IDENTIFYING REMARKABLE RESEARCHERS USING CITATION NETWORK ANALYSIS

EPHRANCE ABU UJUM

FACULTY OF SCIENCE
UNIVERSITY OF MALAYA
KUALA LUMPUR

2014

IDENTIFYING REMARKABLE RESEARCHERS USING CITATION NETWORK ANALYSIS

EPHRANCE ABU UJUM

DISSERTATION SUBMITTED IN FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE

INSTITUTE OF MATHEMATICAL SCIENCES
FACULTY OF SCIENCE
UNIVERSITY OF MALAYA
KUALA LUMPUR

2014

UNIVERSITI MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: **EPHRANCE ABU UJUM**

I.C./Passport No.: **780923-12-5195**

Registration/Matric No.: **SGP070002**

Name of Degree: **MASTER OF SCIENCE**

Title of Project Paper/Research Report/Dissertation/Thesis ("this Work"):

"IDENTIFYING REMARKABLE RESEARCHERS USING CITATION NETWORK ANALYSIS"

Field of Study: **MATHEMATICAL MODELING**

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

(Candidate Signature) Date:

Subscribed and solemnly declared before,

Witness's Signature Date:

Name: **PROFESSOR DR KURUNATHAN A/L RATNAVELU**
Designation: **PROFESSOR**

ABSTRACT

Experts or authorities within a research field exhibit specific traits in how they publish as well as in how they are cited by others. An analysis of such citation dependencies requires a network approach whereby a researcher's impact depends not only on the number of citations he/she has accumulated (over a given period of time) but also on the prominence of researchers who depend on their work. This thesis shall explore how to distinguish researchers based on temporal patterns of their publication and citation records.

As intuition may suggest, the influence of a researcher is proportional to the number of citations he/she has acquired as well as the influence of his/her citing authors. Authority can also be conferred to a researcher by virtue of his/her (co)authored works that continue to accrue citations long after the year of publication.

In this thesis, experts or authorities are identified using the "temporal citation network analysis" approach of Yang, Yin, and Davison (2011). This method assigns a high influence score to researchers who are still actively and persistently publishing, have long publication track record, and are heavily cited (especially by influential peers).

As a case study, the method proposed by Yang and co-workers shall be used to identify authorities within the ISI Web of Knowledge category of "BUSINESS, FINANCE" spanning the period 1980-2011 inclusive. The thesis shall also explore a modification of this method to predict rising stars within the same dataset.

ABSTRAK

Pakar dalam sesebuah bidang penyelidikan menunjukkan ciri-ciri khusus dalam cara mereka menerbitkan artikel dan juga dalam cara mereka dirujuk penyelidik lain. Analisa kebergantungan pemetikan perlu didekati dengan menggunakan konsep rangkaian di mana impak seseorang penyelidik tidak hanya bergantung kepada jumlah pemetikan yang diperolehi (dalam suatu jangka masa tertentu), tetapi juga pada kewibawaan penyelidik-penyelidik lain yang bergantung kepada karya dan ciptaannya.

Disertasi ini meneliti cara membezakan penyelidik dengan mengeksploitasikan pola batas waktu dalam rekod penerbitan dan pemetikan mereka. Seperti yang dicadangkan intuisi, pengaruh seseorang penyelidik berkadar terus dengan jumlah pemetikan yang diperolehi serta pengaruh penyelidik yang memetik artikelnya. Kewibawaan turut diberikan kepada seseorang penyelidik menerusi karya kongsi yang menerima pemetikan beterusan walau bertahun lama sejak tahun penerbitan.

Disertasi ini akan mengenalpasti pakar dengan menggunakan kaedah "temporal citation network analysis" yang disarankan oleh Yang et al. (2011). Kaedah ini memberi skor pengaruh yang tinggi kepada penyelidik yang masih aktif dan menerbitkan artikel secara beterusan, mempunyai rekod penerbitan yang ekstensif, dan juga dipetik secara intensif (terutama sekali daripada kumpulan yang berpengaruh).

Sebagai kes kajian, kaedah yang disarankan oleh Yang et al. akan digunakan untuk mengenalpasti pakar-pakar dalam kategori subjek "BUSINESS, FINANCE" daripada pangkalan data ISI Web of Knowledge dalam jangka waktu merentangi tahun 1980 sehingga (dan termasuk) tahun 2011. Disertasi in juga meneliti modifikasi kaedah Yang et al. untuk meramal pakar yang akan datang dengan menggunakan set data yang sama.

ACKNOWLEDGEMENTS

First and foremost, I offer thanks to God for giving me the opportunity to explore ideas. I express great gratitude to my supervisor Professor Dr. Kurunathan Ratnavelu for his unwavering support and belief in me over the years. His guidance and wisdom has made, and will continue to make an enduring impact on my life. I also wish to express heartfelt gratitude to my collaborator and friend, Dr. Choong Kwai Fatt, for relentlessly encouraging the pursuit of this work and other related problems.

I hereby thank the hardworking staff of the Institute of Mathematical Sciences, Faculty of Science, and the Institute of Postgraduate Studies for their invaluable help and counselling especially during various setbacks I faced during the completion of this thesis. I am profoundly indebted to their patience and generosity. Some critical directions in this work was developed under grants RG146/10HNE and RG298/11HNE provisioned under the UMRG scheme, and later through the UM High Impact Research (HIR) grant.

I wish to thank Thomson Reuters for the data used in this thesis, obtained specifically via institutional access to the *Web of Knowledge*. I also wish to acknowledge the ingenuity of the Open Source community, specifically in Linux, R, Perl, Python, Gephi, and LaTeX.

To my friends and colleagues: I am grateful to Chin Jia Hou, Melody Tan, and Tan Hui Xuan for the help they have given me from time to time. A great deal of the ideas and questions pursued in this work stemmed from useful discussions I have had with these wonderful people. All in all, I owe my family for their endless love and understanding, without which this work would have been impossible to accomplish. It is to them that I dedicate this work.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Background

This thesis focuses on the ranking of researchers in terms of published and cited expertise. Though not apparent at first glance, the need to rank is rooted in the need to rationally allocate resources under constraint or uncertainty[1]. When decisions must be made wherein one choice affects (advances or suppresses) future actions, the right considerations and determinations must be taken into account to properly weigh feasible options. Sometimes there is either too much or too little information to go on. For a researcher looking for clues on how to advance his/her research, there is a vast *search space*[2] to explore (McNee et al., 2002). There simply is not enough time available for any one person to effectively sample every data point in the search space, or every connection, for that matter. Furthermore, each choice may bias one's ability to recognise or decide on future choices[3].

The same goes for decision makers in research management: researchers and the work they produce are routinely weighed and sorted by importance to reflect the scarcity

[1]Researchers want to find relevant literature with minimal time and effort. For a given collection, one can reasonably guess what these are based on the importance signalled by other researchers. On the other hand, decision makers in research management are interested in identifying important workers to support based on available funding and resources.

[2]In terms of the number of published works to keep track of, the works cited by those works, and so on, up to the earliest available works. It is also common to track work published by a particular researcher (or group of researchers), which, at the time of writing, numbers in the millions (alive or dead). In spite of this, not all researchers and their work can, or need to be considered as they may not be relevant to the task at hand. Thus, ranking items by relevance and/or importance is one key strategy to filter out vast amounts of unnecessary/irrelevant information.

[3]This can be attributed to the Matthew effect which states that "the rich get richer and the poor get poorer" (Merton, 1968; Gladwell, 2008). Given that moments in life are strung together by a series of choices, one's disposition changes (is reinforced or weakened) through the course of action taken. Hence there exist *opportunity costs* i.e. the forfeiture of potential gains from unchosen alternatives, among which potentially includes the ability to progressively judge and make better choices (or recover from bad ones).

of available resources (Moed, 2008). What's more, it is often unclear what the expected payoff is specific to a given choice, or whether the expected payoff can even be met. Hence, it is essential to prioritise available options based on tangible evidence, or lacking that, on reasonably accurate or descriptive indicators. In this, data mining is useful to assign value to available options based on a given set of assumptions and data. This information can then be used to help organise (sort) the search space[4] and to inform the decision making process.

Before proceeding, perhaps some perspective is in order. Suppose it takes an average researcher a minimum of one hour to effectively search and read a paper. If one dedicated 3 hours a day to keep apprised of new literature, this totals to $3 \times 365 = 1095$ new papers covered in a year. In contrast, there are, for example, 18,300 Google Scholar-indexed articles in 2013 containing the phrase "global financial crisis" (at the time of writing), hence an average researcher may cover roughly 6% of that literature. Of course not all of this research is actually relevant to any one researcher, and no two papers are thoroughly read in an equal amount of time, but the point here is that because of the sheer volume of available information (new and old), compromises are difficult to avoid. One has to take in a manageable number of items fulfilling some evaluation criteria and effectively discard the bulk of those that don't.

Furthermore, this decision (filtering) process also takes a non-trivial amount of time and so one has to rely on available "indicators" to shortcut the task. For research papers, this is routinely done by checking the number of citations received or by discriminating papers by the authority of its authors (or even their institutional affiliation). The tricky part is when some discarded items or authors offer useful or relevant information but are inadvertently missed out because the indicator(s) used are not comprehensive enough to

[4]Specific to the ranking of authors to research papers, the search space (of authors and their published work) can be organised in terms of authority and quality (or trust and reputation).

include such instances.

Similar constraints are also faced when conducting a performance assessment of research staff. If decision makers are not themselves expert in the fields they manage, selecting candidates based on indicators like the number of publications, number of citations, impact factor of journals, and h-index quite often does a good enough job, bearing in mind of course that these indicators are only as good as the assumptions they are based on. For one thing, the number of publications suggests productivity and not necessarily the quality of the publications or authors themselves.

Also, the number of citations to a paper measures its "citedness", the number of times in which it has been referenced by other papers. Some citations may actually consist of self-citations, that is, citations received by an author by him/herself in his/her successive works. While this is a crucial component in advancing one's research, it is misleading to infer impact when one predominantly receives citations from him/herself instead of from others. This raises further questions: supposing that a citation received by a paper signals impact or importance, then which ones really matter, which ones matter less, and which ones are done purely out of convenience? When asked this way, a citation count seems far too simple to properly capture the complex nuances associated with impact.

Since a person's career in research is not merely the sum of his/her publications or citations, I wanted to study how available data can be used to "mine" the reputation of authors based on how they publish[5] as well as how they influence others. To achieve this goal, I constructed document and author citation networks using articles indexed under Thomson ISI's subject category of "BUSINESS, FINANCE" as a case study. I then used a method proposed by Yang et al. (2011) in a paper entitled "Award prediction with temporal citation network analysis", which specifically assigns a high influence score to researchers who are still actively and persistently publishing, have long publication track

[5]How long, how often, and when.

record, and are heavily cited (especially by influential peers).

This method can be used to identify active experts, predict prospective award (grant) recipients, and discover articles that can be considered as scientific gems[6] (Chen, Xie, Maslov, & Redner, 2007). If such a method were used for the purpose of research management, young and promising researchers may be put at a disadvantage (due to shorter track record from which to infer future success). To circumvent this issue, I modified the method of Yang and co-workers to identify potential rising stars as well, specifically by adding bias to researchers who are cited by authorities many years their senior (Daud, Abbasi, & Muhammad, 2013).

The objective of this work is twofold. First, I wish to study how network analysis methods can be used to gauge the relative impact of researchers based on publication and citation records. Second, I seek to explore how citation network analysis can be utilised to find novel features that are otherwise easily missed (experts, rising stars, and scientific gems). This procedure is called *feature extraction* (Cukierski, Hamner, & Yang, 2011). Ultimately, the knowledge gained from this study should lend some insight on how to write customised code for automated discovery of important documents and authors from large sets of bibliometric data.

This thesis is organised as follows: Chapter 2 describes how the source data was collected and parsed to construct article citation networks and author citation networks. This chapter will also cover the methods used to score researchers and documents based on their location within a structure of citation links, as well as propose a set of screening criteria for determining persons of interest. Chapter 3 provides an analysis of the networks constructed and a listing of researchers that fulfil the set of screening criteria proposed in Chapter 2. The limitations of the methods used shall also be covered in Chapter 3, along with a discussion on alternative applications as well as possible future directions. The

[6]Possesses a modest citation count but plays an important role in the progression of a research field.

thesis is concluded in Chapter 4.

1.2 Literature Review

This section presents a literature review beginning with key concepts used for scoring researchers using conventional bibliometric/scientometric approaches. This is then followed by a review of network analytic approaches, specifically those used in citation networks.

1.2.1 Quantifying authority and expertise

One of the overlapping goals of bibliometric and scientometric research is to measure research output and impact based on publication or citation index data (Pritchard, 1969; Tague-Sutcliffe, 1992; Van Raan, 1997), often referred to as bibliometric data. In principle, the ability to measure provides some basis to compare or discriminate certain quantifiable attributes between entities in research (individual persons, institutions, countries, documents, publications, etc). Though useful to its practitioners and advocates, bibliometric and scientometric methods are not without its detractors. Both fields have drawn criticism for the abuse of bibliometric data (Cameron, 2005), and in other instances for the questionable application or misinterpretation of statistical analyses (Bornmann, Mutz, Neuhaus, & Daniel, 2008; Adler, Ewing, & Taylor, 2009; Silverman, 2009).

Despite such resistance, bibliometric assessments have become a part of modern research culture (Lawrence, 2003), with terms like "publish or perish" (Silen, 1971; Harzing, 2010), "university rankings" (Liu & Cheng, 2005; Usher & Savino, 2007), "impact factor" (Garfield, 2006), and "*h*-index" (Hirsch, 2005) becoming increasingly emphasised in one form or another within national or institutional research policy. Whether for the utilitarian purpose of enhancing public image or to achieve improvements in research funding allocations, bibliometrics and scientometrics provide (to some extent) the means to obtain 'insight' into the inter- and intra-organisational state of affairs pertaining to re-

search (Van Raan, 1997; Hood & Wilson, 2001). To what degree that insight reflects the realities of research is of course, still subject to debate.

With respect to the evaluation of individual persons, or more specifically, researchers, there exist a number of bibliometric/scientometric approaches which I shall describe in the following subsections. For the most part, my interest lies in determining useful and practical ways to discriminate authority or expertise. Before proceeding, some clarification is necessary with regard to what indicates authority or expertise in bibliometric data. In particular, an expert may be prolific (i.e. highly productive), signaling a prodigious propensity to contribute to the existing body of knowledge (Shockley, 1957; Merton, 1988), as well as a perseverance to overcome the hurdles of peer review (Wright, 2001; Harrison, 2004; Bornmann, 2008; Fulda, 2008). However, this is by no means a necessary condition.

It can be argued that a strong indicator of expertise or authority is the ability to significantly exert influence upon others[7] (Kleinberg, 1999). On the one hand, some consistency is expected so that sporadic yet influential collaborations of an average researcher with many coauthors does not overly suggest expertise, especially if single-author works by the former generates dramatically less influence on average (Hirsch, 2005). On the other hand, one-off works that influence other influential works should carry more weight (in terms of indicating expertise) compared to those that influence less influential works (Chen et al., 2007). Based on these considerations, some judgements can be made on which indicators best characterise expertise.

[7] A telling sign of this can be seen in how scientists receive differential recognition for their work based on how they are located in a stratified system. This is termed the Matthew Effect (Merton, 1968). According to Cole (1970), "[...] lesser quality papers by high-ranking scientists receive greater attention than papers of equal quality by low-ranking scientists".

1.2.1 (a) Publication and citation count

On its own, the total number of papers (N_p) is a reasonable indicator for a researcher's productivity. However, one cannot simply infer quality from the quantity of papers produced. To this end, the total number of citations ($\sum_{j=1}^{N_p} C_j$) can be used to indicate impact, though not without considering factors that may actually inflate or exaggerate this value. For example, it is rather presumptuous to assume the influence of a researcher from just one highly cited paper obtained through a one-off collaboration (whether with highly prominent coauthors or otherwise). It is also conceivable to inflate the total citation count through a preponderance of review articles; these are known to acquire more citations (on average) compared to articles based on original work.

A seemingly reasonable alternative to sole reliance on either publication or citation count is to calculate the mean average impact of a researcher as *citations per paper* ($\sum_{j=1}^{N_p} C_j/N_p$). Such a metric however can be inflated by a high total citation count (from a highly skewed citation sequence) or through a small publication count (which corresponds to low productivity). Since it is unintuitive to penalise high productivity, this approach is far from ideal[8].

1.2.1 (b) Author ordering effects

A researcher's reputation within the research community is hard to measure, though under some circumstances, author ordering (authorship position) may provide some hints. To follow this line of reasoning, it is important to clarify under what circumstances author ordering entails significant information on the reputation of its constituent workers. To echo a question posed by Fehr and Schneider (2007): "Do authors (and policy makers) care about author ordering?" One can expect that the answer is in the affirmative in cases where intellectual credit is usually assigned to the first author, whereby he or she is

[8]To overcome this, one could perhaps use $score := \log(N_p) \sum_{j=1}^{N_p} C_j/N_p$. The purpose of the logarithmic term is to provide some bias towards researchers with higher publication count.

assumed to have rendered the most significant contribution towards the development of the work and its publication (Gaeta, 1999; Tscharntke, Hochberg, Rand, Resh, & Krauss, 2007). Furthermore, first author status is commonly associated with higher prestige in the context of academic promotion or reward mechanisms. In some circles, the last author position confers seniority status.

Under what circumstances does author ordering indicate status? This is clearcut in the case of three or more authors, that is, whenever author ordering breaks from alphabetical (or reverse alphabetical) listing. However, it is entirely possible for ordering by status to coincide with some alphabetical ordering, though the occurrence of such cases should dramatically decrease with the size of the collaboration. The case of two authors is inherently tricky since the listing may be in ascending or descending order, except for cases where a common convention is widely-adopted and the probability that any two authors going against that convention is sufficiently low to be neglected.

There are circumstances where alphabetical listing is prevalent over ordering by status. This is typically the case for economics journals in which lexicographic ordering is the norm and not the exception. Engers et al. (1999) posit that such norms emerge due to signalling "equilibrium between authors and the market". Specific to journals in the category of "BUSINESS, FINANCE", it is found that this dataset[9] exhibits a strong preference for lexical author ordering (see table 1.1). Hence, it is difficult, if not impossible, to ascertain the authority or expertise of researchers publishing in this category based on patterns in their authorship position.

Table 1.1: Number of articles in 30 journals under the "BUSINESS, FINANCE" dataset, that maintain forward/reverse alphabetical ordering at least 50% of the time. Each journal has at least 50 articles co-authored by 2 or more workers over the period 2005–2010.

Journal	Forward	Reverse	Lexical	%Lexical	Non-Lexical

[9]Consisting primarily of journals dedicated to the field of financial economics.

ACCOUNT FINANC	60	2	62	76.54	19
ACCOUNT ORG SOC	32	4	36	58.06	26
ACCOUNT REV	125	2	127	88.19	17
AUDITING-J PRACT TH	47	4	51	82.26	11
CONTEMP ACCOUNT RES	74	7	81	90.00	9
EUR FINANC MANAG	54	3	57	87.69	8
FINANC ANAL J	45	2	47	75.81	15
FINANC MANAGE	77	1	78	88.64	10
J ACCOUNT ECON	73	-	73	94.81	4
J ACCOUNT RES	49	-	49	94.23	3
J BANK FINANC	326	9	335	76.66	102
J BUS FINAN ACCOUNT	96	6	102	72.86	38
J CORP FINANC	91	2	93	91.18	9
J EMPIR FINANC	54	5	59	88.06	8
J FINANC	183	1	184	96.34	7
J FINANC ECON	222	-	222	96.52	8
J FINANC QUANT ANAL	93	1	94	94.00	6
J FUTURES MARKETS	72	3	75	71.43	30
J INT MONEY FINANC	90	4	94	83.19	19
J MONETARY ECON	111	-	111	96.52	4
J MONEY CREDIT BANK	98	2	100	89.29	12
J PORTFOLIO MANAGE	64	2	66	56.41	51
J REAL ESTATE FINANC	72	7	79	67.52	38
J RISK INSUR	45	6	51	65.38	27
J RISK UNCERTAINTY	32	3	35	62.50	21
NATL TAX J	42	2	44	83.02	9
QUANT FINANC	74	5	79	69.30	35
REAL ESTATE ECON	52	-	52	78.79	14
REV FINANC STUD	199	-	199	96.60	7
WORLD ECON	63	2	65	65.66	34

1.2.1 (c) Impact factor

By convention, evaluations of researchers depend not only on the number of papers or their authorship position, but also on the impact of the journals they publish in (Lawrence, 2003). The operating assumption behind this reasoning is that it takes considerable skill and resourcefulness to publish in a prestigious journal. Conversely, the prestige of a journal can be quantified in terms of how it attracts the most important work (Garfield, 1996), the bulk of which is presumably produced by the most important

researchers.

In 1971, the Institute for Scientific Information (now known as Thomson ISI), attempted the first systematic analysis of the 'network of journal information transfer' as well as the first published calculation of a journal's relative impact as an 'average citation rate per published article' (Garfield, 1972). This measure, called the journal impact factor score – or impact factor(IF), for short – can be calculated for each journal i in year t as:

$$IF_t^i = \frac{n_t^i}{A_{t-1}^i + A_{t-2}^i} \tag{1.1}$$

where n_t^i is the number of times in census year t that volumes published in the 2-year target window $t-1$ and $t-2$ of journal i are cited, while A_t^i is the number of *citable items*[10] published in journal i in year t (Garfield, 2006; Althouse, West, Bergstrom, & Bergstrom, 2009). IF scores are provided under Thomson ISI's Journal Citation Reports (JCR) database. This measure forms part of the basis of ISI's internal decision making on which journals to include and exclude within their database (Garfield, 1999).

Over time, the impact factor has been adopted for other uses beyond its original purpose: libraries use it as a bibliometric indicator to determine the purchase of journals within a given budget; publishers use it to monitor and make quantitative comparisons across journals as well as journal editors; and administrators use it to determine rank, promotion, and salary within a faculty (Rogers, 2002). The latter is most relevant to the subject matter of this thesis. Given the publication history of some target researcher X,

[10]ISI designates research articles, technical notes and reviews as "citable" items. "Non-citable items" include editorials, letters, news items, and meeting abstracts, and thus these document types do not contribute to the denominator of Equation (1.1). It is important to note that the choice of countable items in the numerator can be unclear (Dong, Loh, & Mondry, 2005).

an *individual impact factor profile* can be computed as:

$$score(X) := \sum_{t \in T} \sum_{q \in Q(t)} IF_t^i(q) \tag{1.2}$$

Here, $Q(t)$ denotes the set of papers published by researcher X at year t, while time T is either the set of years in which X has actively published, or alternately, a predefined census period. Note that this expression[11] implicitly assumes that article positioning by authors is a reliable predictor for their expertise.

Although it is tempting it is to infer article quality and, by extension, the reputation of its author(s) based on the publishing journal's prestige (Judge, Cable, Colbert, & Rynes, 2007), it is important to consider just how grounded this practice is (Seglen, 1997; Walter, Bloch, Hunt, & Fisher, 2003; Dong et al., 2005; Williams, 2007). While a top journal accrues impact (or influence) based on the articles it hosts, each constituent article is not necessarily a top article. Smith (2004) studied the effects of deducing the status of an article as a "top article" based on it being published in a "top N journal", where N is an arbitrary integer[12]. Using a sample of articles published in 1996 and a citation window spanning 1996 to 2004 for 15 leading[13] (ISI-indexed) finance journals, Type I and Type II error rates were determined. Specific to a top three journal rule, it was found that a Type I error rate – whereby a top article is rejected by the decision rule – results 44% of the time, while a Type II error rate – whereby a non-top article is identified as a top article – occurs 33% of the time.

The results of the study conducted by Smith (2004) (with respect to its specific pa-

[11] This scoring algorithm takes into account the frequency, as well as the range of journal impact factors the evaluatees have published in (concurrent to the the year of publication). It does not, however, take into account citation counts received for each article published by the evaluatee, and how far above or below they are from the average (and highest) citation count specific to the journals they have published in.

[12] Smith (2004) defines a top article as one in which "The average number of cites is above the median, mean, 90^{th} percentile published, or 95^{th} percentile for a set of leading finance journals".

[13] Selected by highest average number of cites per article.

rameters) suggests that nearly half the time, it is possible to miss a top article in the majority of "non-top 3 journals", while a third of the time, non-top articles may be wrongly designated as a top article simply by being published in a "top 3 journal". Although the prestige of a journal can – to some extent – be inferred from the aggregated importance of the works it hosts (Garfield, 2006), it is misguided to assume that all of its constituent articles are of the same pedigree (Seglen, 1997).

This even more so considering that a high citation count to individual research articles does not necessarily signal its importance or utility, but rather the level of interest the research community has in what these articles have to say (Bornmann & Daniel, 2008). A high level of interest may in fact be a mixture of positive (supportive) and negative (opposing) reactions, hence the underlying sentiment of a citation count cannot be readily ascertained without going into the details of how and why the citations were made in the first place. In light of this, the practice of inferring the reputation of researchers based on where they publish should be given some pause, especially if done without appropriate context (Dong et al., 2005; Scully & Lodge, 2005).

1.2.1 (d) Hirsch Index

The Hirsch index, or h-index, was devised by physicist Jorge E. Hirsch to gauge the overall impact of an individual researcher's publication record down to a single number (Redner, 2010). This is done by assuming that the publication and citation record of an individual contains useful information to "characterise the scientific output of a researcher" (Hirsch, 2005). Given such data, Hirsch proposes the following scoring method:

> A scientist has index h if h of his or her N_p papers have at least h citations each and the other $(N_p - h)$ papers have $\leq h$ citations each.

In this way, a researcher who consistently publishes highly cited papers will score a higher h-index compared to another who publishes equally many papers, yet accumulates a lower overall citation count.

For example, suppose two researchers publish 10 papers each. The first has an h-index of 10 indicating that 10 of his papers have at least 10 or more citations. The second has an h-index of 1 signifying that 1 of his papers has at least 1 or more citations, and the other 9 with zero citations or at most 1 citation each. The h-index for the second researcher is still 1 even if the one paper with $\geq$ 1 citations was actually cited 1000 times. As another example, consider one researcher with 10 papers each accumulating 10 citations, and another with 10 papers accumulating 100 citations each. Despite the seemingly obvious difference, both researchers have an h-index of 10. This raises some important concerns (Lehmann, Jackson, & Lautrup, 2006, 2008; Sidiropoulos, Katsaros, & Manolopoulos, 2007; García-Pérez, 2009; Prathap, 2010). In particular, if one assumes that publication and citation data contain (enough) useful data, the question then becomes, *is enough data accounted for in the h-index (or any) scoring process?*

1.2.2 Identifying authorities and experts on networks

An expert is a person who displays considerable knowledge or skill in a particular area (Chi, 2006). An authority on the other hand, is a broader term referring to prominent sources of information or instruction that includes people (Fiske, 1991; Marlow, 2004; Hirshfield, 2011), institutions (Choe, Lee, Seo, & Kim, 2013), documents (Kleinberg, 1999; Ding, He, Husbands, Zha, & Simon, 2002), and journals (Pinski & Narin, 1976; Medina & Leeuwen, 2012). When viewed as an information spreading process, an authority can be regarded as someone (or something) that exerts significant influence on other persons (or objects/entities). Such linkages can be neatly described as a network structure whereby each node is used to represent a distinct person, object, or entity, and directed

links between nodes signify the presence of connection as well as the directionality of dependence.

Additionally, link weights can be added to each directed link to denote the strength of the dependence. In this way, authorities are quite often easy to spot on a network as these correspond to nodes that occupy prominent positions within the link structure (Shafer, Isganitis, & Yona, 2006). The extent at which a node occupies a prominent position is hereon referred to as its *prominence*, which is mathematically expressed in terms of *node centrality*. There are several notions of prominence which shall be explored below.

1.2.2 (a) Erdös number, degree, closeness, and betweenness

The assignment of Erdös numbers on the co-authorship network of mathematicians provides an illustrative example of node centrality. Co-authorship networks signify the professional network of researchers used for collaboration and referrals. It is essentially a social network, whereby its organisation is shaped to some extent by trust and reputation of workers (Burt, 2005, 2010), as well as their mutual, complementing, or competing interests (Fafchamps, Leij, & Goyal, 2006, 2010; Goyal, 2009; Breslin et al., 2007).

Erdös number.—The Erdös number is computed as the geodesic (shortest path) distance of a mathematician from legendary polymath Paul Erdös[14], who himself is designated with the Erdös number zero (Grossman, 1996). Accordingly, direct collaborators of Erdös are assigned Erdös number 1, the collaborators of his collaborators Erdös number 2, and so on[15]. This numbering scheme generates much appeal as it intuitively codifies the "closeness" of a researcher to having collaborated with an intellectual giant.

