info/History%20of%20the%20Internet/email.html`

Pinfield, S. & Gorman, G. (2004), Self archiving publications, *International Yearbook of Library and Information Management 2004/2005: Scholarly Publishing in an Electronic Era* 2004/2005, 118–145.
Available at `http://eprints.nottingham.ac.uk/142/`

Pinski, G. & Narin, F. (1976), Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics, *Information Processing & Management* 12(5), 297–312.

Potter, W. (1988), Of Making Many Books There Is No End: Bibliometrics and Libraries, *Journal of Academic Librarianship* 14(4), 238a–38c.

Priem, J., Taraborelli, D., Groth, P. & Neylon, C. (2010), Alt-Metrics: A Manifesto.
Available at `http://altmetrics.org/manifesto/`

Pritchard, A. (1969), Statistical bibliography or bibliometrics, *Journal of Documentation* 25(4), 348–349.

Rahm, E. (2008), Comparing the scientific impact of conference and journal publications in Computer Science, *Information Services and Use* 28(2), 127–128.

Redner, S. (1998), How popular is your paper? An empirical study of the citation distribution, *The European Physical Journal B* 4(2), 131–134.

Salton, G. (1987), *A theory of indexing*, Society for Industrial Mathematics.

Schöpfel, J. & Boukacem-Zeghmouri, C. (2010), Assessing the Return on Investments in GL for Institutional Repositories, *in* D. Farace & J. Schöpfel, eds, Grey Literature in Library and Information Studies, pp. 227–238.

Seglen, P. (1997), Why the impact factor of journals should not be used for evaluating research, *British Medical Journal* 314(7079), 497–502.

Sidiropoulos, A. & Manolopoulos, Y. (2006), Generalized comparison of graph-based ranking algorithms for publications and authors, *Journal of Systems and Software* 79(12), 1679–1700.

Simkin, M. & Roychowdhury, V. (2005), Stochastic modeling of citation slips, *Scientometrics* 62(3), 367–384.

Small, H. (1973), Co-citation in the Scientific Literature: A New Measure of the Relationship Between Two Documents, *Journal of the American society for information science* 24(4), 265–269.

Smith, L. (2002), A tutorial on Principal Components Analysis.
Available at `http://staf.cs.ui.ac.id/WebKuliah/citra/2004/principal_components.pdf`

Swan, A. & Brown, S. (2004*a*), Authors and Open Access publishing, *Learned Publishing* 17, 219–224.
Available at `http://eprints.ecs.soton.ac.uk/11003/`

Swan, A. & Brown, S. (2004*b*), JISC/OSI journal authors survey report, Technical report, JISC/OSI.
Available at `http://ie-repository.jisc.ac.uk/274/`

Swan, A. & Brown, S. (2005), Open access self-archiving: An author study, Technical report, University of Southampton.
Available at `http://eprints.ecs.soton.ac.uk/10999/`

Syfret, M. (1948), The Origins of the Royal Society, *Notes and Records of the Royal Society of London* 5(2), 75–137.

Testa, J. & McVeigh, M. (2004), The impact of Open Access Journals. A citation study from Thomson ISI, *Thomson Scientific, Philadelphia.*
Available at `http://www.thomsonscientific.jp/event/oal/impact-oa-journals.pdf`

Thelwall, M. (2003), Can Google's PageRank be used to find the most important academic Web pages?, *Journal of Documentation* 59(2), 205–217.

Thelwall, M. (2008), Bibliometrics to webometrics, *Journal of information science* 34(4), 605–621.

Thelwall, M., Vaughan, L. & Björneborn, L. (2005), Webometrics, *Annual review of information science and technology* 39(1), 81–135.

Wang, J., Liu, J. & Wang, C. (2007), Keyword extraction based on PageRank, Springer, pp. 857–864.

Watson, A. (2009), Comparing citations and downloads for individual articles at the Journal of Vision, *Journal of Vision* 9(4), 1–4.

West, J., Althouse, B., Rosvall, M., Bergstrom, C. & Bergstrom, T. (2008), Eigenfactor Score and Article Influence Score: Detailed methods.
Available at `http://www.eigenfactor.org/methods.pdf`

White, S. & Creaser, C. (2004), Scholarly journal prices: selected trends and comparisons, Technical report, Occasional Paper no. 34, Library and Information Statistics Unit, Loughborough University.

Wilson, C. (1999), Infometrics, *Annual Review of Information Science and Technology* 34, 107–247.

Wuchty, S., Jones, B. & Uzzi, B. (2007), The increasing dominance of teams in production of knowledge, *Science* 316(5827), 1036.

Yan, E. & Ding, Y. (2011), Discovering author impact: A PageRank perspective, *Information Processing & Management* 47(1), 125–134.

Zipf, G. (1932), Selective Studies and the Principle of Relative Frequency in Language, *Harvard University Press, Cambridge, MA*.

Zipf, G. (1949), Human Behavior and the Principle of Least Effort, *Addison-Wesley, Cambridge, MA*.