[14]One of the most prolific and influential mathematicians to have ever lived, Erdös amassed over 500 collaborators from the start of his career in 1934, up to his death in 1996. According to personal accounts from his collaborators, Erdös would typically seek the hospitality of a mathematician he knew directly, or whom he was referred to, work feverishly with this host to tackle mathematical problems for several days straight, and upon parting from his host, ask for a recommendation on which mathematician to visit next (Hoffman, 1998).

[15]Up to the largest finite Erdös number, which is 13 (see http://www.oakland.edu/enp/trivia/). Mathematicians who cannot trace a connected path to Erdös are assigned an infinite Erdös number.

This notion of ego-centric centrality yields non-trivial information precisely because the structure of complex networks like the co-authorship network typically exhibit variation in the number of links from one node to the next (displays inhomogeneous connectivity patterns). If the co-authorship network of Paul Erdös were structured as a complete graph (whereby each node is indistinguishably connected to all other nodes), Erdös numbers would remain unchanged (i.e. reveals no new information) if computed from a different root node other than Paul Erdös himself.

However, this sensitivity to the choice of root node makes the computation of Erdös numbers of limiting interest for generic social networks since a better approach would be to have a centrality measure that is globally invariant, that is, a measure that is unchanged on the overall scale no matter where the calculation is started. Thankfully, other approaches are possible by exploiting specific quirks in the link structure of empirical networks (social or otherwise). These quirks are perhaps best described based on discoveries made on large-scale co-authorship networks (Newman, 2001c, 2001b, 2001d):

- *Higher level of clustering* than predicted by random (exponential) network models (Erdös & Rényi, 1959, 1960) due to local clustering (Watts & Strogatz, 1998; Newman, 2001a) and the presence of community structure (Girvan & Newman, 2002; Fortunato, 2010). The global clustering coefficient is given by the number of closed triplets of nodes over the total number of triplets (both open and closed). The probability for closed triplets to appear on a random network is small;

- *Heavy-tailed degree distribution* (highly skewed degree inhomogeneity). For an undirected network like the co-authorship network, the degree centrality, or simply the degree, of a node v, $C_D(v) = \deg(v)$, refers to the number of links attached to it. For directed networks like a citation network, a node can be measured by its *in-degree* (links pointing into a node) as well as *out-degree* (links pointing out of

a node). The degree distribution for co-authorship networks typically exhibit the following properties:

  - Large-scale cases (number of nodes $n \to \infty$) deviate from the Poisson degree distribution predicted by the classic Erdös-Rényi model. The probability of finding a node on a Erdös-Rényi graph having k links is $p(k) \sim \lambda^k e^{-k}/k!$, with average number of links given by $\lambda = np$, in which n is the number of nodes and p is the probability of attaching a link between any two nodes (Erdös & Rényi, 1959, 1960). Deviations from a Poisson degree distribution suggest the presence of self-organising processes that override random linking in the network;
  - Consequently, the tail of the degree distribution approximately fits a power-law $p(k) \sim k^{-\gamma}$ with scaling parameter $\gamma > 0$ (Barabási et al., 2002). Networks with this exact degree distribution are termed *scale-free networks*;
  - In some cases, the tail fits a power-law with exponential cut-off, $p(k) \sim k^{-\tau} e^{-k/k_c}$, where τ and k_c are constants (Newman, 2001b, 2001d). Deviations from a power-law degree distribution may result from two classes of factors: (i) ageing of nodes, or (ii) the presence of linking costs or limited node capacity (Amaral, Scala, Barthélémy, & Stanley, 2000);

- *Exhibits degree assortativity*: nodes with similar degree tend to connect to each other, i.e., high with high, low with low (Newman, 2002);

- *Are in the class of "small world" networks* proposed by Watts and Strogatz (1998): have short average path length presumably due to the presence of shortcuts provided by inter-hub links;

- *Local clustering* is generated through homophily (Kossinets & Watts, 2009):

- *Induced homophily* due to transitivity[16], that is, given that A knows B, and B knows C, there is a strong likelihood for A to know C as well.
- *Choice homophily* due to focal closure[17] which describes the tendency of researchers to join or form communities/groups signifying specializations on a particular field, topic, or sub-topic.

The ability to achieve transitivity or focal closure depends on the ability of similar others to be aware of each other. This is fundamentally a problem of routing (searching) with local information, that is, a question of where to pass information where it is needed, and at what cost[18] (Kleinberg & Raghavan, 2005). Thankfully, in the case of small world research collaboration networks, such referral-passing or query-passing is largely feasible due to small average path lengths between any two nodes (Kleinberg, 2000; Rosvall, Grönlund, Minnhagen, & Sneppen, 2005). Additionally, the searchability of research collaboration networks is crucial to maintain a level of professionalism and trust by propagating the reputation of others. By making perfect anonymity difficult to attain, deviant and fraudulent activities are to some extent disincentivized (Fafchamps et al., 2006).

Degree centrality.—Nodes can be distinguished based on their *degree centrality* whereby the presence of a high-skew in the overall connectivity distribution implies that there exist nodes that act as hubs on the network (Fatt, Ujum, & Ratnavelu, 2010). Such nodes are prominent structural features as they are fewer in number yet connect a large fraction of nodes. This may have some dramatic implications. For example, it was found that a scale-free network is robust to random node removal (failure) but not against tar-

[16]This mechanism is also termed triadic closure (Rapoport, 1953) or triadic completion (Banks & Carley, 1996).

[17]According to the theory of tie formation based on the confluence of "social interaction foci" known as *Focus Theory*, foci – consisting of various groups, contexts, and activities – organize and facilitate opportunities for interpersonal interactions (Feld, 1981; McPherson, Smith-Lovin, & Cook, 2001).

[18]If nodes are incentivized to pass information, then the total budget depends on the effective branching factor of the network defined as "the average number of new neighbors per node encountered in a breadth-first search".

geted attacks on its hubs (Albert, Jeong, & Barabási, 2000). This is to say that a removal of a node on the periphery of the network has little to no effect in disconnecting or increasing the diameter[19] of a network compared to the removal of a single hub. To some extent, this lends creedance to the expectation that hubs play a prominent role in the overall structure (and functioning) of a network.

Closeness centrality.—Another useful notion is the idea that some nodes are "closer" to other nodes (on average) relative to others. Such nodes with high *closeness centrality* can be thought as occupying a prominent position within the link structure especially when it is important to reach out to as many nodes as possible with few intermediaries. For a connected graph (one where any two nodes can be connected by a path to each other), the closeness centrality is defined as:

$$C_C(v) = \frac{1}{\sum_{u \neq v} \sigma_{uv}} \tag{1.3}$$

Here, σ_{uv} denotes the geodesic (shortest path) distance between node u and v. The smaller the summand, the smaller the denominator, and thus the larger the closeness (reach) of node v to all other nodes on the network.

Betweenness centrality.—Since empirical networks are typically sparse (contain large gaps in the link structure), some nodes have a higher tendency to lie "in-between" the shortest paths connecting most other nodes. If links correspond to information pathways, such nodes are indeed prominently positioned since there is a higher likelihood of information to pass through them compared to other more peripheral nodes. The extent at which a node has this property is measured by the *betweenness centrality* measure given

[19]Defined as the largest shortest distance between two nodes on a network.

by:

$$C_B(v) = \sum_{u \neq v \neq w \in V} \frac{\sigma_{uw}(v)}{\sigma_{uw}} \tag{1.4}$$

Similar to Equation (1.3), σ_{uw} denotes the shortest path between two nodes u and w, while $\sigma_{uw}(v)$ is the shortest path between u and w that includes the target node v.

1.2.2 (b) Google PageRank and HITS algorithm

Co-authorship networks are examples of undirected networks, whereby the directionality of links is unspecified (or deemed irrelevant). Directed networks on the other hand, make an important distinction on which way a link goes, that is, into or out of a node. Examples of directed networks include citation networks in bibliometrics, hyperlink structure between webpages on the world wide web (www), predator-prey relationships on a food web, and so on.

While it can be useful to extend the concept of geocentric, degree, closeness, and betweenness centrality to incorporate the directionality of links (to formulate a directed network version of these measures), there are other notions of centrality that introduce more refined ideas about authority. Here, two examples come to mind – these were specifically designed for ranking the prominence[20] of web pages based on their relative influence as indicated by the structure of webpage in-links and out-links. These examples (in chronological order of appearance in the literature) are the Hyperlink-Induced Topic Search algorithm (HITS) and the Google PageRank algorithm.

HITS.—The HITS algorithm starts by introducing two node scores, one called the *authority score*, $a(i)$, and the other called the *hub score*, $h(i)$, for some arbitrary node i (Kleinberg, 1998, 1999; Gibson, Kleinberg, & Raghavan, 1998). Both scores are given

[20] By "prominent", it is meant the extent to which a node stands out within the structure. Such nodes can be deemed "important" more in terms of their role in the overall structure rather than its level of functioning.

the initial value of 1 across all nodes. The score is then propagated[21] in the following manner:

$$a(i) := \sum_{j \to i} h(j) \tag{1.5}$$

$$h(i) := \sum_{j \leftarrow i} a(j) \tag{1.6}$$

These equations define a kind of circular notion of what constitutes a hub and authority. Equation 1.5 defines an authority as a node that is in-linked by many hubs, while Equation 1.6 defines a hub as a node that out-links to many authorities. The higher the score, the higher the node's attribute of being either an authority or hub.

PageRank.—In contrast, the Google PageRank algorithm computes only one prominence score for each node (Brin & Page, 1998). Its notion of assigning prominence is also circular in the following sense: a node is prominent if it is in-linked by other prominent nodes. If one views node prominence in terms of its affinity in propagating influence on the network, then the PageRank algorithm can be viewed as a method that evaluates nodes based on the influence of their nearest neighbours (separated at a distance of 1 link), which depends on the influence of their next to nearest neighbours (2 links away), and so on. That is, a node is influential to the extent that it influences other influential nodes.

Mathematically, the PageRank score of nodes on the network are modelled as stationary values on an extensive Markov chain (Langville & Meyer, 2006). The algorithm is formulated as the following recursion relation:

$$G(i) := \alpha \sum_{j \to i} \frac{G(j)}{k_j} + \frac{1-\alpha}{N} \tag{1.7}$$

[21] Scores are recursively "propagated" in the sense that the value of one node is computed from the value of other nodes that depend on it.

where $\alpha = 1 - d$, in which $0 < d < 1$ is the damping parameter, and N is the number of vertices (nodes) on the world wide web (corresponding to distinct webpages as identified by their uniform resource locators, or URL, for short) . The damping parameter can be understood in terms of the probability of undergoing $k = 1/(1-d) > 1$ consecutive walks prior to teleporting (jumping) to another webpage elsewhere on the world wide web. In their original formulation, Page and Brin set $d = 6$ which corresponds to a random surfer following on average 6 links before jumping to a fresh URL.

PageRank assigns higher prominence to nodes that influence other influential nodes since the PageRank score of a node i is directly proportional to the summand on the right hand side. This sum is greater when: (a) the number of in-links pointing into node i is large, and (b) the sum of PageRank scores of nodes in-linking to i is large. Condition (b) corresponds to the case where a node is deemed influential because it influences other influential nodes (accordingly, an in-link from a webpage with a low PageRank score contributes less to the overall score of the target webpage). Note that the second term on the right hand side of Equation (1.7) corresponds to the "injection" of uniform probability. This term models the process of exiting the current Markov chain and starting a new chain rooted at some other node on the network.

Both PageRank and HITS are in stark difference from the simple counting of in-links to a given webpage (specifically, the in-degree centrality score), in the sense that a simple in-link count does not factor in qualitative differences across the webpages that do the in-linking since it treats all such in-linking webpages equally. This is of special relevance to a discussion on the networks of research papers and authors which shall be covered in Sections 1.2.3 to 1.2.4.

1.2.2 (c) Input-Output Model and Structural Influence

There are several works that serve as intellectual precursors to the intuition that "a node is important if it receives links from other important nodes" (Kleinberg, 1999). It is therefore appropriate to mention them here. Among the earliest of which is *Leontief's Input-Output model*, which describes input-output flows in the economy of a country in terms of the inter-dependency of its domestic sectors (Leontief, 1941).

Input-Output model.—Consider the case of n sectors, denoted by $S_1, S_2, \ldots, S_n$, each producing a unique product (hence, there are n unique products), and furthermore, suppose that consumption equals production across the board (i.e. input equals output). If a_{ij} represents the number of units produced by sector S_i required to produce one unit by sector S_j, d_i is the total number of externally demanded units of S_i (not consumed by any sector), then x_i is the total output of industry S_i such that:

$$\begin{aligned} x_1 &= a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n + d_1 \\ x_2 &= a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n + d_2 \\ &\cdots \\ x_n &= a_{n1}x_1 + a_{n2}x_n + \cdots + a_{nn}x_n + d_n \end{aligned} \tag{1.8}$$

Using matrix notation, one may write:

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix}, d = \begin{pmatrix} d_1 \\ \vdots \\ d_n \end{pmatrix}, x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \tag{1.9}$$

so that,

$$x = Ax + d \tag{1.10}$$

Here, A is termed the input-output matrix, d is the final demand vector, and x is the total output vector. Rewriting Equation (1.10), one obtains:

$$(I - A)x = d \tag{1.11}$$

Provided that the matrix $(I - A)$ is invertible, then what results is a system of linear equations with a unique solution. Given the values of the final demand vector, the required output levels can be determined. Additionally, if the principal minors of $(I - A)$ are all positive, the required output vector x is strictly non-negative; this is known as the Hawkin-Simons condition (Hawkins & Simon, 1949). That is,

$$x = (I - A)^{-1}d \tag{1.12}$$

The total output vector can be treated as a prominence score for each node (sector) in the economy, whereby the highest production levels are attributed to sectors that coincide with the highest direct and indirect dependency flows.

Structural influence.—With a few modifications, this model can be used as a basis for determining cliques[22] in a social network (Forsyth & Katz, 1946; Luce & Perry, 1949). Such works lead to the class of *structural influence* models in bibliometrics which are aimed at finding the most prominent journals within the structure of inter-journal influence (Salancik, 1986; Johnson & Podsakoff, 1994; Baumgartner & Pieters, 2003; Wakefield, 2008). The basic idea is that of information transmission along network ties (social or otherwise). Take for example the propagation of a rumour. In this, there are two

[22] According to Hubbell (1965), "A clique can be intuitively defined as a subset of members who are more closely identified with one another than they are with the remaining members of their group."

important considerations. First, some nodes (individuals) are more influential than others and hence are more effective at propagating rumours. Second, given that social connectivity is typically inhomogeneous from one node (individual) to the next, the rumour is likely to be shared between members within the same clique rather than between those associated to different cliques (Hubbell, 1965).

These effects can be modelled as follows. As shown by Festiger (1949), given a binary matrix C (in which its elements are either 0 or 1), the element on the i-th row and j-th column corresponding to the k-th power of C gives the number of walks (chains) of length k that can be traced from node i through intermediaries to j. That is,

$$\text{\#walks of length-}k\text{ spanning }i\text{ to }j = (C_{ij})^k \tag{1.13}$$

Here, C encodes the adjacency (connectivity) of nodes on the social network such that nodes i and j are connected if and only if $C_{ij} = 1$, unconnected if $C_{ij} = 0$, and $C_{ii} = 0$. Note that for "influence" problems on social networks, the adjacency matrix is generally asymmetric and therefore the associated network contains unreciprocated links.

To find the extent at which a rumour is transmitted on the social network given by adjacency matrix C, one needs to further consider that the rate of propagation not only depends on structural details of the underlying social network but may also depend on the context of the rumour, as well as the appeal of that rumour to specific groups. To this end, a rumour can be treated as a signal by introducing a parameter $0 \leq a \leq 1$ that idealises the non-attenuation of the signal across links, whereby complete attenuation (weakening) is given by $a = 0$, and the absence of attenuation is given by $a = 1$. The *transmissibility* of the signal can then be written as the following matrix equation (Katz, 1953):

$$T = aC + a^2C^2 + \cdots + a^kC^k + \cdots = (I - aC)^{-1} - I \tag{1.14}$$

where T has elements t_{ij}, with column sums $t_i = \sum_i t_{ij}$. Let t be a column vector with elements t_i and u be a column vector with unit elements. Hence, $t' = u'[(I - aC)^{-1} - I]$. By multiplying the right hand side of Equation (1.14) by $(I - aC)$ the following expression is obtained:

$$t'(I - aC) = u' - u'(I - aC) = au'C \tag{1.15}$$

Transposing this equation gives:

$$(I - aC')t = aC'u \tag{1.16}$$

Since $C'u$ is the column vector whose elements are the column sums of C (since they are row sums of C', and recalling that u is a column unit vector), one may write $C'u = s$ so that:

$$\left(\frac{1}{a}I - C'\right)t = s \tag{1.17}$$

Given a, C, and s, one may numerically solve[23] the system of linear equations above to obtain t (column sum vector for the transition matrix T that underlies the signal transmission process). This dispenses with the need to compute powers of C and the infinite matrix sum in Equation (1.14). The column vector t gives a measure of the prominence of a node within a structure of influence ties. A modification of this method for finding influential journals is as discussed and demonstrated in Section 3.2.

1.2.3 Citation network of research papers

If each research paper can be viewed as a distinct publication event triggered by a set of preceding events, then a document citation network can be viewed as a web of

[23] Using Gaussian elimination.

interconnected publication events (Garfield, 1970). More precisely, a citation network is a directed graph with nodes representing papers and directed links representing citations from citing articles to cited articles. The number of citations to a specific paper therefore corresponds to the in-degree of the node representing it on the document citation network (Chen et al., 2007). Before going deeper into what citation networks can tell us, it is important to clarify just what a citation entails. According to Egghe and Rousseau (1990):

> [...] a reference is the acknowledgement that one document gives another, while a citation is the acknowledgement that one document receives from another. So, 'reference' is a backward-looking concept while 'citation' is a forward-looking one.

Acknowledgements to past works are made to recognize, support, challenge, or refute those works (Hanney et al., 2005) in various degrees, varying from the thorough to the perfunctory (Krampen, Becker, Wahner, & Montada, 2007). Citation diversions may also occur. These correspond to the "citing of content but the altering of its meaning in a manner that diverts its implications" (Greenberg, 2009). By taking these ambiguities into account, it becomes clear that highly-cited papers confers popularity (Redner, 1998) – in the sense of fame or infamy – but not necessarily authority.

One of the first prototypical studies on citation networks was conducted by Garfield, Sher, and Torpie (1964) to map the chronological development and interdependency of intellectual milestones leading to the discovery of DNA[24]. With the advent of network science, theoretical studies on the global structure and dynamics of citation networks were explored (Bilke & Peterson, 2001; Vázquez, 2001; Jeong, Néda, & Barabási, 2003; Hajra & Sen, 2005). While other studies were conducted to ascertain significant small-, as

[24] According to Garfield et al. (1964), "[...] the use of citation data for constructing historical maps was given great impetus by Dr. Gordon Allen when he prepared a bibliographic citation network diagram demonstrating the chronological relationship and citational linkages among a group of papers on the staining of nucleic acids. Allen's citation network diagram provided a useful model of scientific literature and simultaneously provided, in a two-dimensional topological display, the historical development of the subject matter covered by the fifteen papers in his bibliography."

well as intermediate-scale details of large empirical networks; e.g., the Physical Review family of journals (Chen et al., 2007; Chen & Redner, 2010), the field of sustainability science (Kajikawa, Ohno, Takeda, Matsushima, & Komiyama, 2007), and research literature on organic LEDs (Kajikawa & Takeda, 2009), among others.

From an informetric standpoint, some nodes are more prominent than others due to their position within a structure of relationships. This positional advantage can be estimated directly from a network using a number of appropriate node centrality measures, e.g. closeness (Sabidussi, 1966), betweenness (Freeman, 1977), constraint (Burt, 1995), hub and authority score (Kleinberg, 1999), PageRank (Brin & Page, 1998), etc. While such measures are applicable to directed or undirected networks in general (social, informational, technological, or biological), they are particularly useful in citation networks as they can be used to differentiate works by their intellectual significance (Redner, 2005; Chen et al., 2007) as well as by their function in the literature (Rosvall, Axelsson, & Bergstrom, 2009; Rosvall & Bergstrom, 2010; Chen & Redner, 2010; Herrera, Roberts, & Gulbahce, 2010).

As an aside, a recent preprint by Bertsimas, Brynjolfsson, Reichman, and Silberholz (2014) demonstrates how logistic regression models can be used to prospectively predict the future "value" of a researcher based on data available at the time of publication. This is done by taking advantage of robust features in the document citation network in conjunction with the co-authorship network. These features typically characterise highly cited papers/authors. As an interesting application, the Bertsimas and co-workers show how this method can be used to assess a young researcher's future impact using data from the first 5 years of his/her career. This is similar in spirit to one of my goals in this thesis.

1.2.4 Citation network of authors

A similar approach to mapping citations between documents can also be done between authors of documents. Such networks are termed author citation networks. In comparison to document citation networks, there are relatively fewer works done on this topic. One such paper found that excluding self-citations, 43.8% of co-authoring and 13.5% of non-coauthoring researchers tend to reciprocate citations in 116 years of Physical Review articles spanning 1893 to 2009 (Martin, Ball, Karrer, & Newman, 2013).

In another paper by Ding (2011b), it was found that "[. . .] productive authors tend to directly coauthor with and closely cite colleagues sharing the same research interests; they do not generally collaborate directly with colleagues having different research topics, but instead directly or indirectly cite them; and highly cited authors do not generally coauthor with each other, but closely cite each other".

In a paper by Radicchi, Fortunato, Markines, and Vespignani (2009), the authors modelled the spreading of scientific credit as a diffusion process, specifically, a biased random walk combined with random credit distribution between nodes. The resulting PageRank-like algorithm, dubbed the *Science Author Rank Algorithm* (SARA), was used as a basis for ranking researchers of Physical Review articles spanning the years 1893 to 2006. As a benchmark, 16 of the top 20 ranked scientists based on papers published and cited in 1967–1974, and similarly, 6 in 2003–2004 were found to be recipients of prestigious prizes in physics[25].

Elsewhere, Życzkowski (2010) defined the "weighting factor" of a scientist using components of the normalized leading eigenvector of the coupling matrix for any given author citation network, which is incidentally a method similar in construction to the

[25]This includes the Nobel Prize, Wolf Prize, Boltzmann Medal, Dirac Medal, and Planck Medal. For the period 1967–1974, 12 of the top 20 ranked scientists are recipients of the Nobel Prize in Physics. In contrast, there are only two Nobel laureates, i.e. P.W. Anderson and S. Weinberg, who make the cut in the 2003–2004 test period. The SARA rankings are available online at http://www.physauthorsrank.org/authors/show.

PageRank algorithm.

Additional work on author citation networks was explored in Ding (2011a). In this paper, popularity and prestige scores were computed based on ISI-indexed articles published by scholars in Information Retrieval (IR) from 1956–2008. The results indicated that popularity rank and prestige rank[26] were highly correlated with a weighted PageRank score.

[26] According to Ding (2011a), the popularity of a researcher is defined as "the number of times he is cited (endorsed) in total, and prestige as the number of times he is cited by highly cited papers."

CHAPTER 2

METHODOLOGY

The premise of this thesis is that experts and academic icons can be identified from a large sample of researchers by analysing networks constructed from their publication and citation data. The general idea is to rank researchers by some relative influence score that is determined based on who and how they cite. Such inter-author citation dependencies (linkages) define an author citation network (ACN). The relative influence score of a researcher then corresponds to the centrality score of his/her representative node on the author citation network, using some algorithm that designates nodes as influential if they influence other influential nodes on the network. By sorting the centrality scores from highest to lowest value, we can then associate the top X ranks to the most expert or authoritative researchers ($X \ll N$, the total sample size). However, there are a number of steps required to arrive at this ranking. This chapter will describe these steps, beginning with a description of the definitions and notation used.

2.1 Definitions and notation

In order to put forward the concepts used in this work, we need to draw on the framework of network theory for both the representation of connections in bibliometric data and its analysis. This section shall focus on clarifying network-theoretic (graph-theoretic) definitions and notations used throughout this thesis.

2.1.1 Basic definitions

In the simplest sense, a network describes the connectivity between objects as an abstract configuration of dots called nodes (representing those objects) and connections between those nodes signified by lines called links. While this representation has a natural

visual or graphical quality, it is by no means limited by it, since its structure and its manipulation can be defined by purely algebraic means.

Definition 1 (Network)**.** *A network, or graph, is a mapping of pairwise relationships, connections, or ties between a set of objects (entities).*

Mathematically, a network can be represented by a graph G consisting of a set of nodes (vertices) V, joined by lines representing links (edges) E. This is formally expressed as $G = (V, E)$. The number of nodes is given by $|V| = N$, while the number of links is given by $|E| = M$. Each node represents a specific object or entity. Accordingly, each link represents a specific connection between a pair of objects or entities.

Each link must be anchored between two nodes, hence, there can be no dangling links. Furthermore, we can associated to each graph a $N \times N$ binary adjacency matrix, $A = (a_{ij})$ with elements $a_{ij} = 1$ if there exists a single link joining node i to j, otherwise $a_{ij} = 0$. For graphs without loops (without self-linking nodes), the diagonal of A is zero, that is, $a_{ii} = 0$.

Definition 2 (Undirected graph)**.** *A graph with exclusively symmetric links between its nodes (in the sense that the directionality of links are not specified). Unless otherwise specified, the term "graph" refers to "undirected graph".*

Definition 3 (Simple graph)**.** *A graph is simple if it is unweighted, undirected, and contains no loops or multiple edges.*

Undirected graphs are used to map symmetric relationships between a set of objects. As such, it follows that entries on the adjacency matrix A for undirected graphs have the property that $a_{ij} = a_{ji}$. For a simple graph, only one link is permitted between any pair of connected nodes, that is, $a_{ij} = 1$ and $a_{ji} = 1$ refer to the same link. Accordingly, a simple graph with N nodes may have up to $\binom{N}{2} = N(N-1)/2$ links.

Definition 4 (Directed graph)**.** *A graph with exclusively asymmetric links between its nodes. Unless otherwise specified, a "directed graph" will hereon be referred to as a "digraph".*

Digraphs are used to map flow relationships between a set of objects, whereby an out-going link from node i to j is not necessarily reciprocated by an in-coming link back from node j to i. Accordingly, entries on the adjacency matrix for directed graphs are generally asymmetric, that is, $a_{ij} \neq a_{ji}$ for $i \neq j$. In the case of a simple digraph, for any pair of connected nodes there can be at most one in-coming link and one out-going link. Hence, a simple digraph with N nodes may have up to $N(N-1)$ links.

The adjacency matrix can be generalised as a connection matrix W whereby each matrix element w_{ij} signifies the connection strength between node i to j. Depending on the purpose, these weights can either be real or complex, and can either be strictly nonnegative or span a range of positive and negative values. It follows that a normalised connection matrix is stochastic.

Definition 5 (Complete graph)**.** *A simple undirected graph where every pair of distinct nodes are connected by a unique link.*

Definition 6 (Complete digraph)**.** *A directed graph where every pair of distinct nodes are connected by a pair of unique links (one in each direction).*

Definition 7 (Subgraph)**.** *A subgraph of a graph $G = (V, E)$ is a graph $G' = (V', E')$ with node set $V' \subseteq V$ and link set $E' \subseteq E$.*

2.1.2 Network properties

Each network is defined by the configuration (structure) of its nodes and links. Qualitative and quantitative attributes may be assigned to individual nodes and links, or groups

of nodes and links, as well as the entire network. Such attributes characterise the local, meso (intermediate), and global scale properties of a network.

2.1.2 (a) Local properties

Definition 8 (Degree)**.** *The degree k of a node is given by the number of links connected to it. For digraphs, we shall distinguish the number of in-links and out-links by its "in-degree" and "out-degree", respectively.*

Definition 9 (Distance)**.** *The distance between two nodes i and j is given by the number of intermediate links between them, denoted by $d(i, j)$.*

Definition 10 (Geodesic)**.** *The shortest paths between pairs of nodes on a network are known as geodesics. The shortest path length between nodes i and j is denoted by σ_{ij}.*

Definition 11 (Clustering coefficient)**.** *The clustering coefficient C is the ratio of existing links between a node's nearest neighbours relative to the maximum number of inter-neighbour links. For a node i with k links and e inter-neighbour links, the clustering coefficient is computed as $C_i = 2e/[k_i(k_i - 1)]$.*

2.1.2 (b) Global properties

Definition 12 (Size)**.** *The size of a network is given by the number of nodes $|V| = N$.*

Definition 13 (Density)**.** *The density ρ of a network is defined as the ratio of the number of edges $|E| = M$ to the number of possible edges, given by the binomial coefficient $\binom{N}{2}$, so that $\rho = 2M/[N(N - 1)]$.*

Definition 14 (Average degree)**.** *The average number of links per node. It is defined as $\langle k \rangle = 2M/N$.*

Definition 15 (Diameter)**.** *The diameter D of the graph is given by the longest geodesic, i.e. the shortest path length between the most two distant nodes on the network.*

2.2 Data

The study dataset consists of 126 ISI-indexed journals within the Journal Citation Reports (JCR) subject category of "BUSINESS, FINANCE" from 1980 to 2011[1]. The parameters for the source data are as tabulated in Table 2.1.

Table 2.1: Source data parameters

Source	ISI *Web of Knowledge*, "BUSINESS, FINANCE" SSCI subject category
Download date	25^{th} June 2012
Document type(s)	Article
Years covered	1980 – 2011
Time period	32 years
No. of source journals	126
Total number of articles	62,467

2.2.1 Processing the data

The source data must be preprocessed before it can be used. To this end, we shall borrow the guidelines set by information visualisation wizard, Ben Fry, for handling data. This consists of the following phases (Fry, 2007) :

Acquire	Obtain the data, whether from a file on a disk or a source over a network.
Parse	Provide some structure for the data's meaning, and order it into categories.
Filter	Remove all but the data of interest.
Mine	Apply methods from statistics or data mining as a way to discern patterns or place the data in mathematical context.
Represent	Choose a basic visual model, such as a bar graph, list, or tree.
Refine	Improve the basic representation to make it clearer and more visually engaging.
Interact	Add methods for manipulating the data or controlling what features are visible.

[1]Non-ISI databases are not used in this study. The reason for this is that other data providers – such as Scopus – while broader in source journal coverage, do not have the required depth in time coverage. Furthermore, we find that ISI data have highly consistent structure, making it amenable for automated analysis.

For the purpose of this thesis, we shall only focus on the first four phases, that is, *Acquire*, *Parse*, *Filter*, and *Mine*. We can safely dispense with the last three phases (i.e. *Represent*, *Refine*, and *Interact*), since these are relevant only in the construction of an interactive information visualisation software tool, which we do not go into here.

2.2.1 (a) Acquire

This phase consists of the acquisition of ISI data – hereon referred to simply as the "study dataset" – which we source via the *Web of Knowledge* database interface. While the effectiveness of any data-driven analysis is necessarily dependent on the method(s) used, they are also conditional on the completeness and quality of the source data. With respect to the former, the boundary specification of the data must be scrutinised.

As shown in Table 2.1, we have made several deliberate choices in the time coverage (1980–2011), source journal selection (ISI-indexed journals in the category "BUSINESS, FINANCE"), and document type (article). Any publications outside of these boundaries are effectively neglected by our subsequent analyses. The next crucial step is to interpret the data (the *Parse* phase). This is then followed by a determination of how much of the data is usable for the intended analysis (the *Filter* phase). Filtering is necessary to shed some light on the conditions under which any results are obtained.

2.2.1 (b) Parse

In this phase, we scan through the study dataset to interpret its implicit structure. For Export Format Version 1.0, this is very straightforward since the data fields for each publication record are marked by a 2-character tag (AU, PY, J9, etc), followed by a space, and then by the corresponding data value which typically consists of either a character string or integer.

Each record in ISI data represents a specific publication. The attributes of each record are specified using a two character field tag, as shown in Figure 2.1. This is

called the **ISI Export Format Version 1.0**. A sample of a "complete" ISI record is as shown in Figure 2.2. An important feature of this data is that it answers the *who*, *what*, *where*, *when*, *which*, and *how many* aspects of a specific publication (signifying the multidimensional nature of bibliometric data). To wit:

Who?	Author field (AU, AF), responding author (RP)
What?	Title of publication (TI), document type (DT), language (LA), abstract (AB), keywords (ID), database accession number (UT)
Where?	Source journal name (SO, J9), source journal ISSN (SN), author address (C1)
When?	Publication year (PY)
Which?	Volume (VL), issue (IS), beginning page (BP), end page (EP), references used (CR), field of study (WC, SC)
How many?	Times cited (TC)

While each record is given a unique accession number (UT), these are not expressly linked to other records. For example, the record marked WOS:000168780100005 (A'Hearn B, 2001, J Monetary Econ, V47, P321) actually refers to the record corresponding to WOS:000074353500004 (Burnside C, 1998, J Monetary Econ, V41, P513) in the cited reference field, CR. Such connections are not made explicit in the data.

As we will find, the ability to connect one record to other records – specifically citation linkages between articles – is an important feature which we will need to build ourselves. In the simplest sense, citations are expressed in tuples of the form:

(*source_citing_article*, *target_cited_article*)

By parsing the data, we can extract links between ISI records. We will first need to index accession numbers for all publications in the study dataset. As previously mentioned, these are denoted by the data field (line) tagged as UT. An indexing scheme can be built using a combination of available tags that sufficiently identify any given article.

A simple choice to use for a unique identifier (ID string) is:

$$\text{UT} \equiv \text{AU[1], PY, J9, VL, BP}$$

$$\text{WOS:000074353500004} \equiv \text{BURNSIDE C, 1998, J MONETARY ECON, V41, P513}$$

where string elements AU[1] denotes the first author listed in the author field tag AU, PY is the publication year, J9 is the 29-character source journal abbreviation (in contrast to the complete source journal name field, SO), VL is the source journal volume, and BP denotes the beginning page number of the article. This is a deliberate choice as article cited references in C1 are recorded in exactly this format (see Figure 2.2).

The construction of such an index makes it convenient to look-up the accession number for entries in the C1 field. It must be noted however that the C1 field lists all items in the publication's bibliography, some of which may not be ISI-indexed (lies outside the boundary specification of the dataset).

FN	File Name	ZR	"Total Times Cited Count (WoS, BCI, and CSCD)"
VR	Version Number	NR	Cited Reference Count
PT	Publication Type (J=Journal; B=Book; S=Series)	TC	Times Cited
AU	Authors	PU	Publisher
AF	Author Full Name	PI	Publisher City
BA	Book Authors	PA	Publisher Address
CA	Group Authors	WC	Web of Science Category
GP	Book Group Authors	SC	Subject Category
TI	Document Title	SN	International Standard Serial Number (ISSN)
RID	ResearcherID Number	BN	International Standard Book Number (ISBN)
BE	Editors	D2	Book Digital Object Identifier (DOI)
SO	Publication Name	J9	29-Character Source Abbreviation
SE	Book Series Title	JI	ISO Source Abbreviation
BS	Book Series Subtitle	PD	Publication Date
LA	Language	PY	Year Published
DT	Document Type	VL	Volume
CT	Conference Title	IS	Issue
CY	Conference Date	PN	Part Number
HO	Conference Host	SU	Supplement
CL	Conference Location	SI	Special Issue
SP	Conference Sponsors	BP	Beginning Page
DE	Author Keywords	EP	Ending Page
ID	Keywords Plus®	AR	Article Number
AB	Abstract	PG	Page Count
C1	Author Address	P2	Chapter Count in a Book
RP	Reprint Address	DI	Digital Object Identifier (DOI)
EM	E-mail Address	GA	Document Delivery Number
FU	Funding Agency and Grant Number	UT	Accession Number
FX	Funding Text	ER	End of Record
CR	Cited References	EF	End of File

Figure 2.1: ISI data field tags

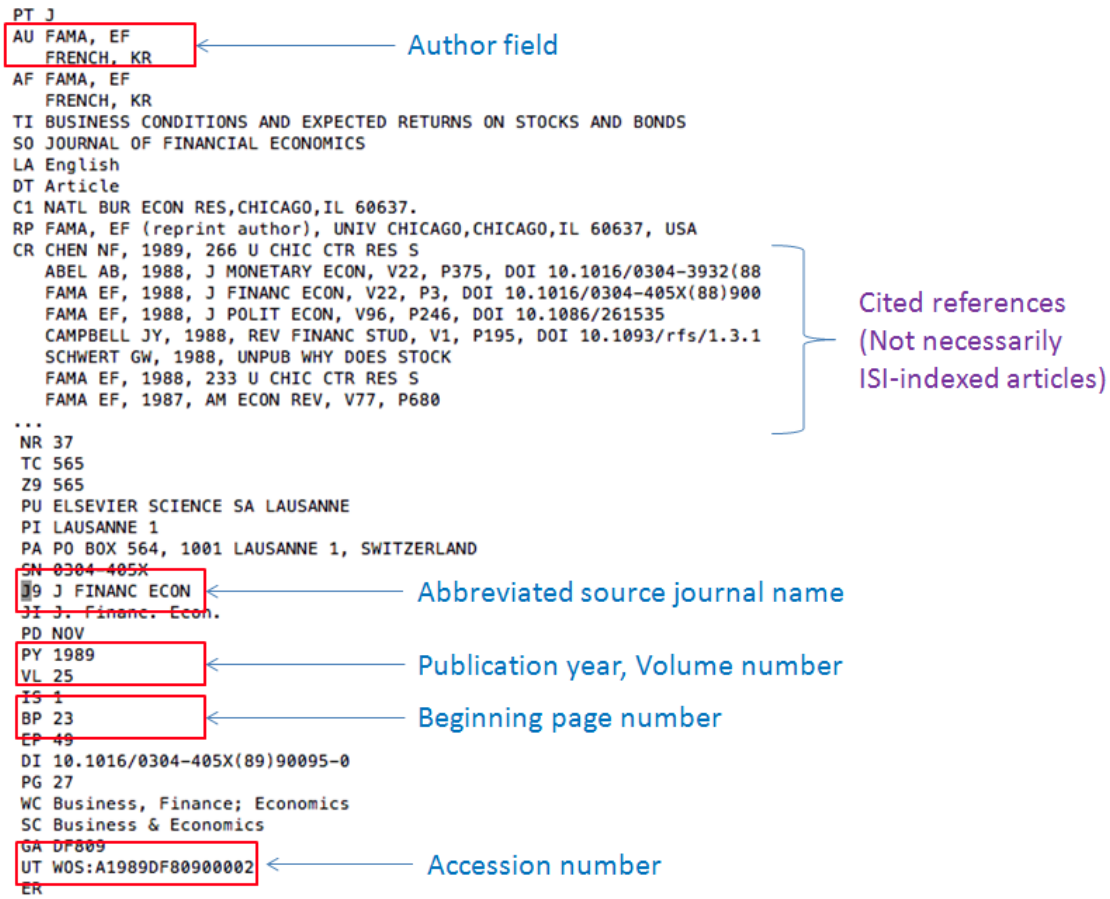

Figure 2.2: Sample ISI data

2.2.1 (c) *Filter*

The next phase involves removing irrelevant portions of the data. Accordingly, all five string elements in the ID string must be non-empty to uniquely identify any given article. For example, if each string element is specified except the beginning page, we may not be able to differentiate between two articles associated to the same first author who published in the same journal, year, and volume. Disambiguation issues will occur for any cited reference where there is one or more missing elements from the ID string.

Another potential issue occurs when a cited reference listed in CR does not match the expected ID string reconstructed from the AU, PY, J9, VL, and BP fields. For example, the publishing year is mistakenly out of source data bounds (1986 is recorded as 1896). Such typographical errors in CR are easier to exclude than to rectify. In a sense, this is a good thing as the resulting citation links are maximally filtered from ambiguous or erroneous ties. On the downside, this may underreport the frequency of citations effected by such errors[2] thus underestimating the "actual impact" of that citation.

Article coverage—As mentioned above, we can only extract connections between articles that have proper author (AU), publication year (PY), abbreviated source journal name (J9), volume (VL), beginning page (BP), and cited reference (CR) fields[3]. A detailed breakdown of article coverage by source journal is as shown in Table 2.2.

Out of the 62,467 articles available in the study dataset, approximately 99% (61,848) were deemed usable during the data preparation phase. This means that 624 articles are left out of the analysis consisting of: (i) 237 articles published anonymously (from which unique authorship cannot be ascertained); (ii) 381 articles without a cited reference field[4];

[2]See Simkin and Roychowdhury (2005). Sometimes the process of citation copying will propagate improperly specified citations.

[3]These five fields uniquely identify articles that are cited (have in-links on the citation graph), while the the CR field is necessary to determine references to other articles (corresponding to out-links on the citation graph).

[4]Of these CR-less articles, 360 are rightfully ignored as they have zero citations at download-date and are therefore unconnected to any other ISI publication in the entire *Web of Knowledge* database. Such

and (iii) 6 articles with all six required fields but were somehow incorrectly left out by the data filtering program. The $\sim$1% loss in data is assumed to have negligible impact on the analysis employed.

Journal coverage—We now need to determine the extent to which journals covered in the study dataset are distributed. By virtue of Bradford's law Bradford (1985), journals within a research field – when arranged by decreasing order of number of publications – can be arranged into three equally sized groups (referred to as *Bradford zones*) according to the ratio $1 : n : n^2$, where n denotes the proportional size of the partition.

Based on the source data, it was found that approximately 3.2% of "BUSINESS, FINANCE" journals make up a third of the total number of articles, 12.9% the second one-third, and the remaining 83.9% makes up the rest (corresponding to a ratio of approximately $1 : 4 : 16$).

While not an exact empirical law, Bradford's law attempts to quantify the observation that some journals have a larger footprint compared to others within the same field. In effect, this disproportionate concentration of publishing activity may cause the some researchers to limit their literature search primarily within the first two thirds, which when done collectively in the community – and over a protracted period in time – contributes to the obscuration of articles (and journals) that lie in the tail of the Bradford distribution.

As it is not necessarily the case that journals with the most items host articles most worthy of readership Nicolaisen and Hjørland (2007), citation-based measures like the impact factor reveal useful insight into the magnitude of impact (relative to number of articles contributed) within a discipline or its constituent subareas.

---

instances if included, would correspond to isolates on the document citation network. The remaining 21 CR-less articles with non-zero citations were overlooked by the strict CR inclusion criteria (these correspond to articles with a number of in-links maximally bounded by the number of citations received and strictly no out-links). This could affect the analysis if those citations are actually contained within "Business, Finance" indexed journals over the period of interest (19 are cited once, while the remaining three are cited 2, 3, and 8 times, respectively). In hindsight, the CR inclusion criteria is an unnecessary complicating step.

At the time of writing, the 2011 JCR Social Science Edition gives these journals a median impact factor of 0.799. The top five journals by (decreasing) impact factor are *Review of Financial Studies* (4.748), followed by *The Journal of Finance* (4.218), *Journal of Financial Economics* (3.725), *Journal of Accounting and Economics* (3.281), and *Accounting, Organizations and Society* (2.878). Hence, we can expect a sizeable fraction of citations within the study dataset to be contained within these five journals.

Table 2.2: Coverage of articles and citations within the "Business, Finance" study dataset. See text for details.

Source journal name	Abbreviated form	#Articles	#Articles covered	%Articles covered	#Cited articles	%Cited articles	Cites received	#Cites covered	%Cites covered
ABACUS-A JOURNAL OF ACCOUNTING AND BUSINESS STUDIES	ABACUS-J ACCOUNT BUS	163	163	100.00	38	23.31	285	93	32.63
ABACUS-A JOURNAL OF ACCOUNTING FINANCE AND BUSINESS STUDIES	ABACUS	111	111	100.00	34	30.63	191	74	38.74
ABACUS-NEW YORK	ABACUS-NEW YORK	125	124	99.20	2	1.60	70	3	4.29
ACCOUNTING AND BUSINESS RESEARCH	ACCOUNT BUS RES	82	82	100.00	33	40.24	109	56	51.38
ACCOUNTING AND FINANCE	ACCOUNT FINANC	188	188	100.00	60	31.91	210	98	46.67
ACCOUNTING HORIZONS	ACCOUNT HORIZ	83	83	100.00	24	28.92	110	49	44.55
ACCOUNTING ORGANIZATIONS AND SOCIETY	ACCOUNT ORG SOC	875	875	100.00	679	77.60	13041	4300	32.97
ACCOUNTING REVIEW	ACCOUNT REV	1212	1212	100.00	989	81.60	23206	13503	58.19
AREUEA JOURNAL-JOURNAL OF THE AMERICAN REAL ESTATE & URBAN ECONOMICS ASSOCIATION	AREUEA J	235	235	100.00	178	75.74	2879	1443	50.12
ASIA-PACIFIC JOURNAL OF ACCOUNTING & ECONOMICS	ASIA-PAC J ACCOUNT E	63	63	100.00	2	3.17	14	4	28.57
ASIA-PACIFIC JOURNAL OF FINANCIAL STUDIES	ASIA-PAC J FINANC ST	184	184	100.00	62	33.70	145	92	63.45
AUDITING-A JOURNAL OF PRACTICE & THEORY	AUDITING-J PRACT TH	462	462	100.00	323	69.91	3851	2149	55.80
AUSTRALIAN ACCOUNTING REVIEW	AUST ACCOUNT REV	116	116	100.00	22	18.97	60	26	43.33
BANKING LAW JOURNAL	BANKING LAW J	620	619	99.84	49	7.90	298	50	16.78

continued on next page . . .

... Table 2.2 continued from previous page

Source journal name	Abbreviated form	#Articles	#Articles covered	%Articles covered	#Cited articles	%Cited articles	Cites received	#Cites covered	%Cites covered
BARCLAYS REVIEW	BARCLAYS REV	57	57	100.00	0	0.00	3	0	0.00
BRITISH TAX REVIEW	BRIT TAX REV	245	245	100.00	0	0.00	40	0	0.00
BULLETIN FOR INTERNATIONAL FISCAL DOCUMENTATION	B INT FISCAL DOC	249	249	100.00	13	5.22	24	14	58.33
CONTEMPORARY ACCOUNTING RESEARCH	CONTEMP ACCOUNT RES	274	274	100.00	185	67.52	2070	1279	61.79
CORPORATE GOVERNANCE-AN INTERNATIONAL REVIEW	CORP GOV-OXFORD	13	13	100.00	0	0.00	12	0	0.00
EMERGING MARKETS REVIEW	EMERG MARK REV	74	74	100.00	38	51.35	139	64	46.04
EUROPEAN ACCOUNTING REVIEW	EUR ACCOUNT REV	139	139	100.00	64	46.04	403	198	49.13
EUROPEAN FINANCIAL MANAGEMENT	EUR FINANC MANAG	190	190	100.00	122	64.21	780	351	45.00
EUROPEAN JOURNAL OF FINANCE	EUR J FINANC	141	141	100.00	23	16.31	79	30	37.97
FEDERAL RESERVE BANK OF ST LOUIS REVIEW	FED RESERVE BANK ST	165	164	99.39	59	35.76	535	125	23.36
FINANCE A UVER	FINANC A UVER	219	219	100.00	36	16.44	138	57	41.30
FINANCE A UVER- CZECH JOURNAL OF ECONOMICS AND FINANCE	FINANC UVER	33	33	100.00	7	21.21	27	12	44.44
FINANCE A UVER-CZECH JOURNAL OF ECONOMICS AND FINANCE	FINANC UVER	197	197	100.00	45	22.84	189	58	30.69
FINANCE AND STOCHASTICS	FINANC STOCH	258	258	100.00	156	60.47	2549	601	23.58
FINANCE AND TRADE REVIEW	FINANC TRADE REV	12	12	100.00	0	0.00	3	0	0.00
FINANCE RESEARCH LETTERS	FINANC RES LETT	106	106	100.00	20	18.87	73	22	30.14

continued on next page ...

... Table 2.2 continued from previous page

Source journal name	Abbreviated form	#Articles	#Articles covered	%Articles covered	#Cited articles	%Cited articles	Cites received	#Cites covered	%Cites covered
FINANCIAL ANALYSTS JOURNAL	FINANC ANAL J	404	404	100.00	250	61.88	2139	1121	52.41
FINANCIAL MANAGEMENT	FINANC MANAGE	951	951	100.00	644	67.72	8565	4048	47.26
FINANZARCHIV	FINANZARCHIV	136	136	100.00	23	16.91	180	31	17.22
FISCAL STUDIES	FISC STUD	194	194	100.00	50	25.77	780	81	10.38
FORBES	FORBES	12525	12176	97.21	6	0.05	514	6	1.17
GENEVA PAPERS ON RISK AND INSURANCE THEORY	GENEVA PAP RISK INS	110	110	100.00	55	50.00	508	153	30.12
GENEVA PAPERS ON RISK AND INSURANCE-ISSUES AND PRACTICE	GENEVA PAP R I-ISS P	376	376	100.00	111	29.52	636	202	31.76
GENEVA RISK AND INSURANCE REVIEW	GENEVA RISK INS REV	60	60	100.00	25	41.67	97	39	40.21
HOUSING FINANCE REVIEW	HOUSING FINANC REV	116	116	100.00	57	49.14	447	238	53.24
IKTISAT ISLETME VE FINANS	IKTISAT ISLET FINANS	149	149	100.00	22	14.77	56	31	55.36
IMF ECONOMIC REVIEW	IMF ECON REV	26	26	100.00	6	23.08	27	6	22.22
IMF STAFF PAPERS	IMF STAFF PAPERS	266	266	100.00	114	42.86	1727	288	16.68
INSTITUTIONAL INVESTOR	INST INVESTOR	1023	986	96.38	6	0.59	30	6	20.00
INTERNATIONAL FINANCE	INT FINANC	78	78	100.00	15	19.23	118	20	16.95
INTERNATIONAL INSOLVENCY REVIEW	INT INSOLV REV	31	31	100.00	1	3.23	3	1	33.33
INTERNATIONAL JOURNAL OF CENTRAL BANKING	INT J CENT BANK	104	104	100.00	29	27.88	144	42	29.17
INTERNATIONAL JOURNAL OF FINANCE & ECONOMICS	INT J FINANC ECON	326	326	100.00	128	39.26	1514	329	21.73

continued on next page ...

... Table 2.2 continued from previous page

Source journal name	Abbreviated form	#Articles	#Articles covered	%Articles covered	#Cited articles	%Cited articles	Cites received	#Cites covered	%Cites covered
INTERNATIONAL JOURNAL OF HEALTH CARE FINANCE & ECONOMICS	INT J HEALTH CARE FI	73	73	100.00	8	10.96	87	10	11.49
INTERNATIONAL MONETARY FUND STAFF PAPERS	INT MONET FUND S PAP	486	486	100.00	306	62.96	6499	1167	17.96
INTERNATIONAL REVIEW OF ECONOMICS & FINANCE	INT REV ECON FINANC	237	237	100.00	84	35.44	352	132	37.50
INVESTMENT ANALYSTS JOURNAL	INVEST ANAL J	48	48	100.00	0	0.00	16	0	0.00
JASSA-THE FINSIA JOURNAL OF APPLIED FINANCE	JASSA	84	84	100.00	0	0.00	2	0	0.00
JOURNAL OF ACCOUNTANCY	J ACCOUNTANCY	1476	1469	99.53	81	5.49	401	106	26.43
JOURNAL OF ACCOUNTING & ECONOMICS	J ACCOUNT ECON	640	640	100.00	577	90.16	23800	14475	60.82
JOURNAL OF ACCOUNTING AND PUBLIC POLICY	J ACCOUNT PUBLIC POL	296	296	100.00	158	53.38	1113	475	42.68
JOURNAL OF ACCOUNTING RESEARCH	J ACCOUNT RES	930	930	100.00	813	87.42	24102	14240	59.08
JOURNAL OF BANKING & FINANCE	J BANK FINANC	2533	2533	100.00	1975	77.97	24856	10557	42.47
JOURNAL OF BEHAVIORAL FINANCE	J BEHAV FINANC	82	82	100.00	10	12.20	45	13	28.89
JOURNAL OF BUSINESS FINANCE & ACCOUNTING	J BUS FINAN ACCOUNT	381	381	100.00	233	61.15	1377	687	49.89
JOURNAL OF COMPARATIVE BUSINESS AND CAPITAL MARKET LAW	J COMP BUS CAP MARK	38	38	100.00	1	2.63	32	2	6.25
JOURNAL OF COMPUTATIONAL FINANCE	J COMPUT FINANC	40	40	100.00	3	7.50	22	6	27.27
JOURNAL OF CORPORATE FINANCE	J CORP FINANC	423	423	100.00	283	66.90	3534	1400	39.62
JOURNAL OF CORPORATE TAXATION	J CORP TAX	394	394	100.00	29	7.36	60	37	61.67

continued on next page ...

. . . Table 2.2 continued from previous page

Source journal name	Abbreviated form	#Articles	#Articles covered	%Articles covered	#Cited articles	%Cited articles	Cites received	#Cites covered	%Cites covered
JOURNAL OF CREDIT RISK	J CREDIT RISK	49	49	100.00	7	14.29	19	9	47.37
JOURNAL OF DERIVATIVES	J DERIV	91	91	100.00	25	27.47	129	50	38.76
JOURNAL OF ECONOMICS AND BUSINESS	J ECON BUS	412	412	100.00	140	33.98	1310	322	24.58
JOURNAL OF EMPIRICAL FINANCE	J EMPIR FINANC	221	221	100.00	79	35.75	387	144	37.21
JOURNAL OF FINANCE	J FINANC	2258	2258	100.00	2141	94.82	125645	64469	51.31
JOURNAL OF FINANCIAL AND QUANTITATIVE ANALYSIS	J FINANC QUANT ANAL	1159	1159	100.00	967	83.43	21211	11037	52.03
JOURNAL OF FINANCIAL ECONOMETRICS	J FINANC ECONOMET	88	88	100.00	32	36.36	277	67	24.19
JOURNAL OF FINANCIAL ECONOMICS	J FINANC ECON	1663	1663	100.00	1530	92.00	98052	52505	53.55
JOURNAL OF FINANCIAL INTERMEDIATION	J FINANC INTERMED	260	260	100.00	184	70.77	2938	1495	50.88
JOURNAL OF FINANCIAL MARKETS	J FINANC MARK	196	196	100.00	142	72.45	1603	999	62.32
JOURNAL OF FINANCIAL RESEARCH	J FINANC RES	326	326	100.00	209	64.11	1360	702	51.62
JOURNAL OF FINANCIAL SERVICES RESEARCH	J FINANC SERV RES	253	253	100.00	142	56.13	1347	549	40.76
JOURNAL OF FINANCIAL STABILITY	J FINANC STABIL	88	88	100.00	28	31.82	171	57	33.33
JOURNAL OF FUTURES MARKETS	J FUTURES MARKETS	1388	1388	100.00	981	70.68	8834	4560	51.62
JOURNAL OF INDUSTRIAL ECONOMICS	J IND ECON	896	896	100.00	419	46.76	16988	1072	6.31
JOURNAL OF INTERNATIONAL FINANCIAL MANAGEMENT & ACCOUNTING	J INT FIN MANAG ACC	38	38	100.00	5	13.16	11	6	54.55
JOURNAL OF INTERNATIONAL MONEY AND FINANCE	J INT MONEY FINANC	1299	1299	100.00	813	62.59	14044	4520	32.18

continued on next page . . .

... Table 2.2 continued from previous page

Source journal name	Abbreviated form	#Articles	#Articles covered	%Articles covered	#Cited articles	%Cited articles	Cites received	#Cites covered	%Cites covered
JOURNAL OF MONETARY ECONOMICS	J MONETARY ECON	1629	1629	100.00	1179	72.38	50514	10510	20.81
JOURNAL OF MONEY CREDIT AND BANKING	J MONEY CREDIT BANK	1523	1523	100.00	975	64.02	20455	5424	26.52
JOURNAL OF OPERATIONAL RISK	J OPER RISK	102	102	100.00	44	43.14	316	171	54.11
JOURNAL OF PENSION ECONOMICS & FINANCE	J PENSION ECON FINAN	85	85	100.00	20	23.53	74	28	37.84
JOURNAL OF PORTFOLIO MANAGEMENT	J PORTFOLIO MANAGE	1353	1351	99.85	596	44.05	5052	2327	46.06
JOURNAL OF REAL ESTATE FINANCE AND ECONOMICS	J REAL ESTATE FINANC	599	599	100.00	374	62.44	4451	1543	34.67
JOURNAL OF REAL ESTATE RESEARCH	J REAL ESTATE RES	119	119	100.00	58	48.74	254	131	51.57
JOURNAL OF REAL ESTATE TAXATION	J REAL ESTATE TAX	509	509	100.00	49	9.63	98	55	56.12
JOURNAL OF RISK	J RISK	63	63	100.00	10	15.87	40	14	35.00
JOURNAL OF RISK AND INSURANCE	J RISK INSUR	904	904	100.00	591	65.38	6113	2743	44.87
JOURNAL OF RISK AND UNCERTAINTY	J RISK UNCERTAINTY	535	535	100.00	344	64.30	10984	1514	13.78
JOURNAL OF RISK MODEL VALIDATION	J RISK MODEL VALIDAT	48	48	100.00	11	22.92	17	15	88.24
JOURNAL OF TAXATION	J TAX	2913	2692	92.41	382	13.11	830	634	76.39
JOURNAL OF THE AMERICAN REAL ESTATE AND URBAN ECONOMICS ASSOCIATION	J AM REAL ESTATE URB	76	76	100.00	66	86.84	843	445	52.79
LLOYDS BANK ANNUAL REVIEW	LLOYDS BANK ANNU REV	116	116	100.00	0	0.00	220	0	0.00
MANAGEMENT ACCOUNTING RESEARCH	MANAGE ACCOUNT RES	77	77	100.00	25	32.47	150	59	39.33
MANAGERIAL FINANCE	MANAGE FINANC	44	44	100.00	1	2.27	4	1	25.00

continued on next page ...

. . . Table 2.2 continued from previous page

Source journal name	Abbreviated form	#Articles	#Articles covered	%Articles covered	#Cited articles	%Cited articles	Cites received	#Cites covered	%Cites covered
MATHEMATICAL FINANCE	MATH FINANC	347	347	100.00	224	64.55	6017	1463	24.31
MSU BUSINESS TOPICS	MSU BUS TOP-MICH ST	43	43	100.00	2	4.65	141	2	1.42
NATIONAL TAX JOURNAL	NATL TAX J	1149	1149	100.00	644	56.05	9019	2046	22.69
NATIONAL WESTMINSTER BANK QUARTERLY REVIEW	NATL WESTM BANK Q R	256	251	98.05	0	0.00	335	0	0.00
NORTH AMERICAN JOURNAL OF ECONOMICS AND FINANCE	N AM J ECON FINANC	69	69	100.00	18	26.09	109	27	24.77
PACIFIC-BASIN FINANCE JOURNAL	PAC-BASIN FINANC J	90	90	100.00	21	23.33	91	28	30.77
PUBLIC FINANCE QUARTERLY	PUBLIC FINANC QUART	452	452	100.00	194	42.92	2018	426	21.11
PUBLIC FINANCE REVIEW	PUBLIC FINANC REV	145	145	100.00	31	21.38	409	41	10.02
PUBLIC FINANCE-FINANCES PUBLIQUES	PUBLIC FINANC	399	399	100.00	151	37.84	1263	321	25.42
QUANTITATIVE FINANCE	QUANT FINANC	597	597	100.00	179	29.98	2057	422	20.52
QUARTERLY REVIEW OF ECONOMICS AND BUSINESS	Q REV ECON BUS	350	350	100.00	98	28.00	1141	177	15.51
QUARTERLY REVIEW OF ECONOMICS AND FINANCE	Q REV ECON FINANC	186	186	100.00	52	27.96	630	114	18.10
REAL ESTATE ECONOMICS	REAL ESTATE ECON	405	405	100.00	292	72.10	3297	1491	45.22
REAL ESTATE TAXATION	REAL ESTATE TAX	9	9	100.00	0	0.00	0	0	0.00
REVIEW OF ACCOUNTING STUDIES	REV ACCOUNT STUD	129	129	100.00	90	69.77	929	581	62.54
REVIEW OF BUSINESS AND ECONOMIC RESEARCH	REV BUS ECON RES	195	195	100.00	20	10.26	171	29	16.96
REVIEW OF DERIVATIVES RESEARCH	REV DERIV RES	43	43	100.00	5	11.63	18	7	38.89
REVIEW OF FINANCE	REV FINANC	89	89	100.00	40	44.94	299	151	50.50

continued on next page . . .

... Table 2.2 continued from previous page

Source journal name	Abbreviated form	#Articles	#Articles covered	%Articles covered	#Cited articles	%Cited articles	Cites received	#Cites covered	%Cites covered
REVIEW OF FINANCIAL STUDIES	REV FINANC STUD	1068	1068	100.00	922	86.33	33108	17221	52.01
REVISTA ESPANOLA DE FINANCIACION Y CONTABILIDAD-SPANISH JOURNAL OF FINANCE AND ACCOUNTING	REV ESP FINANC CONTA	93	93	100.00	5	5.38	10	5	50.00
SCHWEIZERISCHE ZEITSCHRIFT FUR SOZIALVERSICHERUNG	SCHWEIZ Z SOZIALVERS	44	44	100.00	0	0.00	0	0	0.00
TAXES	TAXES	1055	1055	100.00	110	10.43	302	140	46.36
THREE BANKS REVIEW	THREE BANKS REV	59	59	100.00	0	0.00	2	0	0.00
WORLD BANK ECONOMIC REVIEW	WORLD BANK ECON REV	477	477	100.00	231	48.43	10124	849	8.39
WORLD ECONOMY	WORLD ECON	1378	1378	100.00	414	30.04	7013	780	11.12

2.2.1 (d) Mine

The fourth and last phase in our methodology consists of the actual data mining and mathematical modelling. This process is quite elaborate and therefore deserves an entire section of its own. We shall cover the details in Section 2.3. Before proceeding to that, we need to note a few caveats on the extraction of citation linkages from the cited reference field (C1).

2.2.2 Extracting citations

Citations to each article (valid at the download date) are specified by the TC, or times cited field. For a given cited article C, the value of TC (some integer ≥ 0) is incremented whenever references are made by other ISI-indexed documents[5], for example, $R_1 \rightarrow C, R_2 \rightarrow C, \ldots, R_{TC} \rightarrow C$. Such correspondences can be represented as citation linkages on what we shall refer to as the document citation network (DCN).

Definition 16 (Document citation network/graph)**.** *This is the network (graph) of scientific documents (articles, letters, reviews, books, etc) in which each node is a distinct document, and directed links point from the referencing (citing) document to the referenced (cited) node. Directed trees (or chains) on this graph represent citation flows (i.e. the intellectual lineage) between any two nodes.*

Since ISI data specifies cited references made by each indexed document under the CR field, *some* fraction of the corresponding TC count can be retrieved as in-links on the document citation network. We say *some* because a complete reconstruction of TC count from CR fields necessarily depends on the completeness of the set of cited and citing documents. By sampling only cited references that correspond to documents of the type "Article", we can expect to extract the number of article citations $k \leq TC$. Furthermore, the extent of article-citation-extraction in the CR field will also be dependent on the source

[5]Including non-article document types.

journals and time period covered. According to the source data, the citation and in-link distribution is as follows:

Table 2.3: Citation and in-degree statistics.

11	articles with more than	1000	citations	1	article with more than	1000	in-links
50	articles with more than	500	citations	5	articles with more than	500	in-links
208	articles with more than	250	citations	42	articles with more than	250	in-links
1019	articles with more than	100	citations	309	articles with more than	100	in-links
2718	articles with more than	50	citations	972	articles with more than	50	in-links
11978	articles with more than	10	citations	5776	articles with more than	10	in-links
43144	articles with fewer than	5	citations	25539	articles with fewer than	5	in-links
6276	articles with exactly	1	citation	6964	articles with exactly	1	in-link
28151	articles with exactly	0	citations	10634	articles with exactly	0	in-links

There are three main factors that can be attributed to this glaring disparity between citation and in-degree distribution. These are: (i) time coverage, (ii) journal coverage, and (iii) errors in extracting citations from cited reference data. In terms of time coverage, the choice of study period excludes all articles and references published before 1980 and after 2011. On the one hand, references made to ISI articles prior to 1980 must be ignored since we do not possess publication data for those articles. This artificial cutoff inflates the importance of articles published in 1980 (with respect to those citing it in subsequent years) by making it seem as if those articles do not depend on any prior works. On the other hand, citations made from articles published in 2012 will necessarily be missed, which may add to the disparity between the number of citations reported by ISI at the download date with those traced from cited references available between 1980 to 2011.

Citation counts recorded by ISI are restricted to citations made by indexed publications from ISI-indexed journals or conference proceedings. Accordingly, we should be able to perfectly match up the TC count of a publication with its corresponding in-degree centrality on the document citation network, provided that we have complete time, journal, and conference proceedings coverage for the entire *Web of Knowledge* database. Since complete database access is not readily available, we should expect some difficulty in matching up the exact number of in-links with the TC count for any given ISI-indexed publication. Specific to our purposes, journals *outside* that cite those *inside* of the JCR

subject category of "BUSINESS, FINANCE" are necessarily left out, and thus contribute to gaps in the in-link structure[6].

Citation extraction may itself pose similar issues. Errors in processing or retrieving citations from the C1 field contribute towards adding false positive in-links or the incorrect omission of in-links (false negative links). This may either be due to mistakes in the extraction algorithm (software code) or due to erroneous entries in the cited reference field. The former is within our control, and thus it is important to avoid any oversight associated purely with the extraction process itself. The latter on the other hand, is a more difficult issue since the presence of a single character error in either the author, publication year, source journal abbreviation, volume number, or beginning page number, can render the resulting ID string unusable.

For example, a single character difference in the first author string such as "FAMA EF, 1993, J FINANC ECON, V33, P3", is treated as distinct from say, "FAMA E, 1993, J FINANC ECON, V33, P3", unless measures are taken to anticipate and account for such occurrences. Essentially, this is a problem of *author disambiguation*: how to correctly identify the same individual with different names (synonyms), as well as distinguish different individuals with the same name (homonyms). This is an important open problem[7] such that the KDD Cup 2013 competition[8], focused on the "Author Disambiguation Challenge"[9]. This competition was jointly organised by Microsoft Research in an effort to augment its Academic Search platform.

Accordingly, any error in an extracted reference from the C1 field (not just in the first

[6]These may typically include journals in the JCR subject category of "BUSINESS", "ECONOMICS", and "MANAGEMENT".

[7]Since the accuracy of any resulting analyses depends on the correctness of the data handling (parsing). The resolution of such a problem requires the use of machine learning algorithms such as that employed by Li et al. (2013).

[8]Under the Association for Computing Machinery's Special Interest Group on Knowledge Discovery and Data Mining, or ACM SIGKDD.

[9]Specifically, on synonym disambiguation.

author name) creates a disambiguation problem, specifically, how to correctly associate an erroneous reference with the correct accession number. Since the focus of this thesis is to prototype a network analysis method for ranking researchers, for the sake of simplicity, we assume that such errors in the study dataset are negligible.

2.3 Network analysis

In this section, we discuss the construction and mining of networks from bibliometric data. The basic idea is to score (rate) nodes based on their location within the resulting link structure, that is, we wish to determine the extent at which a node is either peripheral or central to a given network. While this depends on the structure of the network itself (and its underlying bibliometric data), it also depends on how we wish to look at it. A scoring algorithm can be designed to pick out a desired trait by assigning higher scores to nodes that exhibit prominence (i.e. structurally stand out) in that trait. These scores can then used as a basis to rank each corresponding entity represented by the node set.

2.3.1 Document citation network (DCN)

Description and construction—A document citation network $D = (V, E)$ consists of n nodes $V = \{v_1, \ldots, v_n\}$ representing research papers (journal articles and/or conference papers) with m directed links $E = \{e_1, \ldots, e_m\}$ between nodes representing citation linkages between papers. If a paper p_2 cites paper p_1 in its bibliography a corresponding link $v_2 \rightarrow v_1$ is added to D to reflect that association. Each link should respect strict time ordering, that is, an older paper should not cite a newer paper. For each paper in the study dataset, we assign a directed link from each citing paper to each target cited paper listed within its cited reference (hence, D is a directed graph).

Matrix representation—The connectivity of a document citation network (DCN) can be expressed by the binary adjacency matrix A, of which, each element describes the

presence or absence of a citation flow from paper $i \to j$:

$$A = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nn} \end{pmatrix} \tag{2.1}$$

Each citing paper points to a specific source citing paper exactly once, hence, a citation flows from i to j (for all $i \neq j$) if and only if $A_{ij} = 1$, otherwise $A_{ij} = 0$. The citation network is simple by design – papers cannot reference themselves, hence, D contains no loops (this corresponds to setting $A_{ii} = 0$). Additionally, there are no multiple links connecting the source and target nodes since each paper can make references to other documents exactly once. Since time ordering must be preserved, an older paper j cannot cite a newer paper i unless appearing roughly around the same time.

Discriminating nodes—The key idea here is that: (i) the connectivity generally varies from one node to the next (on real world networks), and (ii) some nodes are more central to the network than others. Centrality scores provide some means to quantitatively discriminate between nodes although the resulting hierarchy (ranking) changes according to the underlying emphasis of what kind of centrality is being measured. For example, the number of citations received by a publication corresponds to the in-degree centrality of its representative node on the DCN. An assortment of measures can be designed and implemented depending on desired features we wish to highlight. Some examples are described in Appendix A.

Since we wish to take into account the influence of citing papers and not just their number, a PageRank centrality approach is best suited for this purpose (see Equation (1.7)). Since older papers that are still cited years after its initial publication signals some con-

tinuing importance, the citation age – that is, the time elapsed between the citing paper and the cited paper – must be taken into account somehow (Redner, 2005). We describe such a scheme in the following paragraph.

Assigning link weights—Recall that a simple citation count unrealistically treats all citation sources as equal, regardless of the citation age. Clearly, this oversimplifies the situation and therefore we need a method that resolves the quality of papers based on *how* they are cited. Following Yang et al. (2011), a temporal score can be assigned to each citation link in which it is assumed that the importance of a cited paper is greater the larger the time gap between its publication year and that of its citing paper. Given that paper p_i published in year y_i, cites paper p_j published in an earlier year y_j such that $y_i - y_j \geq 0$, the citation influence ratio (CIR) of paper p_j on p_i can be expressed as:

$$CIR(p_{ij}) = \beta_1(1 - \beta_2^{y_i - y_j}) \tag{2.2}$$

where β_1 is a scaling parameter and β_2 ($0 < \beta_2 < 1$) is the decay base.

Setting $\beta_1 = 1$ and $\beta_2 = 0$ reduces Equation (2.2) to a full citation count from paper j to i. Based on experiments conducted by Yang and co-workers on ACM SIG publications, it was found that best performance is obtained when $\beta_1 = 1$ and $\beta_2 = 0.9$. Setting $\beta_1 = 1$ is appropriate since we wish to discount the citation count, and not to increase it. Setting $\beta_2 \to 1$ is necessary to discount the citation exponentially according to the time elapsed.

Scoring individual papers—We now have enough information to compute an influence score for each paper based on its prominence on the DCN. For this purpose, we use the weighted PageRank approach as formulated below:

$$PR(i) = \alpha \sum_{j \to i} PR(j) \cdot P(j,i) + \frac{(1-\alpha)}{N} \tag{2.3}$$

This equation is similar in formulation as Equation (1.7), with the exception of the propagation factor $P(j,i)$ on the right hand side which serves to bias the random walk on links with a high $P(j,i)$ value. Specific to the DCN we constructed, we shall equate the propagation factor with the citation influence ratio of paper j on i. To wit:

$$P(i,j) = CIR(p_{ij}) = \beta_1(1-\beta_2^{y_i-y_j}) \tag{2.4}$$

Hence, the influence score of paper (node) i is:

$$\begin{aligned} PR(i) &= \alpha \sum_{j \to i} PR(j) \cdot CIR(p_{ji}) + \frac{(1-\alpha)}{N} && (2.5) \\ &= \alpha \sum_{j \to i} PR(j) \cdot \beta_1(1-\beta_2^{y_j-y_i}) + \frac{(1-\alpha)}{N} && (2.6) \end{aligned}$$

Note that according to the notation above, node j cites (in-links with) node i. Following Chen et al. (2007), we set the parameter $\alpha = (1-d) = 0.5$ to model a random researcher sequentially following $k = 1/(1-d) = 2$ citation chains on average before jumping to a new paper (node) on the DCN.

Next step: Scoring individual researchers.—In Equation (2.2), we have introduced a bias to each cited paper that assigns greater importance the larger the citation age (elapsed time) with its citing counterpart. The next logical step is to use this information to score individual researchers. Intuitively, the influence of a researcher should depend on the quality of their work, as well as the influence of researchers citing that work. We describe such a scheme in Section 2.3.2 and Section 2.3.3.

2.3.2 Author citation network (ACN)

Description and construction—Author citation linkages can be determined by cross-referencing links on the DCN with the associated author data. Suppose that a citing

paper X has a set of authors $A = \{a_1, a_2\}$ and its cited paper Y has a set of authors $B = \{b_1, b_2, b_3\}$. It follows that author citation linkages derived from the document citation link $X \rightarrow Y$ are all ordered pairings connecting set A to B, that is, each element in set A is assigned a directed link to all elements in set B so that:

$$\{a_1 \rightarrow b_1, a_1 \rightarrow b_2, a_1 \rightarrow b_3, a_2 \rightarrow b_1, a_2 \rightarrow b_2, a_2 \rightarrow b_3\}$$

Such linkages can be encoded as an author citation network $G = (V', E')$ consisting of N nodes $V' = \{v'_1, \ldots, v'_N\}$ representing distinct author keywords and M directed links $E' = \{e'_1, \ldots, e'_M\}$ between nodes representing citation linkages between authors. If an author a_2 cites author a_1 a corresponding link $v'_2 \rightarrow v'_1$ is added to G to reflect that association. We further require that author self-cite links are suppressed (ignored) in order to maintain a simple (loopless) directed graph.

Assigning link weights—Accordingly, the citation influence (CI) from researcher a_j (in paper j) to researcher a_i (in paper i) can be defined as a function of its link weight on a_{ij}. Yang et al. (2011) proposes quantifying the citation influence between authors based on the citation influence ratio between their respective papers as previously defined in Equation (2.2):

$$CI(a_{ij}) = \sum_{p_{ij}:a_i \rightarrow a_j} CIR(p_{ij}) \tag{2.7}$$

where p_j is any paper authored by researcher a_j citing some paper p_i, which in turn, is authored by researcher a_i (the direction of the arrow in $a_i \rightarrow a_j$ indicates that a_i cites, or depends on, a_j).

Interpretation—Equation (2.7) asserts that higher influence is assigned to a cited author a_i the larger the number of citing items p_j (from a_j) and the larger the time gap for each individual citing item. The former assertion is proportional to a citation count. The fractional nature of the CIR of each paper contributing to the overall CI score reflects

how we choose to value researchers that author papers with a typically long citation age.

Scoring individual researchers: Coarse-Grain scheme—Having defined link weights on the ACN, we can now use the same weighted PageRank algorithm defined in Equation (2.3) to score each node. We shall refer to this scheme as the Coarse-Grain (CG) link weighting scheme, since it is based purely on citation data (does not require publication data to compute prominence scores). According to this scheme, we formulate the prominence score of each node as:

$$\begin{aligned} PR^{(C)}(i) &= \alpha \sum_{j \to i} PR^{(C)}(j) \cdot P^{(C)}(j,i) + \frac{(1-\alpha)}{N} && (2.8) \\ &= \alpha \sum_{j \to i} PR^{(C)}(j) \cdot CI(a_{ji}) + \frac{(1-\alpha)}{N} && (2.9) \\ &= \alpha \sum_{j \to i} PR^{(C)}(j) \cdot \left(\sum_{p_{ji}: a_j \to a_i} CIR(p_{ji}) \right) + \frac{(1-\alpha)}{N} && (2.10) \end{aligned}$$

in which we have set the propagation factor as:

$$P^{(C)}(i,j) = CI(a_{ij}) = \sum_{p_{ij}: a_i \to a_j} CIR(p_{ij}) \tag{2.11}$$

The superscript C is used to denote the CG scheme.

Similar to the DCN formulation, $\alpha = (1-d) = 0.5$, so that a random researcher searches a neighbourhood within $k = 1/(1-d) = 2$ degrees of separation on average before seeking information elsewhere (although it is fundamentally a social network, we assume that the ACN is another kind of informational network due to the information diffusion component of the PageRank algorithm).

Important remarks—Note that the DCN PageRank scores are not required in the CG PageRank calculations, only the citation influence ratio of papers, CIR. Furthermore,

this CG scheme is purely based on citation data alone. The publication traits of individual researchers – specifically, their publishing history and tendencies – are not taken into account. For this, we need to consider the Yang-Yin-Davison (YYD) link weighting scheme as described in the following subsection.

2.3.3 Yang-Yin-Davison link weighting scheme

Targeted features—In contrast to the Coarse-Grain scheme, the Yang-Yin-Davison scheme (YYD) attempts to score researchers not only through the impact of their work but also on how they publish their work. The reasoning for this is that the CG scheme thus far only computes "raw" influence. Hence, high scoring researchers obtained from computing the CG scheme are influential researchers, regardless of their publication traits. It can be argued however that authorities and experts exhibit additional traits in their publication profile, specifically that they are typically: (i) long-established in their field; (ii) are highly continuant in their work; and (iii) are still active in the present (Yang et al., 2011).

Individual Temporal Importance—Given these considerations, a high relative impact score should be assigned to researchers who have continuant and numerous long-standing contributions that are cited by far more recent works (by other researchers) in the literature. Yang and coworkers thus constructed a temporal score based on three aspects of a researcher's academic activity. This score is termed the individual temporal importance (ITI) and is expressed as:

$$ITI_i = CareerTime_i \times \left(\frac{1}{LastRestTime}\right) \times \left(\frac{1}{PubInterval}\right) \tag{2.12}$$

where, relative to researcher (author) a_i, $CareerTime$ is the number of years spanning the first and last publication, $LastRestTime$ is the number of years since the last publi-

cation (relative to the present), and $PubInterval$ is the average number of years between two consecutive publications.

Assigning link weights—In the YYD scheme, each citation link $a_i \rightarrow a_j$ on the author citation network (ACN) must combine citation and publication information. However, it is also possible that a_j and a_i have jointly co-authored papers together, and if such is the case, Yang and co-workers reasoned their proximity should be reflected in their corresponding link weight. To address this, they proposed using the following link weight:

$$w(a_{ij}) = (NumCo(a_{ij}) + CI(a_{ij})) \times ITI_j \tag{2.13}$$

where $NumCo(a_{ij})$ is the number of co-authored papers shared between a_i and a_j. CI is the citation influence defined in Equation (2.7).

Removal of coauthor term—There is a dimensional problem in Equation (2.13), specifically, two quantities of incompatible units are being added together: coauthor count ($NumCo(a_{ij})$) with a fractional citation count $CI(a_{ij})$. For this reason, we remove the coauthor term and define the modified YYD link weight as:

$$w^*(a_{ij}) = CI(a_{ij}) \times ITI_j \tag{2.14}$$

Since ITI has units of $[time]^{-1}$ (see Equation (2.12)), the modified YYD link weight dimensions of *citation count over time*.

Scoring individual researchers: YYD scheme—The next step is to propagate the "temporal authority" from some citing author a_i to some cited author a_j. The temporal authority should capture temporal characteristics of a researcher's publication and citation profile. This is the key difference that distinguishes the YYD scheme from the CG discussed in Section 2.3.2. The link weights are normalised over the entire network by

defining the propagation probability from a_i to a_j using:

$$P^{(Y)}(i,j) = \frac{w^*(a_{ij})}{\sum_{k:i\to k} w^*(a_{ik})} \tag{2.15}$$

As summarized in Yang et al. (2011): "author a_i will propagate more authority to author a_j [...] if a_i has greater citation influence on a_j, or if a_j has greater individual temporal importance."

$$\begin{aligned} PR^{(Y)}(i) &= \alpha \sum_{j\to i} PR^{(Y)}(j) \cdot P^{(Y)}(j,i) + \frac{(1-\alpha)}{N} && (2.16) \\ &= \alpha \sum_{j\to i} PR^{(Y)}(j) \cdot \left(\frac{w(a_{ji})}{\sum_{k:k\to i} w(a_{ki})} \right) + \frac{(1-\alpha)}{N} && (2.17) \end{aligned}$$

Similar to the CG scheme, $\alpha = (1-d) = 0.5$, so that a random researcher searches a neighbourhood within $k = 1/(1-d) = 2$ degrees of separation on average before seeking information elsewhere (the YYD scheme/algorithm is run on the same author citation network).

2.3.4 Goodness of prediction

In terms of evaluating the accuracy of scores and rankings produced, this can be done when *ground truth* is available, which in this case it is not. We do not have an absolute reference point to say with 100% certainty that one researcher has more impact than another[10]. What we are able to do is generate quantitative judgements based on certain assumptions. This is much like how PageRank does not actually rank – in an ontological sense – the best to worst webpages, it can only provide an epistemological model of what may be the case based on certain justifications and beliefs (explicit or implicit). Following

[10]We cannot tell whether the positioning of a researcher is off by its "actual" value, since we do not have, and quite conceivably, cannot attain this data. The best outcome from doing this work is to corroborate the presence or absence of some effect or an alternative listing. This evidence must be interpreted and put into context with other relevant information. If it is consistent with some auxiliary data, we are perhaps making some progress.

this reasoning, evaluation of the *goodness-of-prediction* will not be covered in this thesis. We shall instead look into ways of integrating rankings produced by different methods to infer additional insight. While we cannot be certain how spot on such inferences are, we shall demonstrate how these can be used to help us obtain a better grasp of the data.

2.4 Outline of Methodology

To clarify on what was discussed in this chapter, we provide several schematics to outline the methodology used in this thesis. We note three key assumptions underlying the work presented. First, there is enough useful information recorded in publication and citation data to make quantitative comparisons between researchers as shown in Figure 2.3. Second, researchers responsibly and comprehensively cite intellectual influences leading to the work they publish (MacRoberts & MacRoberts, 2010, 1996; Greenberg, 2009); see Figure 2.4. And third, a sufficient amount of latent information can be extracted from citation links to infer influence flows between researchers; see Figure 2.5. The outline of the Coarse-Grain (CG) and Yang-Yin-Davison (YYD) link weighting schemes are as depicted in Figure 2.6 and Figure 2.7, respectively.

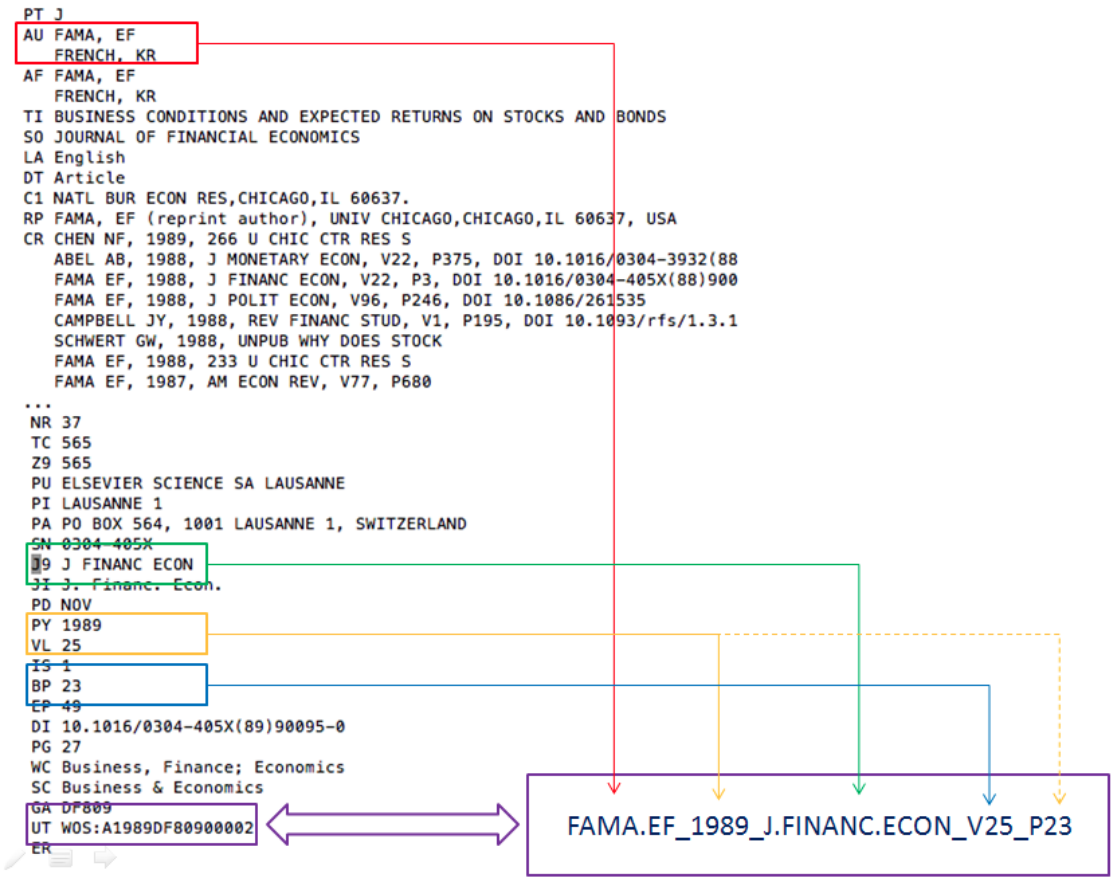

Figure 2.3: Parsing ISI data.

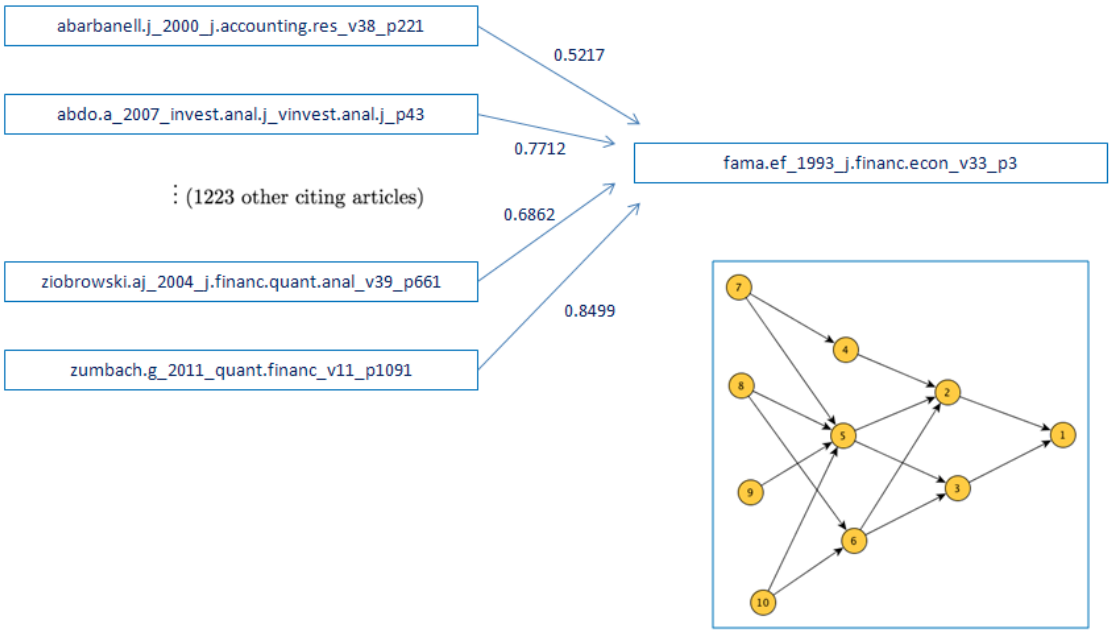

Figure 2.4: Snapshot of document citation network (DCN) centred on one paper, i.e. "fama.ef_1993_j.financ.econ_v33_p3". Numerical values on links corresponds to CIR values. Inset: illustration of hierarchical structure due to time ordering of papers on the DCN.

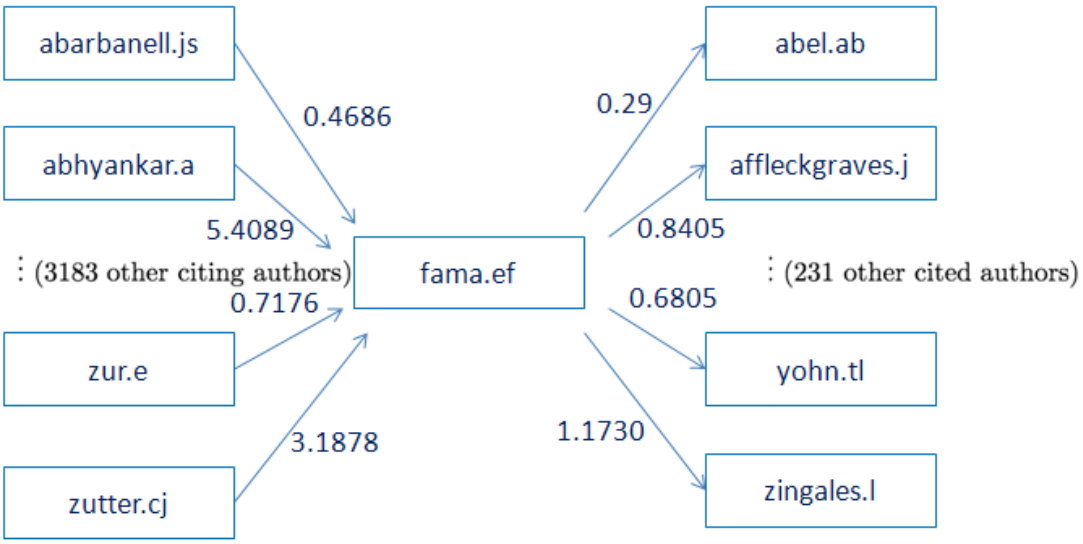

Figure 2.5: Snapshot of author citation network (ACN) centred on one author, i.e. "fama.ef". Numerical values on links corresponds to CI values.

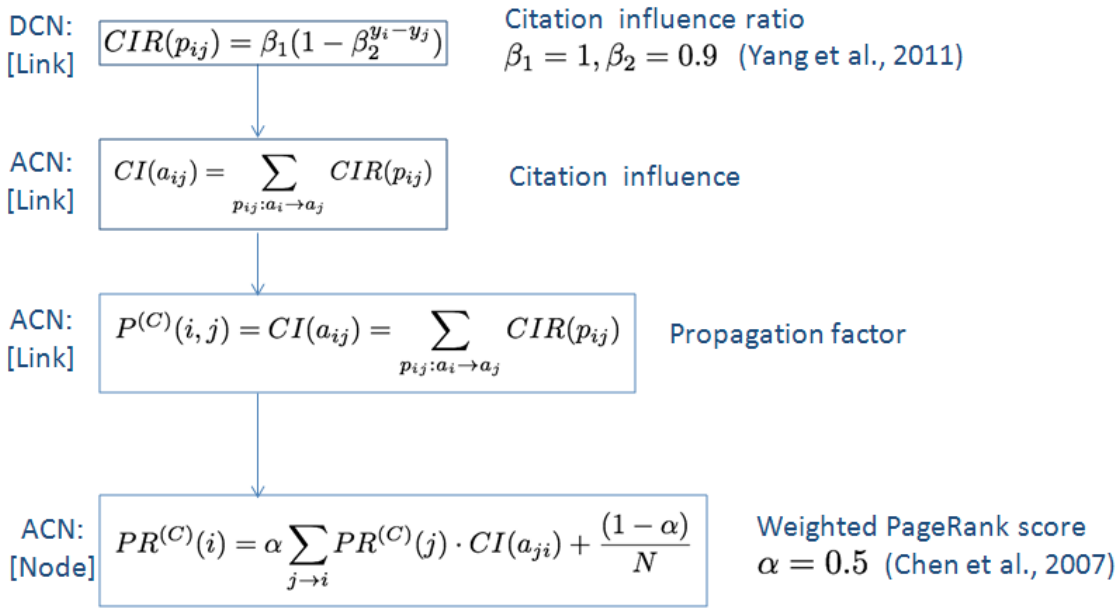

Figure 2.6: Outline of Coarse-Grain (CG) scheme.

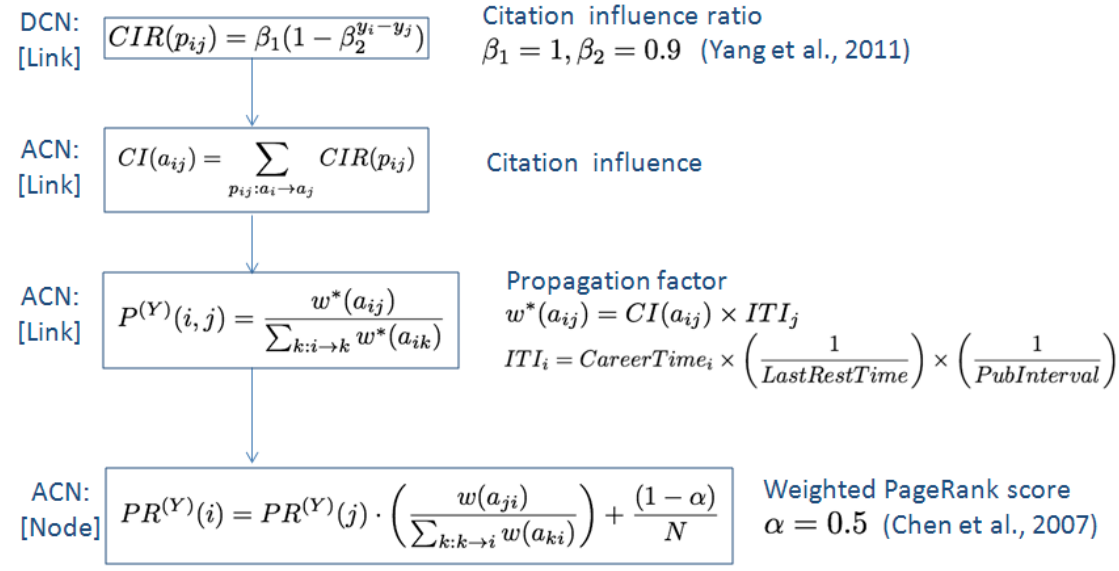

Figure 2.7: Outline of YYD scheme.

2.5 Software used

Calculations were carried out using a mixture of custom-made Perl, Bash, R, and Python scripts:

- Perl: This programming language was used to script general data processing tasks. See Section 2.2.1.

- R: The network analysis package *igraph* (Csardi & Nepusz, 2006) was used to handle large-scale network data. The PerformanceAnalytics package (Carl et al., 2009) was used to produce scatter plots with statistical correlation values.

- Python: The network analysis package NetworkX (Hagberg, Schult, & Swart, 2008) was used to compute weighted PageRank scores.

- Bash: This Unix shell was used to shift data inputs and outputs between Perl, R, and Python.

- Gephi: This open source graph visualisation software tool was used to visualise the networks studied (Bastian et al., 2009). Furthermore, Gephi has an efficient implementation of the community detection algorithm proposed by Blondel et al. (2008). This greatly assisted in improving the layout of graphs by grouping and colouring nodes that belong to the same cluster.

CHAPTER 3

ANALYSIS

In this chapter we will show the results of our analysis according to the methodology described in Chapter 2. This chapter is split into three sections. The first and second section covers the identification of important papers on the document citation network (DCN) and experts/authorities on the author citation network (ACN), respectively. The third section covers how we can modify the Yang-Yin-Davison method to identify rising stars on the author citation network.

3.1 Document citation network

The constructed document citation network (DCN) has properties as shown in Table 3.1. Not all nodes belong to the same connected component, and therefore we choose the giant weakly connected component (GWCC) to compute citation influence ratio (CIR) scores for links and weighted PageRank scores for nodes (see Equation (2.2) and Equation (1.7), respectively). Since the GWCC consists of roughly 95.9% of all nodes, as well as, 99.6% of all links on the DCN, we expect that the effect of omitting all other components is negligible. A plot of the giant component of the DCN is shown in Figure 3.1.

Visually, we can see that the DCN exhibits some community structure, whereby citations within the same community are more intense than between disparate communities. This is likely due the clustering of papers and their references over time to maintain existing paradigms or through the formation of new topics and research areas. It is also quite possible that the community structure is a manifestation of clustering behaviour between researchers to form invisible colleges (Crane & Kaplan, 1973; Zuccala, 2005).

Table 3.1: Properties of document citation network (DCN).

Nodes	36,043
Links	265,058
No. of connected components	543
Size of giant weakly connected component	34,590
Links on giant weakly connected component	264,110
Giant Weakly Connected Component (GWCC)	
Density	2.2×10^{-4}
Average path length	5.0
Diameter	17
Transitivity	0.074
Mean degree	7.6
Median degree	2.0
Maximum degree	1227
Modularity	0.644
No. of resolved communities	22

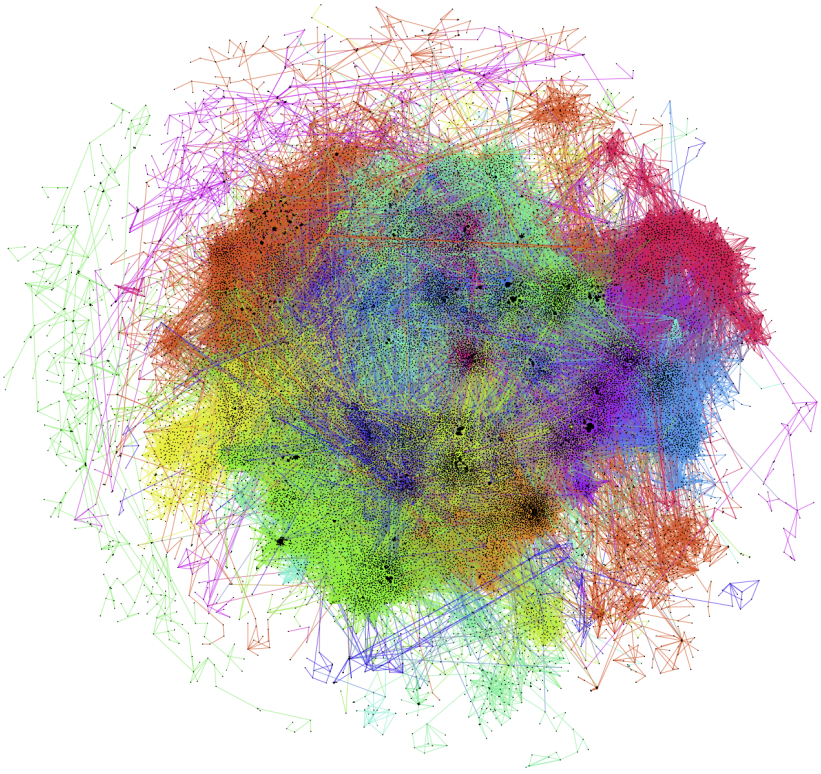

Figure 3.1: Giant weakly connected component of document citation network (DCN). Nodes are color-coded via community detection method of Blondel et al. (2008) and plotted using an open source graph visualisation and exploration tool called Gephi (Bastian et al., 2009).

We list the top 20 cited articles by in-link count in Table 3.2 and the top 20 Google PageRank articles based on a weighted PageRank with CIR link weights in Table 3.3. CIR weights are computed using Equation (2.2). For clarity, we denote the rankings associated with the column "Cite Rank" as $CiteRank$, and "Google Rank" as $GoogleRank$. We have two options to assess the observed rank permutation on both tables: (i) we can either look for agreement in the form of shared top-ranked items, or (ii) we can look for stark disagreements between both rankings.

With respect to option (i), we can find some items with $|CiteRank - GoogleRank| < 10$. Such instances can be used to corroborate the ability of either method (citation count or PageRank) to pick out important papers. A good agreement between both methods is obtained when a paper has high Google# precisely because it possess many in-links, thus inflating the summand in Equation (1.7) and Equation (2.3). What's far more interesting is option (ii) since this signifies cases of stark disagreements between $CiteRank$ and $GoogleRank$. In particular:

- Items with $GoogleRank/CiteRank > 10$ correspond to papers that are highly cited but have low prominence score, hence, such items signify potentially overvalued papers.

- Items with $CiteRank/GoogleRank > 10$ correspond to papers that are not highly cited but have high prominence score. Such instances signify undervalued papers in general, and in the case of highly ranked papers by $GoogleRank$, scientific gems (Chen et al., 2007).

Overvalued papers—According to Table 3.2, papers corresponding to $CiteRank$ #4, #5, #7, and #11 appear overvalued, though only around a factor of 10. Aside from being ranked in the top 20 by PageRank, what these papers have in common is that each have accrued more than 300 citations and they are (co)authored by known prize winners

in the field (see Table 3.10). These prize winners include Sheridan Titman (University of Texas at Austin), Eugene F. Fama (The University of Chicago Booth School of Business), and Kenneth R. French (Tuck School of Business at Darthmouth). The exception to this is Mark M. Carhart who has published 3 ISI-indexed articles, in which the last two were published in 2002. The inter-connectivity of top 20 cited papers on the DCN is as displayed in Figure 3.2.

The selection criteria we used to detect overvalued papers, specifically that:

$$GoogleRank/CiteRank > 10$$

is an ad hoc choice that may yield false positives in the sense that some papers may actually be important, just not as important as some others in terms of influencing influential work. Clearly, for any positive hits obtained from the above criteria, it is only plausible to say that such papers have *overvalued citation counts relative to their PageRank score*.

There is also the question of near hits (or misses) like the paper at #20 by $CiteRank$ in Table 3.2 yields $171/20 = 8.55$, a value which could be considered close enough to raise a flag according to our selection criteria. To be fair, "raising flags" are all that can be done when applying such heuristics. That is to say, any selection criteria designed on an ad hoc basis (i.e. without empirical support) should only be used to assist in detecting papers with anomalous (suspicious) features[1]. Ultimately, verification must be done manually and by employing all relevant information.

Undervalued papers—According to Table 3.3, papers corresponding to $GoogleRank$

[1]Especially in the case of unravelling highly cited works that have secured unfounded authority via citation distortions (Greenberg, 2009). Greenberg describes citation distortions as follows: "Primary data that weakened or refuted claims on which the belief was based were ignored (citation bias) and a small number of influential papers and citations exponentially amplified supportive claim over time without presenting new primary data (amplification). Certain related claims were invented as fact. The combined effects of these citation distortions resulted in authority of the belief (acceptance of it) according to social network theory."

#4, #6, #12 #15, and #18 have $CiteRank/GoogleRank > 10$. The common factor to these five papers is the that they are all published in the early 1980s and have citation counts (in-link count) in the range of 91 to 310. This points out to the citation age bias built into Equation (2.2); CIR is large the greater the gap between citing article and cited article, which is incidentally easier to achieve for older articles in the dataset. When combined with the artificial time coverage cut-off in the study dataset (exclusion of articles published before 1980), we can expect the earliest papers which are cited by many influential papers to obtain higher PageRank scores. We find that this is indeed the case since most papers in the top 20 list by PageRank score are from the 1980s, 3 in the 1990s, and none from the 2000s.

Despite this, we do find that not all items in the 80s are stuck to the early years (about four papers appear in the mid-80s). More importantly, the top 20 list exhibits strong rank permutation relative to $CiteRank$. The question remains, are papers marked by asterisks in Table 3.3 *scientific gems*? This is a loaded question since the selection criteria chooses exactly those high-ranked papers by PageRank algorithm that have a moderate or low citation count. This means that the evidence is not in the numbers (the number of citations) but rather in the quality of the articles that cite it[2]. For example, the paper ranked at #4 (reinganum.mr_1981) by PageRank is cited by the paper at rank #17 (debondt.wfm_1985), which is itself a highly prominent paper. We show a snapshot of the citation network for top 20 papers listed in Figure 3.3 below.

[2]Recall that the PageRank algorithm defines a node as prominent if it is cited by prominent nodes. See Equation (1.7).

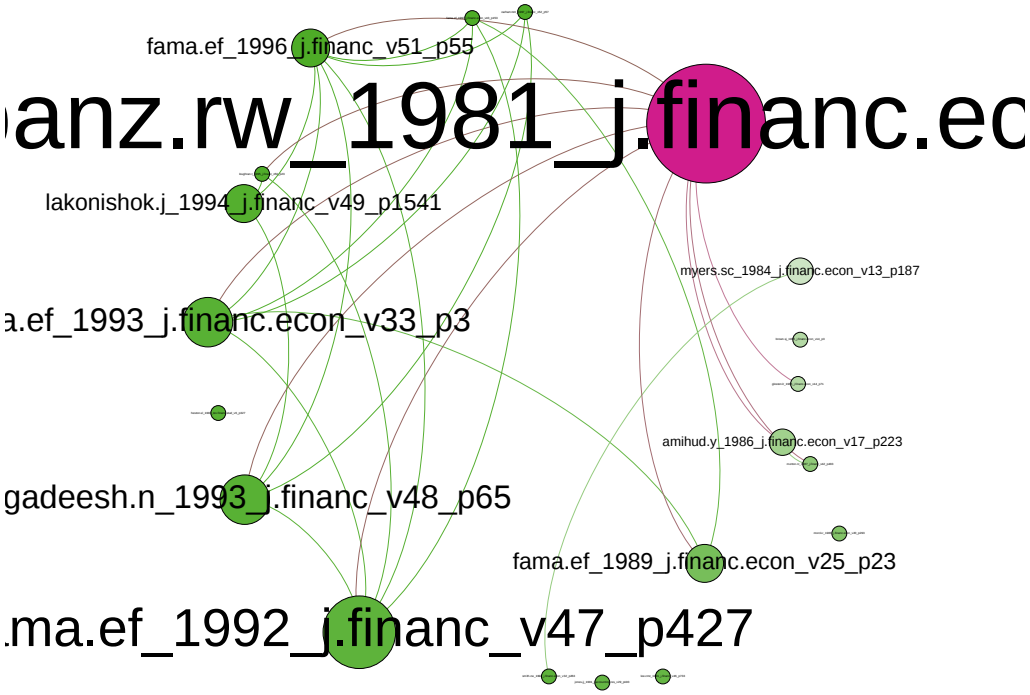

Figure 3.2: Document citation network (DCN) for nodes in the top 20 list by citation count (in-degree centrality). Nodes are color-coded by year and sized by citation count on the entire DCN. Plotted with Gephi.

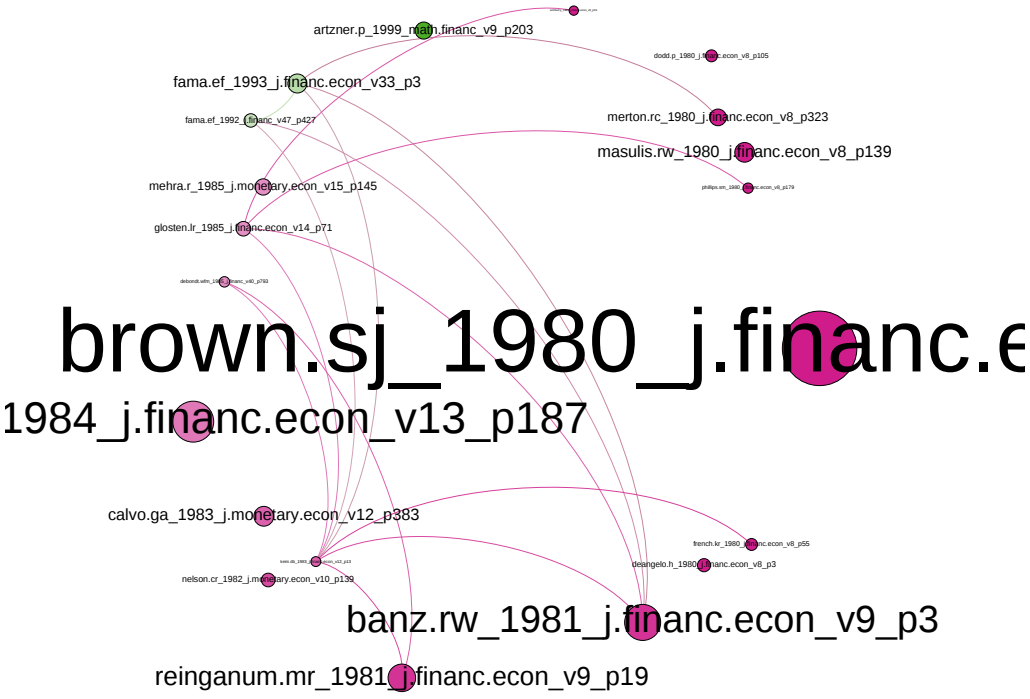

Figure 3.3: Document citation network (DCN) for nodes in the top 20 list by PageRank. Nodes are color-coded by year and sized by PageRank score on the entire DCN. Plotted with Gephi.

Table 3.2: The top 20 cited articles. JAR, JF, JFE, and RFS denote the journals *Journal of Accounting Research*, *The Journal of Finance*, *Journal of Financial Economics*, and *Review of Financial Studies*, respectively. The asterisk (*) denotes articles with $PageRank$-to-$CiteRank$ ratio larger than 10.

Cite Rank	Cites	Google Rank	Google # ($\times 10^{-3}$)	Publication SO	VL	BP	PY	Title	Author(s)
1	1227	7	1.28	JFE	33	3	1993	Common risk-factors in the ret...	E.F. Fama and K.R. French
2	878	13	1.10	JF	47	427	1992	The cross-section of expected ...	E.F. Fama and K.R. French
3	857	2	1.96	JFE	13	187	1984	Corporate financing and invest...	N.S. Majluf and S.C. Myers
4*	614	59	0.53	JF	52	57	1997	On persistence in mutual fund ...	M.M. Carhart
5*	505	61	0.52	JF	48	65	1993	Returns to buying winners and ...	N. Jegadeesh and S. Titman
6	477	11	1.14	JFE	14	71	1985	Bid, ask and transaction price...	L.R. Glosten and P.R. Milgrom
7*	452	127	0.32	JFE	43	153	1997	Industry costs of equity	E.F. Fama and K.R. French
8	426	40	0.64	JFE	20	293	1988	Management ownership and marke...	R. Morck, A. Shleifer and R.W. Vishny
9	406	32	0.73	JFE	14	3	1985	Using daily stock returns - th...	S.J. Brown and J.B. Warner
10	400	34	0.69	JFE	17	223	1986	Asset pricing and the bid ask ...	Y. Amihud and H. Mendelson
11*	390	133	0.31	JF	51	55	1996	Multifactor explanations of as...	E.F. Fama and K.R. French
12	372	88	0.41	JFE	32	263	1992	The investment opportunity set...	C.W. Smith and R.L. Watts
13	362	27	0.76	RFS	6	327	1993	A closed-form solution for opt...	S.L. Heston
14	357	114	0.35	JF	46	733	1991	Inferring trade direction from...	C.M.C. Lee and M.J. Ready
15	351	55	0.54	JF	42	483	1987	A simple-model of capital-mark...	R.C. Merton
16	348	3	1.81	JFE	9	3	1981	The relationship between retur...	R.W. Banz
17	342	112	0.36	JAR	29	193	1991	Earnings management during imp...	J.J. Jones
18	340	113	0.35	JF	49	1541	1994	Contrarian investment, extrapo...	J. Lakonishok, A. Shleifer A and R.W. Vishny
19	338	60	0.52	JFE	25	23	1989	Business conditions and expect...	E.F. Fama and K.R. French
20	336	171	0.26	JF	50	23	1995	The new issues puzzle	T. Loughran and J.R. Ritter

Table 3.3: The top 20 articles by Google PageRank score. JF, JFE, JME, and MF denote the journals *The Journal of Finance*, *Journal of Financial Economics*, *Journal of Monetary Economics*, and *Mathematical Finance*, respectively. The asterisk (*) denotes articles with $CiteRank$-to-$PageRank$ ratio larger than 10.

Google Rank	Google # ($\times 10^{-3}$)	Cite Rank	Cites	Publication SO	VL	BP	PY	Title	Author(s)
1*	3.01	23	310	JFE	8	205	1980	Measuring security price perfo...	S.J. Brown and J.B. Warner
2	1.96	3	857	JFE	13	187	1984	Corporate financing and invest...	N.S. Majluf and S.C. Myers
3	1.81	16	348	JFE	9	3	1981	The relationship between retur...	R.W. Banz
4*	1.54	84	185	JFE	9	19	1981	Misspecification of capital-as...	M.R. Reinganum
5	1.30	46	238	JME	12	383	1983	Staggered prices in a utility-...	G.A. Calvo, GA
6*	1.30	226	114	JFE	8	139	1980	The effects of capital structu...	R.W. Masulis
7	1.28	1	1227	JFE	33	3	1993	Common risk-factors in the ret...	E.F. Fama and K.R. French
8	1.22	45	240	MF	9	203	1999	Coherent measures of risk	P. Artzner, F. Delbaen, J.M. Eber and D. Heath
9	1.21	60	222	JFE	8	323	1980	On estimating the expected ret...	R.C. Merton
10	1.19	22	315	JME	15	145	1985	The equity premium - a puzzle	R.Mehra and E.C. Prescott
11	1.14	6	477	JFE	14	71	1985	Bid, ask and transaction price...	L.R. Glosten and P.R. Milgrom
12*	1.11	153	139	JME	10	139	1982	Trends and random-walks in mac...	C.R. Nelson and C.I. Plosser
13	1.10	2	878	JF	47	427	1992	The cross-section of expected ...	E.F. Fama and K.R. French
14	1.09	111	166	JFE	8	3	1980	Optimal capital structure unde...	H. Deangelo and R.W. Masulis
15*	1.06	333	96	JFE	8	105	1980	Merger proposals, management d...	P. Dodd
16	1.06	136	147	JFE	8	55	1980	Stock returns and the weekend ...	K.R. French
17	1.01	21	328	JF	40	793	1985	Does the stock-market overreac...	W.F.M. Debondt and R. Thaler
18*	1.00	365	91	JFE	8	179	1980	Trading costs for listed optio...	S.M. Phillips and C.W. Smith
19	0.99	63	218	JFE	12	13	1983	Size-related anomalies and sto...	D.B. Keim
20	0.97	174	130	JFE	8	31	1980	Dealership market - market-mak...	Y. Amihud and H. Mendelson

One interesting pattern for top listed papers in Table 3.2 and Table 3.3 is the exclusivity of a small set of source journals in which these papers are published in. Specifically, we find that top listed articles tend to be published in Journal of Finance (JF) and Journal of Financial Economics (JFE). In Section 2.2.1 (c), we have mentioned that both JF and JFE have 2011 impact factor (IF) scores of 4.218 and 3.725, respectively, yet the Review of Financial Studies (RFS), with its impact factor of 4.748, hardly makes an appearance in either top 20 list. To investigate why this is the case, we need to review the limitations associated to how the impact factor is calculated.

Here, we highlight three main concerns. First, IF cannot readily be used to form unbiased judgments across different fields. This follows from the empirical observation that citation rates vary from one specialization to the next, and therefore, a 2-year target window may provide insufficient time to accumulate publications or citations comparable to those in fast-paced fields (Althouse et al., 2009). A 5-year version of impact factor is provided by ISI, however, these are not widely used.

Second, the formulation of the IF score allows for the manipulation of journal editors (Falagas & Alexiou, 2008; Archambault & Larivière, 2009; Pontille & Torny, 2010). This typically involves editorial strategies that increase the numerator of Equation (1.1) while minimizing the denominator or keeping it unchanged. One strategy, exploits the inclusion of journal self-citations as a means to boost the n_t^i term. While a version of the impact factor without self-citations is provided by ISI, these are often over-looked. Although sometimes necessary (Leslie, 2005), publication delays have also been identified as a strategy to inflate IF calculations (Tort, Targino, & Amaral, 2012).

Third, impact factors are prone to misinterpretation. Reports of increases in impact factor – generally assumed to indicate improvement – must factor in background inflation to account for the tendency of reference lists to grow longer over time coupled with the increasing trend for a significant proportion of those references to cite recent items within

the 2-year target period for IF calculations (Althouse et al., 2009; Neff & Olden, 2010).

Given these intricacies, the usage of alternative measures may help put the impact factor into proper context. For example, pairwise inter-journal citation flows have been tabulated to determine rankings for a small set of journals that include and exclude journal self-citation effects (Borokhovich, Bricker, & Simkins, 1994; Ratnavelu, Fatt, & Ujum, 2012). This approach naturally has an underlying network interpretation: one need only map each journal as a distinct node, with inter-journal flows signifying citations from articles in one journal to another. These flows are represented by directed links pointing from source citing journals to target cited journal, with link weights signifying the magnitude of the citation flow rate or volume.

From a network analysis perspective, journal measures can be constructed based on eigenvector centrality[3]. One effective centrality approach is based on an input-output approach to clique identification (Hubbell, 1965). Kleinberg (1999) traces the first such application to the ranking of journals in physics produced by Pinski and Narin (1976). Salancik (1986) later introduced a variation of the same method which he termed the *structural influence measure*. Salancik's method was subsequently used to determine the relative influence of journals in management (Johnson & Podsakoff, 1994), marketing (Baumgartner & Pieters, 2003), and accounting (Wakefield, 2008). We shall use Salancik's method to rank the influence of journals in the following section.

[3]A significant application can be found in the Eigenfactor and Article Influence score developed by West, Bergstrom and Bergstrom (Bergstrom, 2007; West, Bergstrom, & Bergstrom, 2010). As of 2007, both scores have been adopted by ISI to supplement the JCR impact factors (Franceschet, 2010).

3.2 Journal citation network

Construction—A journal citation network can be constructed by defining the following journal adjacency matrix:

$$B = \begin{pmatrix} B_{11} & B_{12} & \cdots & B_{1N} \\ B_{21} & B_{22} & \cdots & B_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ B_{N1} & B_{N2} & \cdots & B_{NN} \end{pmatrix} \tag{3.1}$$

where $B_{IJ} \in \mathbb{N}$ corresponds to the total citations made from articles i published in citing journal I to articles j published in source cited journal J. This can be done by counting the number of times citing articles in journal I reference any cited article in journal J:

$$B_{IJ} = \sum_{i \in I, j \in J} A_{ij} \tag{3.2}$$

Here, A_{ij} denotes the article adjacency matrix. Journal self-citations may take up non-zero values, hence $B_{II} \geq 0$. The journal network for "Business, Finance" is as depicted in Figure 3.4.

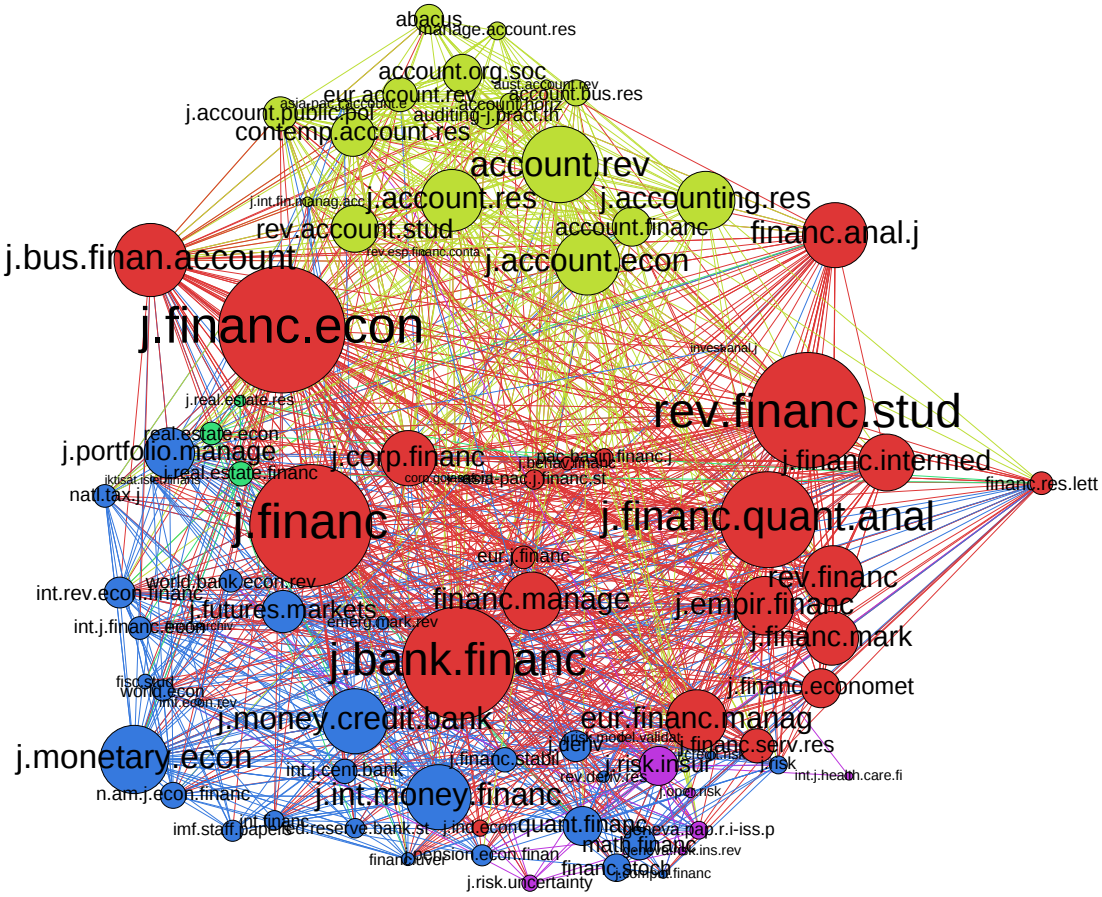

Figure 3.4: Journal citation network for “Business, Finance” (1980–2011). Community detection was carried out using the hierarchical optimization of modularity method developed by Blondel et al. (2008). Community (module) membership is as listed in Table 3.4. Plotted with Gephi.

Table 3.4: Module membership for journals in Figure 3.4.

Module 1 (light green)	**Module 2 (red)**	**Module 4 (blue)**
abacus	j.bus.finan.account	j.futures.markets
account.financ	j.financ.econ	j.portfolio.manage
account.rev	eur.financ.manag	j.money.credit.bank
j.accounting.res	financ.manage	int.rev.econ.financ
account.bus.res	j.financ	j.monetary.econ
account.horiz	rev.financ.stud	j.int.money.financ
account.org.soc	asia-pac.j.financ.st	natl.tax.j
auditing-j.pract.th	j.bank.financ	emerg.mark.rev
contemp.account.res	j.corp.financ	imf.staff.papers
eur.account.rev	j.financ.mark	int.j.financ.econ
j.account.econ	j.financ.quant.anal	n.am.j.econ.financ
j.account.public.pol	pac-basin.financ.j	world.bank.econ.rev
j.account.res	financ.anal.j	world.econ
j.int.fin.manag.acc	eur.j.financ	j.deriv
manage.account.res	j.financ.serv.res	math.financ
rev.account.stud	j.ind.econ	fed.reserve.bank.st
aust.account.rev	rev.financ	quant.financ
asia-pac.j.account.e	j.behav.financ	j.financ.stabil
rev.esp.financ.conta	j.empir.financ	financ.stoch
Module 3 (purple)	corp.gov-oxford	int.financ
int.j.health.care.fi	j.financ.intermed	j.comput.financ
j.risk.insur	j.financ.economet	financ.uver
geneva.pap.r.i-iss.p	financ.res.lett	finanzarchiv
j.oper.risk	invest.anal.j	fisc.stud
j.risk.uncertainty	j.risk.model.validat	iktisat.islet.finans
geneva.risk.ins.rev		imf.econ.rev
Module 5 (dark green)		int.j.cent.bank
j.real.estate.financ		j.pension.econ.finan
real.estate.econ		j.risk
j.real.estate.res		j.credit.risk
		rev.deriv.res
		jassa

Structural influence score—Given N source journals, we can use the index of structural influence (Salancik, 1986; Johnson & Podsakoff, 1994) to measure inter-journal influence, formulated for each journal J as:

$$x_J = D_{1J}x_1 + D_{2J}x_2 + \cdots + D_{NJ}x_N + d_J \tag{3.3}$$

Here, d_J is defined as the J-th element in the $N \times 1$ intrinsic importance vector $\vec{d}$, which is set to a value of 1.0 for each journal J. This serves two functions: first, it quantifies the notion that no one journal is intrinsically more important than any other; and second, it sets up the calculation so that the base-line value for the structural influence score is 1.0. The total citations from journal $I \rightarrow J$ are given by the dependency matrix D with elements derived from Equation 3.2:

$$D_{IJ} = \frac{B_{IJ}}{\sum_J B_{IJ}} \tag{3.4}$$

Furthermore, we require that $D_{II} = 0$ so that the structural influence score quantifies a position of power based on the dependency of other journals. The system of simultaneous linear equations describing the structural influence scores for the entire journal network are then written as:

$$\begin{aligned} \vec{x} &= D^T\vec{x} + \vec{d} \\ (I - D^T)\vec{x} &= \vec{d} \\ \vec{x} &= (I - D^T)^{-1}\vec{d} \end{aligned} \tag{3.5}$$

where I is a $N \times N$ identity matrix, D^T denotes the transpose of matrix D, and the $\cdot^{-1}$ superscript denotes matrix inversion. The non-negativity of the structural influence score $\vec{x}$ depends on whether the leading principal minors of $(I - D^T)$ are positive; this is known as the Hawkins-Simon condition (Hawkins & Simon, 1949). If $(I - D^T)$ satisfies

the Hawkins-Simon condition, then $\det(I - D^T) > 0$ implying that $(I - D^T)$ is non-singular.

Journal ranking—The resulting scores and rankings produced using Salancik's index of structural influence is as listed in Table 3.5 and Table 3.6, respectively. The scores are listed to emphasise the degree to which the centrality of one journal differs from another. The ranks are listed to show the permutation of rank scores with respect to the centrality algorithm used. As it turns out, Journal of Finance scores highest by structural influence score, S, followed by Journal of Financial Economics, Review of Financial Studies, and finally Journal of Financial and Quantitative Analysis. These journals make up the top 4 journals in financial economics corresponding to Module 2 in Table 3.4. Interspersed within the rest of the top 10 ranks are journals in monetary economics (Module 4) and accounting research (Module 1).

Table 3.5: Centrality of "Business, Finance" journals based on inter-journal citation links spanning the 5-year period 2007–2011. Journals are listed by decreasing structural influence score, S. C_D, C_C, C_B, denote degree, closeness, and betweenness centrality, respectively. The in and out superscripts denote in-link and out-link versions of the corresponding centrality algorithm. $PR^{0.86}$, $PR^{0.5}$, $auth$, and hub denotes the Google PageRank score with $d = 0.86$, PageRank with $d = 0.5$, HITS authority, and HITS hub score, respectively.

Source journal	C_D^{in}	C_D^{out}	C_D^{tot}	C_C^{in}	C_C^{out}	C_B	C	$PR^{0.85}$	$PR^{0.5}$	$auth$	hub	S
j.financ	79	33	112	0.2072	0.1216	0.0344	0.2770	0.1821	0.0905	1.0000	0.3513	12349.85
j.financ.econ	79	41	120	0.2240	0.1220	0.0657	0.2509	0.1624	0.0833	0.6192	1.0000	10838.15
rev.financ.stud	72	36	108	0.2337	0.1211	0.0420	0.3068	0.1398	0.0668	0.5482	0.9284	9639.13
j.financ.quant.anal	59	34	93	0.2416	0.1208	0.0623	0.3567	0.0324	0.0204	0.1530	0.4765	2089.09
j.monetary.econ	43	18	61	0.2389	0.1229	0.0467	0.2841	0.0283	0.0234	0.0516	0.0435	1451.90
account.rev	47	34	81	0.2331	0.1218	0.0579	0.1937	0.0307	0.0273	0.0940	0.1811	1441.18
j.money.credit.bank	50	22	72	0.2423	0.1154	0.0149	0.2378	0.0284	0.0242	0.0646	0.0831	1413.43
j.account.econ	42	27	69	0.2331	0.1191	0.0280	0.2522	0.0233	0.0183	0.0769	0.2019	1258.98
j.bank.financ	71	58	129	0.2429	0.1223	0.1276	0.2403	0.0288	0.0301	0.0721	0.7735	1205.62
j.accounting.res	43	7	50	0.2529	0.1147	0.0244	0.2592	0.0201	0.0167	0.0677	0.0085	1047.95
j.financ.intermed	41	24	65	0.2500	0.1230	0.0616	0.3596	0.0138	0.0119	0.0521	0.1827	789.53
j.corp.financ	42	30	72	0.2443	0.1225	0.0720	0.4027	0.0124	0.0114	0.0542	0.5846	691.04
j.account.res	35	27	62	0.2382	0.1189	0.0312	0.2654	0.0135	0.0130	0.0387	0.1912	660.68
j.financ.mark	33	23	56	0.2409	0.1239	0.0549	0.3568	0.0104	0.0100	0.0417	0.1372	583.71
contemp.account.res	35	28	63	0.2402	0.1198	0.0634	0.2751	0.0118	0.0132	0.0230	0.1128	478.94
rev.financ	34	31	65	0.2423	0.1230	0.0635	0.3484	0.0087	0.0093	0.0325	0.1256	468.29
j.bus.finan.account	54	38	92	0.2522	0.1215	0.0740	0.2023	0.0117	0.0140	0.0218	0.2039	461.76
financ.manage	40	31	71	0.2416	0.1218	0.0581	0.3888	0.0085	0.0094	0.0374	0.3291	430.43
financ.anal.j	47	23	70	0.2537	0.1204	0.0633	0.2647	0.0089	0.0116	0.0261	0.1106	357.05
j.int.money.financ	46	34	80	0.2471	0.1246	0.0794	0.2128	0.0101	0.0138	0.0348	0.0769	343.76
math.financ	21	11	32	0.2318	0.1176	0.0138	0.3384	0.0094	0.0122	0.0112	0.0129	310.75

continued on next page . . .

... Table 3.5 continued from previous page

Source journal	C_D^{in}	C_D^{out}	C_D^{tot}	C_C^{in}	C_C^{out}	C_B	C	$PR^{0.85}$	$PR^{0.5}$	$auth$	hub	S
rev.account.stud	28	17	45	0.2318	0.1186	0.0282	0.2737	0.0073	0.0097	0.0211	0.0692	275.48
financ.stoch	19	10	29	0.2240	0.1173	0.0185	0.3817	0.0089	0.0119	0.0093	0.0101	272.66
eur.financ.manag	37	22	59	0.2382	0.1218	0.0446	0.2925	0.0063	0.0090	0.0275	0.1128	248.06
j.empir.financ	33	32	65	0.2409	0.1216	0.0439	0.3600	0.0057	0.0085	0.0224	0.2527	220.61
quant.financ	23	32	55	0.2423	0.1236	0.0529	0.2189	0.0062	0.0105	0.0048	0.0546	136.49
j.portfolio.manage	31	18	49	0.2402	0.1159	0.0174	0.2620	0.0050	0.0087	0.0096	0.0574	135.81
account.org.soc	21	20	41	0.2318	0.1208	0.0457	0.2478	0.0065	0.0114	0.0081	0.0128	125.19
fed.reserve.bank.st	21	13	34	0.2493	0.1188	0.0267	0.2584	0.0038	0.0070	0.0066	0.0103	122.35
j.risk.insur	25	25	50	0.2306	0.1222	0.0350	0.2344	0.0075	0.0133	0.0076	0.0410	119.58
j.futures.markets	26	28	54	0.2318	0.1203	0.0236	0.2604	0.0047	0.0086	0.0177	0.0777	115.64
auditing-j.pract.th	18	20	38	0.2211	0.1108	0.0091	0.3708	0.0045	0.0083	0.0061	0.0314	111.86
real.estate.econ	16	22	38	0.2324	0.1250	0.0395	0.3837	0.0051	0.0094	0.0045	0.0254	87.21
rev.acc.stud	14	0	14	0.2867	0.0115	0.0000	0.2394	0.0030	0.0064	0.0039	0.0000	84.17
j.financ.economet	21	17	38	0.2402	0.1194	0.0297	0.2665	0.0035	0.0073	0.0079	0.0264	83.27
eur.account.rev	21	21	42	0.2275	0.1154	0.0030	0.2367	0.0043	0.0086	0.0055	0.0255	76.98
j.real.estate.financ	15	25	40	0.2402	0.1227	0.0505	0.3491	0.0052	0.0099	0.0016	0.0460	69.01
imf.staff.papers	14	7	21	0.2199	0.1130	0.0056	0.2353	0.0034	0.0074	0.0030	0.0039	62.23
int.j.financ.econ	15	18	33	0.2234	0.1204	0.0225	0.2844	0.0033	0.0072	0.0041	0.0100	60.22
j.financ.serv.res	18	28	46	0.2537	0.1257	0.1014	0.3164	0.0028	0.0066	0.0039	0.0457	55.39
int.j.cent.bank	12	18	30	0.2194	0.1222	0.0229	0.3512	0.0028	0.0066	0.0031	0.0172	49.95
world.bank.econ.rev	12	13	25	0.2183	0.1198	0.0218	0.2527	0.0031	0.0073	0.0027	0.0140	49.95
j.risk.uncertainty	9	4	13	0.2234	0.1148	0.0123	0.2685	0.0029	0.0068	0.0018	0.0028	49.03
j.financ.stabil	12	12	24	0.2293	0.1120	0.0045	0.3067	0.0027	0.0065	0.0051	0.0201	48.51
natl.tax.j	11	19	30	0.2363	0.1215	0.0347	0.2046	0.0029	0.0069	0.0014	0.0201	47.17
j.deriv	16	15	31	0.2409	0.1186	0.0118	0.2602	0.0031	0.0072	0.0057	0.0170	43.87

continued on next page ...

... Table 3.5 continued from previous page

Source journal	C_D^{in}	C_D^{out}	C_D^{tot}	C_C^{in}	C_C^{out}	C_B	C	$PR^{0.85}$	$PR^{0.5}$	$auth$	hub	S
financ.res.lett	11	28	39	0.2287	0.1243	0.0689	0.2660	0.0024	0.0064	0.0034	0.0402	32.60
j.oper.risk	9	14	23	0.2245	0.1198	0.0230	0.2503	0.0027	0.0067	0.0011	0.0038	30.32
n.am.j.econ.financ	13	18	31	0.2409	0.1164	0.0163	0.2554	0.0028	0.0070	0.0009	0.0087	29.08
geneva.pap.r.i-iss.p	9	24	33	0.2194	0.1243	0.0360	0.3929	0.0033	0.0077	0.0013	0.0098	28.57
account.horiz	10	23	33	0.2123	0.1199	0.0120	0.3525	0.0024	0.0063	0.0016	0.0319	28.10
j.ind.econ	9	7	16	0.2199	0.1176	0.0115	0.3054	0.0023	0.0062	0.0035	0.0085	26.17
j.real.estate.res	7	17	24	0.2234	0.1210	0.0171	0.5469	0.0030	0.0072	0.0009	0.0121	26.07
int.rev.econ.financ	16	32	48	0.2436	0.1218	0.0476	0.2359	0.0026	0.0068	0.0034	0.0469	25.34
world.econ	11	19	30	0.2211	0.1208	0.0340	0.1865	0.0027	0.0069	0.0005	0.0075	24.48
account.financ	21	31	52	0.2324	0.1193	0.0300	0.1993	0.0031	0.0078	0.0034	0.0712	24.13
j.account.public.pol	18	24	42	0.2211	0.1170	0.0126	0.2612	0.0028	0.0072	0.0023	0.0575	23.70
j.pension.econ.finan	10	17	27	0.2331	0.1227	0.0409	0.2014	0.0022	0.0062	0.0013	0.0154	22.59
abacus	17	6	23	0.2103	0.1173	0.0085	0.1962	0.0026	0.0068	0.0016	0.0009	21.64
eur.j.financ	11	30	41	0.2318	0.1185	0.0243	0.2879	0.0022	0.0061	0.0034	0.0801	17.44
account.bus.res	13	24	37	0.2199	0.1178	0.0074	0.2221	0.0023	0.0064	0.0014	0.0275	17.30
int.financ	9	17	26	0.2257	0.1223	0.0306	0.3481	0.0023	0.0063	0.0003	0.0078	16.99
manage.account.res	8	9	17	0.2123	0.1130	0.0042	0.4941	0.0024	0.0066	0.0004	0.0018	14.80
geneva.risk.ins.rev	5	7	12	0.1991	0.1181	0.0040	0.4492	0.0025	0.0068	0.0002	0.0029	10.16
asia-pac.j.financ.st	9	27	36	0.2205	0.1203	0.0222	0.3444	0.0020	0.0061	0.0017	0.0852	10.15
rev.deriv.res	4	12	16	0.2134	0.1165	0.0081	0.3146	0.0020	0.0061	0.0005	0.0176	8.92
imf.econ.rev	1	12	13	0.1842	0.1173	0.0023	0.2709	0.0019	0.0058	0.0005	0.0073	7.94
fisc.stud	6	7	13	0.2161	0.1198	0.0161	0.2099	0.0023	0.0066	0.0008	0.0007	7.78
j.comput.financ	3	8	11	0.1982	0.1148	0.0026	0.3574	0.0020	0.0061	0.0000	0.0033	6.47
pac-basin.financ.j	8	24	32	0.2166	0.1180	0.0127	0.3515	0.0020	0.0061	0.0022	0.0989	6.17
asia-pac.j.account.e	4	11	15	0.2145	0.1223	0.0254	0.3058	0.0019	0.0059	0.0003	0.0015	6.01

continued on next page ...

... Table 3.5 continued from previous page

Source journal	C_D^{in}	C_D^{out}	C_D^{tot}	C_C^{in}	C_C^{out}	C_B	C	$PR^{0.85}$	$PR^{0.5}$	$auth$	hub	S
j.risk	8	26	34	0.2293	0.1241	0.0837	0.2223	0.0020	0.0062	0.0008	0.0162	5.87
emerg.mark.rev	5	25	30	0.1937	0.1208	0.0082	0.3006	0.0019	0.0060	0.0013	0.0434	3.84
j.behav.financ	5	16	21	0.2251	0.1178	0.0157	0.2958	0.0019	0.0059	0.0005	0.0298	3.53
finanzarchiv	3	13	16	0.2087	0.1199	0.0071	0.2620	0.0019	0.0060	0.0000	0.0104	2.92
int.j.health.care.fi	2	0	2	0.2575	0.0115	0.0000	0.5000	0.0018	0.0059	0.0001	0.0000	2.83
financ.uver	5	25	30	0.2245	0.1232	0.0255	0.2088	0.0019	0.0060	0.0004	0.0115	2.52
aust.account.rev	3	21	24	0.1982	0.1178	0.0043	0.2289	0.0018	0.0059	0.0002	0.0122	2.46
j.int.fin.manag.acc	4	25	29	0.2103	0.1196	0.0141	0.2046	0.0018	0.0059	0.0002	0.0158	2.05
j.credit.risk	2	17	19	0.2057	0.1229	0.0158	0.2377	0.0018	0.0059	0.0004	0.0093	1.79
j.risk.model.validat	1	15	16	0.1878	0.1229	0.0017	0.3767	0.0018	0.0059	0.0000	0.0073	1.08
rev.esp.financ.conta	0	32	32	0.0115	0.1387	0.0000	0.1927	0.0018	0.0058	0.0000	0.0221	1.00
corp.gov-oxford	0	12	12	0.0115	0.1365	0.0000	0.4613	0.0018	0.0058	0.0000	0.0164	1.00
invest.anal.j	0	8	8	0.0115	0.1307	0.0000	0.3060	0.0018	0.0058	0.0000	0.0062	1.00
iktisat.islet.finans	0	19	19	0.0115	0.1431	0.0000	0.2004	0.0018	0.0058	0.0000	0.0055	1.00
jassa	0	12	12	0.0115	0.1367	0.0000	0.2233	0.0018	0.0058	0.0000	0.0026	1.00
int.insolv.rev	0	1	1	0.0115	0.1227	0.0000	1.0000	0.0018	0.0058	0.0000	0.0011	1.00

Table 3.6: Rank of "Business, Finance" journals based on inter-journal citation links spanning the 5-year period 2007–2011. Journals are listed by decreasing structural influence score S.

Source journal	C_D^{in}	C_D^{out}	C_D^{tot}	C_C^{in}	C_C^{out}	C_B	C	$PR^{0.85}$	$PR^{0.5}$	$auth$	hub	S
j.financ	1	8	3	74	35	30	48	1	1	1	6	1
j.financ.econ	2	2	2	50	29	8	29	2	2	2	1	2
rev.financ.stud	3	4	4	30	38	24	59	3	3	3	2	3
j.financ.quant.anal	5	5	5	16	40	12	70	4	8	4	5	4
j.monetary.econ	11	49	19	26	17	20	49	8	7	12	33	5
account.rev	8	6	7	31	33	15	3	5	5	5	13	6
j.money.credit.bank	7	39	10	13	77	56	24	7	6	9	21	7
j.account.econ	13	22	13	32	57	37	30	9	9	6	10	8
j.bank.financ	4	1	1	12	24	1	26	6	4	7	3	9
j.accounting.res	12	78	26	5	81	41	34	10	10	8	68	10
j.financ.intermed	15	31	16	7	15	13	73	11	18	11	12	11
j.corp.financ	14	16	9	10	23	6	81	13	20	10	4	12
j.account.res	18	23	18	27	58	32	41	12	15	14	11	13
j.financ.mark	21	37	21	19	12	16	71	16	23	13	14	14
contemp.account.res	19	18	17	22	50	10	47	14	14	20	16	15
rev.financ	20	14	15	14	16	9	65	21	28	17	15	16
j.bus.finan.account	6	3	6	6	36	5	8	15	11	22	9	17
financ.manage	16	13	11	17	31	14	79	22	27	15	7	18
financ.anal.j	9	36	12	3	44	11	40	19	19	19	18	19
j.int.money.financ	10	7	8	9	8	4	13	17	12	16	24	20
math.financ	28	72	49	40	68	58	62	18	16	25	55	21
rev.account.stud	24	54	30	37	61	36	46	24	25	23	26	22
financ.stoch	34	74	58	51	70	49	77	20	17	27	62	23
eur.financ.manag	17	40	20	28	32	22	52	26	29	18	17	24
j.empir.financ	22	9	14	18	34	23	74	28	33	21	8	25
quant.financ	27	10	22	15	13	17	14	27	22	36	29	26
j.portfolio.manage	23	50	27	23	76	50	39	31	30	26	28	27
account.org.soc	29	44	34	39	41	21	27	25	21	28	56	28
fed.reserve.bank.st	30	64	43	8	59	38	33	35	44	31	61	29
j.risk.insur	26	26	25	41	27	28	19	23	13	30	35	30
j.futures.markets	25	19	23	36	46	43	36	32	32	24	23	31
auditing-j.pract.th	35	45	38	56	85	65	75	33	34	32	38	32
real.estate.econ	39	41	37	35	7	26	78	30	26	37	43	33
rev.acc.stud	44	86	77	1	86	86	25	44	59	40	86	34
j.financ.economet	31	55	39	25	55	35	43	36	38	29	41	35
eur.account.rev	32	42	32	45	78	76	22	34	31	34	42	36
j.real.estate.financ	42	27	35	24	20	18	66	29	24	53	31	37
imf.staff.papers	45	79	68	60	82	71	20	37	37	47	75	38
int.j.financ.econ	43	51	46	52	45	46	50	39	41	38	63	39
j.financ.serv.res	36	20	29	4	6	2	61	48	55	39	32	40
int.j.cent.bank	48	53	56	63	28	45	67	49	53	46	48	41
world.bank.econ.rev	49	65	61	64	52	48	31	40	39	48	54	42

continued on next page ...

... Table 3.6 continued from previous page

Source journal	C_D^{in}	C_D^{out}	C_D^{tot}	C_C^{in}	C_C^{out}	C_B	C	$PR^{0.85}$	$PR^{0.5}$	$auth$	hub	S
j.risk.uncertainty	57	84	80	54	79	61	44	46	48	51	79	43
j.financ.stabil	50	67	64	43	84	72	58	52	57	35	46	44
natl.tax.j	51	46	54	29	37	29	9	45	47	57	45	45
j.deriv	40	61	51	21	60	63	35	42	40	33	49	46
financ.res.lett	52	21	36	44	9	7	42	57	60	43	36	47
j.oper.risk	58	63	65	49	51	44	28	51	52	61	76	48
n.am.j.econ.financ	46	52	50	20	75	52	32	50	45	63	66	49
geneva.pap.r.i-iss.p	59	34	44	62	10	27	80	38	36	60	64	50
account.horiz	55	38	45	69	48	62	69	59	61	55	37	51
j.ind.econ	60	80	75	61	69	64	55	61	63	41	67	52
j.real.estate.res	66	58	63	53	39	51	86	43	43	62	58	53
int.rev.econ.financ	41	11	28	11	30	19	21	54	49	45	30	54
world.econ	53	47	55	57	42	31	1	53	46	68	70	55
account.financ	33	15	24	34	56	34	5	41	35	44	25	56
j.account.public.pol	37	32	31	55	73	60	37	47	42	49	27	57
j.pension.econ.finan	56	56	59	33	21	25	7	64	64	58	53	58
abacus	38	83	66	72	71	66	4	55	50	54	84	59
eur.j.financ	54	17	33	38	62	42	51	65	66	42	22	60
account.bus.res	47	33	40	59	66	69	15	60	58	56	40	61
int.financ	61	57	60	46	25	33	64	63	62	73	69	62
manage.account.res	63	75	71	70	83	74	84	58	56	70	81	63
geneva.risk.ins.rev	68	82	83	76	63	75	82	56	51	77	78	64
asia-pac.j.financ.st	62	24	41	58	47	47	63	67	70	52	20	65
rev.deriv.res	72	68	74	68	74	68	60	69	69	67	47	66
imf.econ.rev	80	69	78	81	72	78	45	75	81	69	72	67
fisc.stud	67	81	79	66	53	53	12	62	54	64	85	68
j.comput.financ	75	76	84	78	80	77	72	68	68	80	77	69
pac-basin.financ.j	64	35	48	65	64	59	68	70	67	50	19	70
asia-pac.j.account.e	73	73	76	67	26	40	56	74	78	74	82	71
j.risk	65	25	42	42	11	3	16	66	65	65	51	72
emerg.mark.rev	69	28	52	79	43	67	54	71	71	59	34	73
j.behav.financ	70	60	67	47	65	55	53	76	75	66	39	74
finanzarchiv	76	66	73	73	49	70	38	72	72	79	60	75
int.j.health.care.fi	78	87	86	2	87	87	85	77	77	78	87	76
financ.uver	71	29	53	48	14	39	11	73	73	72	59	77
aust.account.rev	77	43	62	77	67	73	18	79	76	76	57	78
j.int.fin.manag.acc	74	30	57	71	54	57	10	78	74	75	52	79
j.credit.risk	79	59	70	75	18	54	23	80	79	71	65	80
j.risk.model.validat	81	62	72	80	19	79	76	81	80	81	71	81
rev.esp.financ.conta	82	12	47	82	2	81	2	82	82	83	44	82
corp.gov-oxford	84	70	81	84	4	83	83	86	86	87	50	83
invest.anal.j	86	77	85	86	5	84	57	85	85	84	73	84
iktisat.islet.finans	83	48	69	83	1	80	6	83	83	82	74	85
jassa	85	71	82	85	3	82	17	84	84	86	80	86
int.insolv.rev	87	85	87	87	22	85	87	87	87	85	83	87

3.3 Identifying experts and authorities

Having explored the document and journal citation network, we now move on to uncovering remarkable features in author citation networks. A plot of the giant weakly connected component of the ACN is as shown in Figure 3.5. Its properties are as tabulated in Table 3.7 below.

Table 3.7: Properties of author citation network (ACN).

Nodes	12,627
Links	587,611
No. of connected components	34
Size of giant weakly connected component	12,539
Links on giant weakly connected component	587,541
Giant Weakly Connected Component (GWCC)	
Density	3.7×10^{-3}
Average path length	3.5
Diameter	15
Transitivity	0.144
Mean degree	46.9
Median degree	10.0
Maximum degree	3186.0
Modularity	0.395
No. of resolved communities	10

The Spearman rank correlation coefficients[4], r_s, between select node attributes (i.e. career time, last rest time, number of coauthors, number of publications, total citation count, h-index, PageRank score by Coarse-Grain scheme, and PageRank by Yang-Yin-Davison scheme) are as shown in Table 3.8 (and also Figure 3.6).

For this set of attributes, each pair is found to exhibit significant correlation at the level of $p < 0.01$ except for number of citations n_C against last rest time λ ($r_s = -0.02$, $p = 0.06$). This implies that relative to the study data, there is no clear overall pattern

[4]Spearman's rank correlation coefficient is a non-parametric measure of statistical dependence between two variables. Specifically, it is defined as the Pearson correlation coefficient between ranked variables. Given a sample of size n, x_i and y_i are the ranks of the i-th value of scores X and Y, from which one can compute the Spearman correlation coefficient as $r_s = 1 - \left(6 \sum d_i^2\right) / \left(n(n^2 - 1)\right)$, where $d_i = x_i - y_i$ is the difference between ranks (Myers, Well, & Lorch, 2010). In the presence of outliers in the tails of the samples, Spearman correlation is less sensitive compared to Pearson correlation since the contribution of outliers are limited by the value of their rank rather than by their magnitude. Hence, Spearman correlation is a more suitable measure for comparing author attributes, which tend to be heavy-tailed (Clauset, Shalizi, & Newman, 2009).

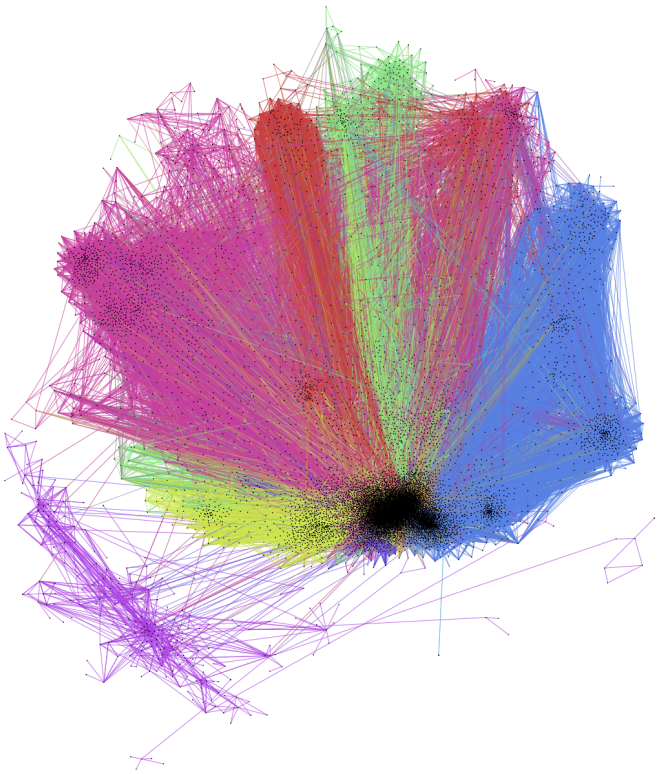

Figure 3.5: Giant weakly connected component of author citation network (DCN). Nodes are color-coded via community detection method of Blondel et al. (2008) and plotted using Gephi.

linking author citation counts with the time elapsed since last publication. However, all other attributes show statistically significant correlation:

- Career time, τ: As expected, a longer career time typically goes hand in hand with a larger publication count (n_P) and larger coauthor count (k), which in theory, allows for more citations accrued, hence τ is also positively correlated with an author's citation count, h-index, as well as PageRank score by Coarse-Grain (PR^C) and Yang-Yin-Davison scheme (PR^Y), respectively. There is a weak negative correlation ($r_s = -0.11, p < 0.01$) between τ and last rest times indicating some tendency for long career times to be accompanied by short last rest times and vice versa.

Table 3.8: Spearman rank correlation coefficient for node attributes on giant component of the author citation network constructed in this study. h-index scores are estimated based on articles limited to journals in the study dataset (i.e. ISI-indexed articles published under the"Business, Finance" subject category spanning the period 1980-2011). Values in the lower triangle correspond to correlation p-values.

	τ	λ	k	n_P	n_C	h^*	PR^C	PR^Y
Career time, τ	-	-0.11	0.42	0.61	0.49	0.50	0.50	0.58
Last rest time, λ	0.00	-	-0.36	-0.23	0.02	-0.04	0.19	-0.13
No. of coauthors, k	0.00	0.00	-	0.67	0.46	0.53	0.37	0.51
No. of publications, n_P	0.00	0.00	0.00	-	0.59	0.71	0.55	0.71
No. of citations, n_C	0.00	0.06	0.00	0.00	-	0.85	0.84	0.79
h-Index, h^*	0.00	0.00	0.00	0.00	0.00	-	0.73	0.75
PageRank for CG scheme, PR^C	0.00	0.00	0.00	0.00	0.00	0.00	-	0.89
PageRank for YYD scheme, PR^Y	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-

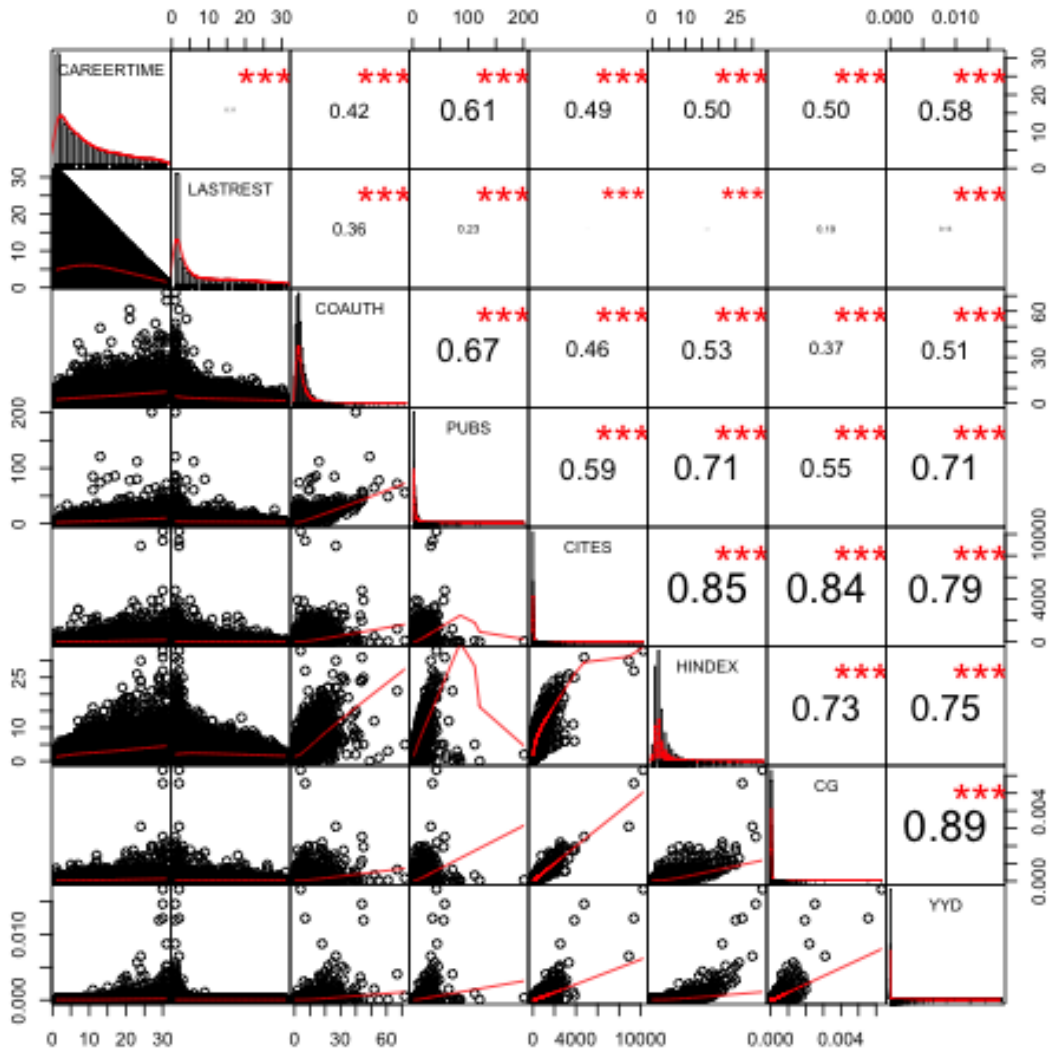

Figure 3.6: Scatterplot of correlation matrix in Table 3.8. Graphic is produced using the PerformanceAnalytics package in R (Carl et al., 2009).

- PR^C and PR^Y score: The Spearman rank correlation coefficient is strongest between PageRank scores for the Coarse-Grain and Yang-Yin-Davison scheme due to similarities in their construction (see Figures 2.6 to 2.7). The observed differences are due to the inclusion of the individual temporal importance (ITI) score when computing link weights according to the Yang-Yin-Davison scheme (see Equation (2.12)).

- Last rest time, λ: Nodes with high PageRank score by Coarse-Grain scheme tend to exhibit longer last rest times. This is because high values of PR^C tend to go to senior researchers who author the earliest influential publications, some of whom are no longer actively publishing. To a limited extent, it appears that authors with short last rest times tend to have more coauthors, more publications, and higher PageRank score by Yang-Yin-Davison scheme.

- Number of coauthors, k: Higher coauthor count tends to signal more publications, each of which has some potential to rack up citations, thus boosting h-index and PageRank score (since these measures are proportional to citation counts).

- Number of publications, n_P: Evidently, there is a sizeable positive correlation between the number of publications by an author with his/her total citations, and by extension, his/her h-index and PageRank score.

- Number of citations, n_C: PageRank scores computed using the Coarse-Grain scheme exhibits stronger (linear) correlation with an author's total citation count compared to the Yang-Yin-Davison scheme since Equation (2.11) is directly proportional to citation count (see Figure 2.6). While similar in construction to the Coarse-Grain scheme, the Yang-Yin-Davison scheme takes into account additional information about the author, namely, his/her temporal characteristics (see Figure 2.7).

- h-index: Evidently, both the Coarse-Grain and Yang-Yin-Davison schemes show similarly strong correlation with h-index, although less in magnitude compared to the correlation of both schemes with citation count. Since the correlation is not perfect, there are differences in the author rankings produced by citation count, h-index, as well as both PageRank scores (PR^C and PR^Y).

The top 20 ranks by weighted PageRank score PR for both the Coarse-Grain (CG) and Yang-Yin-Davison (YYD) scheme are listed in Table 3.9. A common feature for Table 3.9(a) and Table 3.9(b) is that both lists are not strongly ordered by decreasing citation count (as indicated by the number of in-links) and hence provides an alternative take on researcher performance.

Another common feature for the top 20 spots according to the CG and YYD schemes are the similarity in the range, mean and median of τ ($CareerTime$). Here, however, the similarities end. Both methods are seen to produce markedly different rankings in terms of the temporal characteristics of researchers as depicted in the distribution of ITI values. For the top 20 ranks based on the CG scheme, ITI ranges from 0.3 to 57 (mean = 18.1, median = 13.8). In contrast, the top 20 ranks based on the YYD scheme has ITI ranging from 13.0 to 71 (mean = 34.4, median = 33.0).

This can be traced to differences in the range, mean and median for λ ($LastRestTime$) and ϕ ($PubInterval$), which is by design. As expressed in Equations (2.2) to (2.13), the YYD link weight scheme was devised to boost the ranking of highly cited researchers who have long $CareerTime$, short average $PubInterval$, and small $LastRestTime$. Such characteristics more appropriately capture the publication and citation characteristics of academic authorities/experts compared to the simpler link weight scheme based on author citation influence defined in Equation (2.7). We now proceed to make some remarks on the reasonableness of the ranking produced in Table 3.9.

Eugene Fama—For both the CG and YYD schemes, the top position is assigned to "fama, ef" which represents the financial economist Eugene Francis Fama who currently holds the Robert R. McCormick Distinguished Service Professor of Finance chair at the University of Chicago. Fama is widely recognized as the "father of modern finance"[5] for his groundbreaking work on random walk models of stock price movements Fama (1965) and the efficient market hypothesis Malkiel and Fama (1970). He is also the recipient of many honors and awards for his long and distinguished service in finance research (some of which are listed in Table 3.10). Recently, Fama was awarded the 2013 Nobel Memorial Prize in Economic Sciences which he shares with Robert Shiller and Lars Peter Hansen.

Based on Table 3.9, we see that while Fama does not score highest in ITI score, he possesses the largest citation count (and h-index) as indicated by the number of citation in-links. Furthermore, the weighted PageRank score of a given node i is proportional to the number of its in-linking nodes j as well as their individual PageRank score, and the propagation factor with respect to node i, $P(j, i)$. Hence, it is conceivable that Fama gains top rank not only through the sheer number of author citations but also through the influence of his works on other influential workers.

This is easy enough to verify. By tracing links on the ACN network, we have found that all the other top 20 "authoritative" researchers have cited Fama's works. Incidentally, 26 out of Fama's 43 papers in the ISI subject category of "BUSINESS, FINANCE" (over the period 1980-2011) are coauthored with Kenneth Ronald French i.e. "french.kr". Not surprisingly, French is positioned at #3 on the YYD ranking since he and Fama share exactly the same author citations for those 26 jointly-authored works.

René Stulz—In the second rank on the YYD scoreboard is finance professor René M. Stulz ("stulz.rm") of Ohio State University's Fisher College of Business. Stulz is also

[5]The University of Chicago Booth School of Business biography page available at http://www.chicagobooth.edu/faculty/directory/f/eugene-f-fama

a recipient of the Fama-DFA and Jensen prize (see Table 3.10) and a long-time active member of the publishing community in finance. He was previously editor of the Journal of Finance for twelve years (1988-2000) and has published extensively throughout his career in the three most influential journals in finance (citation needed), namely, the Journal of Finance (9 papers though not during his tenure as editor), Journal of Financial Economics (27 papers) and the Review of Financial Studies (9 papers)[6].

Although having slightly less than half the number of in-links compared to Fama, Stulz has nearly three times the ITI score (see Table 3.9). Furthermore, 14 of the YYD top 20 researchers form in-links to Stulz i.e. "fama.ef" (#1), "french.kr" (#3), "titman.s" (#4), "roll.r" (#5), "shleifer.a" (#6), "harvey.cr" (#7), "kothari.sp" (#10), "amihud.y" (#11), "masulis.rw" (#12), "saunders.a" (#13), "ritter.jr" (#15) , "subrahmanyam.a" (#16), "berger.an" (#18), and "bekaert.g" (#19). Hence, the propagation factor from each of these nodes to "stulz.rm" adds considerable weight by virtue of Stulz's high ITI score and the significant citation influence of his citing authors.

When further combined with the sheer number of in-links, this results in a weighted PageRank score that is second only to Eugene Fama – despite having relatively fewer citation in-links than, say, "french.kr" or "shleifer.a". This emphasizes the utility of citation network analysis over traditional methods in terms of detecting important or subtle features within a structure of citation linkages. Table 3.10 lists some other notable prize winning researchers. Among these are finance professors Sheridan Titman, Richard Roll, and Andrei Shleifer listed as 4th, 5th, and 6th most influential by YYD ranking, respectively.

Sheridan Titman—Professor Titman of The University of Texas at Austin, previously held the post of special assistant to the Assistant Secretary of the Treasury for Economic

[6]The Ohio State University Fisher College of Business biography page available at http://fisher.osu.edu/fin/faculty/stulz/.

Policy in Washington D.C. and is currently director of the Energy Management and Innovation Center at University of Texas[7]. Titman co-authored the influential textbook "Financial Markets and Corporate Strategy" and has published leading papers on corporate finance and investments.

Richard Roll—On the other hand, Richard Roll – who currently holds the Distinguished Professor, Joel Fried Chair in Applied Finance at UCLA Anderson School of Management – though ranked at #5, is by all means an intellectual giant. He was awarded the Irving Fisher Prize for the best American dissertation in economics in 1968, is a four-time winner of the Graham and Dodd Award for financial writing, and was also accorded the Leo Melamed Award for best financial research by an American business school professor. Additionally, Roll has published over 100 articles in highly acclaimed peer-reviewed journals since 1966[8].

Andrei Shleifer—Of similar prolificity is Andrei Shleifer of Harvard University, the 1999 recipient of the John Bates Clark Medal , whom, as of August 2013, is reputedly the most cited economist in the world according to RePEc[9]. From these examples, it seems that there is a great deal of information embedded within a structure of author citation linkages, which, in turn, allows us to extract network features that correspond to either highly influential researchers (CG scheme) or authorities/experts (YYD scheme) even in the absence of auxiliary information describing their various accolades and achievements.

[7]The University of Texas at Austin biography page available at http://www.utexas.edu/opa/experts/profile.php?id=393

[8]Curriculum vitae and UCLA Anderson School of Management biography page available at http://www.anderson.ucla.edu/faculty/finance/faculty/roll

[9]Awarded by the American Economic Association for "that American economist under the age of forty who is adjudged to have made a significant contribution to economic thought and knowledge". See http://www.aeaweb.org/honors_awards/clark_medal.php.

Table 3.9: Top 20 ranks by weighted PageRank score. Several notations are used for brevity: ranks are denoted by $R^{(\cdot)}$ for either the CG or YYD link weighting scheme (indicated in superscripted brackets as C and Y, respectively), weighted PageRank scores for either network are denoted in the same way as $PR^{(\cdot)}$, τ is $CareerTime$, λ is $LastRestTime$, ϕ is the publication interval $PubInterval$, ITI is the individual temporal importance, k is the number of coauthors, n_P is the number of publications, and n_C is the number of citation in-links. The asterisk on the column label h* indicates that the h-index was computed based on publication and citation data limited to ISI journal articles indexed under the "BUSINESS, FINANCE" subject category over the period 1980–2011.

(a) Coarse-Grain (CG)

Author keyword	R^C	R^Y	τ	λ	φ	*ITI*	k	n_P	n_C	h*	PR^C ($\times 10^{-3}$)	PR^Y ($\times 10^{-3}$)
fama.ef	1	1	30	2	0.71	21.0	4	43	10255	33	6.32	16.89
french.kr	2	3	30	2	0.86	17.5	7	36	9384	27	5.56	12.47
shleifer.a	3	6	24	2	0.75	16.0	27	33	8913	30	3.10	6.69
stulz.rm	4	2	30	1	0.53	57.0	44	58	4762	31	2.52	14.64
roll.r	5	5	31	1	0.72	43.0	18	44	2547	21	2.21	8.58
stambaugh.rf	6	48	27	3	1.23	7.3	9	23	2569	21	1.99	1.60
warner.jb	7	28	31	1	2.38	13.0	13	14	2811	11	1.97	2.47
myers.sc	8	27	27	1	2.70	10.0	9	11	2960	8	1.96	2.52
titman.s	9	4	29	1	0.53	55.0	45	56	3858	25	1.95	12.16
brown.sj	10	17	31	1	1.55	20.0	17	21	2454	14	1.85	3.60
lucas.re	11	126	27	5	5.40	1.0	2	6	3870	6	1.77	0.67
lakonishok.j	12	20	29	3	0.74	13.0	32	40	3110	25	1.75	3.09
vishny.rw	13	67	22	2	2.20	5.0	6	11	3749	11	1.75	1.23
smith.cw	14	24	30	2	1.03	14.5	19	30	2456	20	1.74	2.75
calvo.ga	15	70	20	9	1.33	1.7	8	16	1500	10	1.64	1.17
ross.sa	16	62	29	3	1.26	7.7	18	24	1698	21	1.57	1.36
amihud.y	17	11	31	1	1.03	30.0	20	31	1905	15	1.54	4.35
masulis.rw	18	12	31	1	1.15	27.0	26	28	1827	18	1.52	4.14
plosser.ci	19	457	12	18	2.00	0.3	3	7	2356	7	1.47	0.21
keim.db	20	209	22	7	1.29	2.4	13	18	1656	14	1.46	0.44

(b) Yang-Yin-Davison (YYD)

Author keyword	R^C	R^Y	τ	λ	φ	*ITI*	k	n_P	n_C	h*	PR^C ($\times 10^{-3}$)	PR^Y ($\times 10^{-3}$)
fama.ef	1	1	30	2	0.71	21.0	4	43	10255	33	6.32	16.89
stulz.rm	4	2	30	1	0.53	57.0	44	58	4762	31	2.52	14.64
french.kr	2	3	30	2	0.86	17.5	7	36	9384	27	5.56	12.47
titman.s	9	4	29	1	0.53	55.0	45	56	3858	25	1.95	12.16
roll.r	5	5	31	1	0.72	43.0	18	44	2547	21	2.21	8.58
shleifer.a	3	6	24	2	0.75	16.0	27	33	8913	30	3.10	6.69
harvey.cr	29	7	23	1	0.59	39.0	25	40	3332	26	1.29	5.72
verrecchia.re	22	8	31	1	0.89	35.0	22	36	2524	22	1.44	5.22
larcker.df	47	9	31	1	0.69	45.0	31	46	2463	25	0.94	5.13
kothari.sp	32	10	24	1	0.73	33.0	33	34	2410	24	1.18	4.67
amihud.y	17	11	31	1	1.03	30.0	20	31	1905	15	1.54	4.35
masulis.rw	18	12	31	1	1.15	27.0	26	28	1827	18	1.52	4.14
saunders.a	93	13	31	1	0.44	71.0	67	72	1198	21	0.65	3.96
whaley.re	30	14	30	1	0.91	33.0	19	34	1874	20	1.23	3.76
ritter.jr	24	15	25	1	1.19	21.0	16	22	2660	17	1.43	3.75
subrahmanyam.a	95	16	20	1	0.48	42.0	25	43	1707	19	0.65	3.66
brown.sj	10	17	31	1	1.55	20.0	17	21	2454	14	1.85	3.60
berger.an	43	18	24	1	0.63	38.0	43	39	2434	24	0.97	3.48
bekaert.g	53	19	19	1	0.59	32.0	22	33	2304	22	0.83	3.17
lakonishok.j	12	20	29	3	0.74	13.0	32	40	3110	25	1.75	3.09

Table 3.10: Prizes won by top 20 authorities/experts listed in Table 3.9(b). The Brattle Group and Smith Breeden prizes are awarded for articles published in the Journal of Finance. Similarly, the Fama-DFA and Jensen prizes are awarded for articles published in the Journal of Financial Economics. Superscripts placed after each author keyword denotes the corresponding YYD rank.

Prize	Year awarded	Placement	Author keyword$^{YYD\ Rank}$	Reference
Brattle Group	1999	Distinguished Paper	shleifer.a^{6}	La Porta et al. (1999)
Fama-DFA	1998	First Prize	fama.ef^{1}	Fama (1998)
Fama-DFA	1998	Second Prize	subrahmanyam.a^{16}	Brennan et al. (1998)
Fama-DFA	1999	First Prize	saunders.a^{13}	Gande et al. (1999)
Fama-DFA	2000	First Prize	roll.r^{5}	Chordia et al. (2000)
Fama-DFA	2004	First Prize	stulz.rm^{2}	Doidge et al. (2004)
Fama-DFA	2004	Second Prize	fama.ef^{1}, french.kr^{3}	Fama and French (2004)
Fama-DFA	2011	Second Prize	saunders.a^{13}	Massoud et al. (2011)
Jensen	2000	Second Prize	shleifer.a^{6}	La Porta et al. (2000)
Jensen	2001	First Prize	harvey.cr^{7}	Graham and Harvey (2001)
Jensen	2002	Second Prize	shleifer.a^{6}	Shleifer and Wolfenzon (2002)
Jensen	2003	First Prize	shleifer.a^{6}	Shleifer and Vishny (2003)
Jensen	2005	First Prize	harvey.cr^{7}	Brav et al. (2005)
Jensen	2006	Second Prize	fama.ef^{1}, french.kr^{3}	Fama and French (2006)
Jensen	2008	First Prize	stulz.rm^{2}	Bargeron et al. (2008)
Jensen	2010	First Prize	ritter.jr^{15}	Gao and Ritter (2010)
Jensen	2010	Second Prize	stulz.rm^{2}	De Angelo et al. (2010)
Smith Breeden	1991	First Prize	ritter.jr^{15}	Ritter (1991)
Smith Breeden	1991	Distinguished Paper	harvey.cr^{7}	Harvey (1991)
Smith Breeden	1992	First Prize	fama.ef^{1}, french.kr^{3}	Fama and French (1992)
Smith Breeden	1995	Distinguished Paper	lakonishok.j^{20}, shleifer.a^{6}	Lakonishok et al. (1994)
Smith Breeden	1997	First Prize	titman.s^{4}	Daniel and Titman (1997)

3.4 Identifying rising stars

A portrait of the scientist as a young man— Some modifications can be made to the YYD scheme implemented thus far to highlight specific features of interest. For example, consider the task of identifying rising stars. One possible approach is to bias the YYD scheme to push up the ranks of researchers with shorter $CareerTime$ and $LastRestTime$. Specifically, this short-age bias could be inserted into Equation (2.7) so that the CIR contribution is especially significant when a senior authority cites a "junior" researcher.

To this end, we propose:

$$CI(a_{ji}) = \sum_{p_{ji}:a_j \to a_i} CIR(p_{ji}) \frac{\nu^{\tau_j - \tau_i}}{(\lambda_i \lambda_j)} \tag{3.6}$$

where τ represents $CareerTime$, λ represents $LastRestTime$, and ν is the age decay base. The influence score contributed from researcher j to i is proportional to the *generation gap* $\tau_j - \tau_i$. The *quiescence product* $\lambda_i \lambda_j$ penalises contributions (citations) to researchers i fitting the following criteria: (1) those who have not published recently (relative to the census year); and/or (2) those who do not work contemporaneously with the citing researcher j.[10] It is possible to modify Equation (3.6) to use a *quiescence ratio* λ_j / λ_i instead so that:

$$CI(a_{ji}) = \sum_{p_{ji}:a_j \to a_i} CIR(p_{ji}) \nu^{\tau_j - \tau_i} \frac{\lambda_j}{\lambda_i} \tag{3.7}$$

[10]For instances where researcher i has short $CareerTime$ relative to researcher j, but has not been active in the field for a number of years (has shifted work outside of the field, is retired, or is deceased).

or *quiescence gap* $|\lambda_j - \lambda_i|$:

$$CI(a_{ji}) = \sum_{p_{ji}:a_j \to a_i} CIR(p_{ji}) \frac{\nu^{\tau_j - \tau_i}}{|\lambda_j - \lambda_i| + 1} \tag{3.8}$$

These expressions however do not penalize instances where both j and i have similarly large $LastRestTime$ and therefore make for a less attractive modification. Dimension analysis dictates that Equation (3.6) has the units of citations/time-squared, but the time factor can easily be removed by replacing λ_i and λ_j each with λ_i/λ_{max} and λ_j/λ_{max} respectively, where λ_{max} denotes the maximum value of $LastRestTime$. This leaves Equation (3.6) dimensionally equivalent to Equation (2.7). We shall refer to predictions (scores) generated by Equation (3.6) as the age-biased YYD link weight scheme, or YYD+, for short. We list the top 20 scores in Table 3.11 and Table 3.12. The affiliations of the top 50 rising stars are included as well in order to highlight the Ivy League university affiliations, where applicable.

Most entries in Table 3.11 are ranked above the 50th position by YYD ranking. The three exceptions correspond to "subrahmanyam.a", "bekaert.g", "graham.jr"; the first two being YYD top 20 entries. These three instances are characterized by long $CareerTime$ (see Table 3.12) and therefore point to a weakness in the scheme. However, taking a closer look at all other entries, we find that these instances (except "o'hara.m") have $CareerTime \leq 11$ which corresponds to a first publication not earlier than the year 2000. Such instances point to the effectiveness of the scheme in indentifying researchers with short $CareerTime$ yet are cited by senior authorities – perhaps a good working definition for an "rising star". One indication that we are on the right track is to find award winners in this list.

Brattle Group Prize—Winners of this award include Jeffrey Wurgler at First Prize in 2002 (Baker & Wurgler, 2002), Heitor Almeida at First Prize in 2008 (Almeida &

Philippon, 2007), and Thorsten Beck for Distinguished Paper in 2010 (Beck, Levine, & Levkov, 2010). For the Fama-DFA Prize we have two-time winner Joseph Chen at Second Place in 2001 (Chen, Hong, & Stein, 2001) and at First Place in 2002 (Chen, Hong, & Stein, 2002). Another two-time winner of the Fama-DFA prize is Viral V. Acharya – once with Lasse H. Pedersen at First Place in 2005 (Acharya & Pedersen, 2005), and yet again at Second Place in 2007 (Acharya, Bharath, & Srinivasan, 2007).

Jensen, Smith-Breeden Prize—Jensen Prize winners include Viral V. Acharya at First Place in 2000 (Acharya, John, & Sundaram, 2000) and Heitor Almeida at Second Place in 2005 (Almeida & Wolfenzon, 2005). Finally, the Smith Breeden Prize winners include Martin Lettau at First Prize in 2001 (Campbell, Lettau, Malkiel, & Xu, 2001), Amir Yaron for Distinguished Paper in 2004 (Bansal & Yaron, 2004), and Lu Zhang at First Prize in 2005 (Zhang, 2005).

Furthermore, if we include Maureen O'Hara ("o'hara") and John R. Graham ("graham.jr") – both of which have $CareerTime \geq 14$ according to the study dataset – we find that O'Hara has won the Smith Breeden Distinguished Prize on three separate occasions i.e. 2000, 2002, and 2003 (Ellis, Michaely, & O'Hara, 2000; Easley, Hvidkjaer, & O'Hara, 2002; O'Hara, 2003), while Graham is a three-time First Place Jensen Prize winner for the years 2001, 2005, and 2006 (Graham & Harvey, 2001; Brav, Graham, Harvey, & Michaely, 2005; Graham & Tucker, 2006). This lends some limited support for the effectiveness of the YYD+ scheme in identifying "rising stars". We shall seek to improve on this scheme (and develop an objective evaluation criteria) in future works.

Table 3.11: Top 20 ranks by weighted PageRank score according to the age-biased YYD link weight scheme (YYD+). The following notations are used for brevity: ranks are denoted by $R^{(\cdot)}$ for the CG, YYD, or YYD+ link weight scheme (indicated in superscripted brackets as C, Y, and Y+, respectively), weighted PageRank scores for the three networks are denoted in the same way as $PR^{(\cdot)}$.

Author keyword	R^C	R^Y	R^{Y+}	PR^C	PR^Y	PR^{Y+}
beck.t	168	52	1	0.50	1.54	5.49
ang.a	295	136	2	0.31	0.61	5.35
graham.jr	119	36	3	0.61	1.96	3.97
zhang.l	1972	547	4	0.09	0.18	3.97
wurgler.j	283	146	5	0.32	0.58	3.48
o'hara.m	240	82	6	0.36	1.02	3.05
xing.yh	2030	1000	7	0.09	0.10	2.76
campello.m	1260	361	8	0.12	0.26	2.42
bekaert.g	53	19	9	0.83	3.17	2.29
lettau.m	431	369	10	0.25	0.25	2.20
acharya.vv	946	230	11	0.15	0.38	2.14
pedersen.lh	770	432	12	0.17	0.22	2.03
kumar.a	3585	1462	13	0.06	0.08	2.00
sadka.r	2123	667	14	0.09	0.15	1.99
lins.kv	936	344	15	0.15	0.26	1.96
almeida.h	1227	414	16	0.13	0.23	1.92
yaron.a	1093	528	17	0.13	0.18	1.82
massa.m	2235	505	18	0.09	0.19	1.82
subrahmanyam.a	95	16	19	0.65	3.66	1.69
chen.j	778	296	20	0.17	0.31	1.63

Table 3.12: Top 20 ranks by weighted PageRank score according to the age-biased YYD link weight scheme (YYD+). The following notations are used for brevity: ranks are denoted by R^{Y+}, while weighted PageRank scores are denoted by PR^{Y+}. Other notations are based on those defined in Table 3.9.

Author keyword	R^{Y+}	τ	λ	φ	*ITI*	k	n_P	n_C	*h**	PR^{Y+} $(\times 10^{-3})$
beck.t	1	11	1	0.44	25	16	26	2143	16	5.49
ang.a	2	9	1	0.56	16	14	17	924	12	5.35
graham.jr	3	15	1	0.56	27	28	28	1651	17	3.97
zhang.l	4	7	1	0.30	23	39	24	422	8	3.97
wurgler.j	5	11	1	0.85	13	9	14	911	10	3.48
o'hara.m	6	14	1	0.58	24	13	25	1105	15	3.05
xing.yh	7	5	1	0.56	9	13	10	282	7	2.76
campello.m	8	9	1	0.56	16	16	17	429	9	2.42
bekaert.g	9	19	1	0.59	32	22	33	2304	22	2.29
lettau.m	10	10	1	1.43	7	7	8	681	6	2.20
acharya.vv	11	11	1	0.50	22	23	23	433	10	2.14
pedersen.lh	12	9	1	1.00	9	6	10	524	9	2.03
kumar.a	13	5	1	0.42	12	11	13	160	5	2.00
sadka.r	14	7	1	0.47	15	11	16	246	9	1.99
lins.kv	15	9	1	0.75	12	17	13	673	11	1.96
almeida.h	16	9	1	0.69	13	13	14	407	9	1.92
yaron.a	17	7	1	1.17	6	12	7	301	6	1.82
massa.m	18	9	1	0.39	23	16	24	262	9	1.82
subrahmanyam.a	19	20	1	0.48	42	25	43	1707	19	1.69
chen.j	20	10	1	0.71	14	20	15	353	4	1.63

Table 3.13: Profile of "rising star" researchers in finance within the top 50 ranks by YYD+ scheme.

Rank	Abbrev. Name	Full Name	PhD (Year Awarded)	Affiliation
1	beck.t	Thorsten Beck	Virginia (1999)	Tilburg
2	ang.a	Andrew Ang	Stanford (1999)	Columbia
4	zhang.l	Lan Zhang	Chicago (2001)	Illinois
4	zhang.l	Lu Zhang	Wharton (2002)	Ohio State
5	wurgler.j	Jeffrey Wurgler	Harvard (1999)	NYU Stern
7	xing.yh	Yuhang Xing	Columbia (2003)	Rice
8	campello.m	Murillo Campello	Illinois (2000)	Cornell
10	lettau.m	Martin Lettau	Princeton (1994)	UC Berkeley
11	acharya.vv	Viral V. Acharya	NYU (2001)	NYU Stern
12	pedersen.lh	Lasse Heje Pedersen	Standford (2001)	NYU Stern
13	kumar.a	Alok Kumar	Cornell (2003)	Miami
14	sadka.r	Ronnie Sadka	Northwestern (2003)	Boston College
15	lins.kv	Karl V. Lins	North Carolina (2000)	Utah
16	almeida.h	Heitor Almeida	Chicago (2000)	Illinois
17	yaron.a	Amir Yaron	Chicago (1994)	Wharton UPenn
18	massa.m	Massimo Massa	Yale (1998)	INSEAD
20	chen.j	Jianguo Chen	Mississippi (1999)	Massey NZ
23	hanlon.m	Michelle Hanlon	Washington (2002)	Michigan
24	rajgopal.s	Shivaram Rajgopal	Iowa (1998)	Emory Goizueta
26	shivakumar.l	Lakshamanan Shivakumar	Vanderbilt (1996)	London Bus. Sch
34	pan.j	Jun Pan	Stanford (2000)	MIT
37	lehavy.r	Reuven Lehavy	Northwestern (1997)	Michigan
38	wu.lr	Liuren Wu	NYU (MPhil 1998)	CUNY
39	brandt.mw	Michael W. Brandt	Chicago (1998)	Duke
41	hail.l	Luzi Hail	Zurich (1996)	Wharton UPenn
43	zhang.y	Zhang Yi	Nebraska (2008)	Texas A & M
45	hennesy.ca	Christopher A. Hennessy	Princeton (2001)	London Bus. Sch.
46	mansi.sa	Sattar A. Mansi	Washington (1999)	Virginia Tech
47	hvidkjaer.s	Soren Hvidkjaer	Cornell (2002)	Copenhagen
48	lowry.m	Michelle Lowry	Rochester (2000)	Penn State
50	yang.j	Jian Yang	Texas A & M (1999)	Colorado
50	yang.j	Jun Yang	Washington (2005)	Indiana

CHAPTER 4

CONCLUSION

In this thesis, I have explored the identification of influential and expert researchers by analysing networks constructed from a large set of publication and citation data. Specifically, I constructed two versions of the same author citation network (to map intellectual flows between researchers) that differs only in the specification of their link weights. The first link weight algorithm, dubbed the Coarse-Grain (CG) scheme, uses only citation cues to determine the (asymmetric) connection strength between researchers. This network was used to generate a list of top 20 highly influential researchers.

In contrast, the second link weight algorithm, dubbed the Yang-Yin-Davison (YYD) scheme, defines the connection strengths using the method outlined in Yang et al. (2011), albeit with a few modifications – I removed the inclusion of co-authorship features since I have found this to introduce a methodological inconsistency. The YYD network was used to generate a list of top 20 experts and/or authorities.

Although some highly central nodes on both networks coincide with award winning researchers, I have found that the YYD scheme produces exactly those influential researchers who are exemplified by long, distinguished, and ever-vibrant careers. The plausibility of a few cases (specifically, the first six ranks on the YYD top 20) were corroborated using online information sources (researcher biodata from faculty webpages and curriculum vitae). Interestingly, both approaches (CG and YYD) generate reasonably good lists of influential researchers even in the absence of information on the reputation or distinctions accorded to such researchers. This suggests that the status of a researcher can be inferred from his/her position within a structure of citation linkages.

I also explored one modification to the YYD link weight scheme i.e. introducing

a bias that boosts the score of productive (as in frequently publishing and active status) researchers with short career time yet are cited by "senior" authorities: such a feature should correspond to a rising star. I have obtained some encouraging results in this respect (as indicated by the appearance of several prize winners), save for several instances of senior authorities creeping into the top 20 ranks. This is one area that I hope to pursue in future works.

In summary, this thesis contributes to existing work in the following manner:

1. The implementation of the YYD method here fixes a dimensional inconsistency within the original formulation, namely, the mixing of author count and citation counts without appropriately matching the dimensions of either quantity (see Equation (2.13)). I have fixed this by removing the coauthor term, since the link weights should be weighted by some measure of citation flow (i.e. influence) rather than co-authorship strength. Furthermore, co-authorship strength is undirected which implies that the coupling between two adjacent nodes i and j is symmetric, to wit, $w_{ij} = w_{ji}$. This is in direct contrast with the concept of a citation flow which is fundamentally directed, implying that in general, the coupling between two adjacent nodes is asymmetric, i.e. $w_{ij} \neq w_{ji}$.

2. I have proposed a method for detecting rising stars in research by introducing a bias that boosts the score of frequently publishing and active status researchers who have a short career timespan, yet are cited by "senior" authorities (see Equation (3.6)).

3. The resulting networks (document citation, co-authorship, and author citation networks) can be used to provide a quantitative survey of some target research field. This was demonstrated on ISI articles within the field of financial economics as shown in Sections 3.1 to 3.4.

As is the case with any mathematical determination, the methods in this thesis are only as good as the assumptions used and the tractability of the issues faced. In this, two notable issues require resolution. First, the selection of parameters (i.e. the citation influence ratio scaling parameters β_1 and β_2 in Equation (2.2), as well as the age-decay exponent ν in Equation (3.6)) is currently determined on a trial-and-error basis. Ideally, a specific value or range of values should be determined analytically rather than by guesswork. Second, I have gauged the "goodness" of the rankings by manually verifying their quality and plausibility over an extremely small subset. The implementation of some suitable objective criteria to measure the performance of the algorithm is left as an interesting and important challenge for future work.

As for assumptions made, I implicitly assume that researchers responsibly cite works they were influenced by. This could be the exception rather than the rule since it may not be practical in most settings to list all possible influences, especially when it can only be done at the expense of conciseness and parsimony. In spite of this, the methods used are able to make sufficient use of available patterns in the data to pick out features otherwise missed by conventional citation analysis (methods based on raw citation counts).

Some further improvements can be made with regard to the implementation of the Yang-Yin-Davison method:

- Include author disambiguation in the data preparation phase in order to identify author keywords that are homonymous (many individuals sharing the same name) and synonymous (individuals who identify themselves with many variations of their name). This is an important extension in order to decrease instances where the influence of a researcher is exaggerated (homonyms) or under-rated (synonyms).

- Account for the contribution of coauthors to each paper. If non-alphabetical ordering is present, author ordering may signal the contribution of coauthors (Moed,

2000). For example, more emphasis (weighting) can be given to the first author, followed by the last author, with middle authors receiving the remainder of the credit. First authors are typically reserved for the author who contributes the most to the publication, while seniority is often conveyed by the last author position (Tscharntke et al., 2007).

- Inclusion of journal weights using either impact factor (Garfield, 1972) or Salancik's structural influence score (Salancik, 1986) when computing the strength of citation flows to account for the importance of the journals. According to Judge et al. (2007), articles published in journals with high citation rates have higher visibility. Since such journals are usually harder to publish in, successful entry may signal the skill of the authors or the relevance of the work. Salancik's method can be easily implemented as demonstrated in Section 3.2.

- Compute separate rankings based on different communities detected from the author citation network (see Figure 3.5). Each community likely corresponds to groups of researchers working within a specific research topic or area and therefore the ranking of authors without accounting for this distinction may under-represent some workers who contribute significantly within (rather than outside of) their respective communities.

- Compute separate rankings based on different topics. Topics can be uncovered by clustering the similarity of article titles and abstracts using a method such as Latent Dirichlet Allocation (Blei, Ng, & Jordan, 2003; Griffiths & Steyvers, 2004). This should allow for a more localised and topic-sensitive ranking of authors.

As a closing remark, it is important to note a conceptual trap in creating scoring methods for the purpose of ranking researchers, especially when such methods are used

as predictors for future success. Such a scheme relies on the assumption that one's importance can be factored down to who they are and their effect upon others. In this, caution must be paid when making a mathematical determination of who is or isn't important as we are often faced with incomplete or imperfect data (knowingly or unknowingly), as well as the possibility that there may exist exceptions to the rule (e.g. exceptional newcomers from obscure parts of the network).

Furthermore, if it is our aim to score a group of people in order to figure out who is more deserving of opportunities, *then by definition, those whom we haven't chosen don't get chosen*[1]. This induces a feedback effect in which not only do the opportunity-rich get richer, but the opportunity-poor get poorer (Merton, 1968; Burt, 1993). Since statistical predictions are made by generalising from the past in order to extrapolate into the future, it is important to keep in mind that the absence of evidence is not evidence of absence. Expectations will hold right up until the point that they don't, and therein lies surprise – eccentric (or anomalous) characters who do not conform to our normal preconceptions of what it takes and means to make an impression. Responsible use of methods like those described in this thesis must always give emphasis to the detection of such outliers. The detection of rising stars as demonstrated in this thesis is one such step in this direction.

[1] As wisely pointed out be the character Larry Fleinhardt in the episode "Sacrifice" in Season 1 of the television series Numb3rs. He explains to the main protaganist of the show, Professor Charles Eppes, on the pitfalls of econometric profiling (on entire neighbourhoods and individuals).

Appendices

APPENDIX A

ALTERNATIVE SCORING METHODS FOR RANKING PAPERS

For most established research fields, the size of the literature is vast and therefore ranking papers by importance (and wherever possible, by relevance) is a practical method to organise the resulting search space. One commonly accepted approach is to count the number of citations received by an article, specifically, the number of citing papers that bibliographically lists a target cited paper as its reference. If citations are mapped as directed links pointing from citing paper j to cited article i, we obtain the citation graph (directed graph or network) $G = (V, E)$, where V is the set of $N = |V|$ nodes (vertices) representing distinct papers, and E is the set of $M = |E|$ directed links (edges) connecting V. In this way, the number of citations to a paper i is simply the in-degree for its corresponding node on G.

While such a measure is straightforward to compute, it ignores two important considerations. First, citation counts do not take into account the function (context) of each citation, i.e., whether it is positive, negative, perfunctory, etc. At best, a citation count measures the popularity of a paper (Redner, 1998) rather than its importance. Second, citations are treated equally regardless of the importance of the citing article. Arguably, a paper's relative importance should increase if cited by many important successive works. Conversely, a paper's relative importance should be diluted if cited by many relatively unimportant works.

Several attempts have been made to address the weaknesses of the citation count. One such approach is the adaptation of webpage ranking methods like Google's PageRank score (Chen et al., 2007; Maslov & Redner, 2008) as well as the Hypertext Induced Topic Search (HITS) authority/hub score (Shimbo, Ito, & Matsumoto, 2007) to rank papers on

a citation network. While both methods are insensitive to the context in which citations are made, it yields alternative rankings that can be used to complement those done via citation count. Furthermore, both HITS and PageRank score nodes based on its position within a structure of ties, thus making it a more appropriate approach to measuring relative prominence than citation count.

Here we demonstrate the usage of the PageRank and HITS algorithm to produce paper rankings using the "BUSINESS, FINANCE" study dataset for the purpose of identifying prominent papers. We then construct two additional scores based on nearest neighbour information of each node. The first, termed the *seminal score*, exploits citation patterns to a target node to measure just how multifaceted its influence is in disparate areas within the literature. The second, termed the *integrative score* exploits referencing patterns of a target node to measure how it bridges previously unconnected works.

PageRank without link weights

The PageRank algorithm forms the basis of Google's massive webpage indexing system (Brin & Page, 1998). Essentially, PageRank models the behaviour of a random surfer visiting one webpage to the next either by walking along directed links between nodes sequentially or restarting the walk at a random node. If $\langle k \rangle$ represents the average number of links the random surfer traverses before jumping (teleporting) to a random node, then successive links are traversed with probability $\alpha = 1 - d$, where d is the damping parameter. Teleportation occurs with probability $1 - \alpha$ which typically occurs after sequentially following $k = 1/(1 - d)$ links on average. The PageRank of paper i is defined as:

$$G_i = \alpha \sum_{j \in \Gamma_{in}(i)} \frac{G_j}{k_j} + \frac{1 - \alpha}{N} \tag{A.1}$$

The first term defines the probability distribution of a random walk from node j to node i with probability $1/k_j$. The second term represents the uniform probability of restarting

the random walk at any node in the network (Chen et al., 2007). The inclusion of this term is necessary since an average user will continuously click-through a small succession of links within a given site before restarting elsewhere when his/her interest is exhausted at that site.

Taken as a whole, the expression in Equation (A.1) can be viewed as a democratic voting process whereby each node distributes their score to other nodes. In this sense, each link from node j to node i *propagates* (contributes) a vote of magnitude G_j/k_j from j to i. The PageRank score $G(i)$ is thus the stationary probability of visiting node i via random walk or teleportation as defined by the scaling parameter α.

There are several important features to PageRank that add to its appeal in measuring the prominence of papers on a citation network. First, the more citations (in-links) to paper (node) i, the larger the sum in the first term of Equation (A.1); hence, the resulting scores should have strong positive correlation with citation count. However, proportionality to citation count alone will not suffice as a robust measure since citation practices vary from one research field to the next, e.g. size of a field, average citation rate, etc. (Maslov & Redner, 2008). This is where the other features of the PageRank algorithm become especially useful.

The second important feature is that citations from prominent papers (as indicated by large PageRank number $G(j)$) contribute more to $G(i)$ than those from less prominent ones. This ensures that older, less-cited papers which play a part in influencing successive prominent works receive an improved (relative) standing compared to that indicated by citation count. Such papers can indeed be considered as *gems* within the literature (Chen et al., 2007).

Third, the contribution of the score from paper j to i, $G(j)$, is diluted the larger the number of references (out-links) of paper j, k_j. This ensures that higher weight is propagated from citing papers that themselves depend on few other references within the liter-

ature. In this sense, PageRank can be used to emphasise works that almost surely shape the direction of successive works. Using the "BUSINESS, FINANCE" study dataset, the top 10 papers are listed in Table A.1.

Table A.1: Top 10 papers by PageRank score $G(i)$ ($\alpha = 0.5$, i.e. $\langle k \rangle = 2$ citation links)

"BUSINESS, FINANCE" dataset				
Paper	Title	Author(s)	Cites	$G(i)$ ($\times 10^{-3}$)
J Financ Econ **8**, 205 (1980)	MEASURING SECURITY PRICE PERFORMANCE	S.J. Brown and J.B. Warner	592	1.894
J Financ Econ **9**, 19 (1981)	MISSPECIFICATION OF CAPITAL-ASSET PRICING - EMPIRICAL ANOMALIES BASED ON EARNINGS YIELDS AND MARKET VALUES	M.R. Reinganum	294	1.695
J Financ Econ **13**, 187 (1984)	CORPORATE FINANCING AND INVESTMENT DECISIONS WHEN FIRMS HAVE INFORMATION THAT INVESTORS DO NOT HAVE	S.C. Myers and N.S. Majluf	1947	1.454
J Financ Econ **9**, 3 (1981)	THE RELATIONSHIP BETWEEN RETURN AND MARKET VALUE OF COMMON-STOCKS	R.W. Banz	603	1.189
J Financ Econ **8**, 3 (1980)	OPTIMAL CAPITAL STRUCTURE UNDER CORPORATE AND PERSONAL TAXATION	H. DeAngelo and R.W. Masulis	322	1.127
J Financ **47**, 427 (1992)	THE CROSS-SECTION OF EXPECTED STOCK RETURNS	E.F. Fama and K.R. French	1561	1.021
J Financ Econ **33**, 3 (1993)	COMMON RISK-FACTORS IN THE RETURNS ON STOCKS AND BONDS	E.F. Fama and K.R. French	2000	1.014
Math Financ **9**, 203 (1999)	Coherent measures of risk	P. Artzner, F. Delbaen, J.M. Eber and D. Heath	1058	0.979
J Monetary Econ **15**, 145 (1985)	THE EQUITY PREMIUM - A PUZZLE	R. Mehra and E.C. Prescott	1103	0.892

Scoring prominence of papers by HITS

The Hyperlink-Induced Topic Search (HITS) algorithm was developed to rank webpages using a hub score y_i and an authority score x_i (Kleinberg, 1999; Ding et al., 2002). It is based on the intuition that a good authority is cited by many good hubs and a good hub cites many good authorities. This circular definition can be expressed in index notation as:

$$x_i = \sum_{j \in \Gamma_{in}(i)} y_j \tag{A.2}$$

$$y_i = \sum_{j \in \Gamma_{out}(i)} x_j \tag{A.3}$$

The main advantage that this algorithm has over PageRank is that it can be used to characterise papers according to two traits. The top 10 "BUSINESS, FINANCE" papers by authority and hub score are as listed in Table A.2 and Table A.3, respectively.

Authority score

Table A.2: Top 10 papers by HITS authority score $A(i)$

"BUSINESS, FINANCE" dataset				
Paper	Title	Author(s)	Cites	$A(i)$
J Financ Econ **33**, 3 (1993)	COMMON RISK-FACTORS IN THE RETURNS ON STOCKS AND BONDS	E.F. Fama and K.R. French	2000	1.0000
J Financ **47**, 427 (1992)	THE CROSS-SECTION OF EXPECTED STOCK RETURNS	E.F. Fama and K.R. French	1561	0.6702
J Financ **48**, 65 (1993)	RETURNS TO BUYING WINNERS AND SELLING LOSERS - IMPLICATIONS FOR STOCK-MARKET EFFICIENCY	N. Jegadeesh and S. Titman	857	0.4610
J Financ **52**, 57 (1997)	On persistence in mutual fund performance	M.M. Carhart	952	0.4373
J Financ **51**, 55 (1996)	Multifactor explanations of asset pricing anomalies	E.F. Fama and K.R. French	679	0.3622
J Financ **49**, 1541 (1994)	CONTRARIAN INVESTMENT, EXTRAPOLATION, AND RISK	J. Lakonishok, A. Shleifer and R.W. Vishny	547	0.3189
J Financ Econ **49**, 283 (1998)	Market efficiency, long-term returns, and behavioral finance	E.F. Fama	546	0.2367
J Financ Econ **43**, 153 (1997)	Industry costs of equity	E.F. Fama and K.R. French	600	0.2301
J Financ **50**, 23 (1995)	THE NEW ISSUES PUZZLE	T. Loughran and J.R. Ritter	479	0.2105
J Financ **40**, 793 (1985)	DOES THE STOCK-MARKET OVERREACT	W.F.M. Debondt and R. Thaler	732	0.2076

Hub score

Table A.3: Top 10 papers by HITS hub score $H(i)$

"BUSINESS, FINANCE" dataset				
Paper	Title	Author(s)	Cites	$H(i)$
J Account Econ **31**, 105 (2001)	Capital markets research in accounting	S.P. Kothari	246	1.0000
J Monetary Econ **49**, 139 (2002)	Investor psychology in capital markets: evidence and policy implications	K. Daniel, D. Hirshleifer and S.H. Teoh	71	0.9617
J Financ **56**, 1533 (2001)	Investor psychology and asset pricing	D. Hirshleifer	231	0.9015
J Financ **55**, 1515 (2000)	Asset pricing at the millennium	J.Y. Campbell	115	0.5999
Eur Financ Manag **14**, 12 (2008)	Behavioural finance: A review and synthesis	A. Subrahmanyam	2	0.5980
J Account Econ **50**, 410 (2010)	Accounting anomalies and fundamental analysis: A review of recent research advances	S. Richardson, I. Tuna and P. Wysocki	4	0.5902
J Corp Financ **16**, 137 (2010)	Share repurchases as a potential tool to mislead investors	K.N. Chan, D.L. Ikenberry, I. Lee and Y.Z. Wang	5	0.4997
J Financ **54**, 1325 (1999)	Conditioning variables and the cross section of stock returns	W.E. Ferson and C.R. Harvey	119	0.4974
Annu Rev Financ Econ **2**, 49 (2010)	Cross-Sectional Asset Pricing Tests	R. Jagannathan, E. Schaumburg and G. Zhou	0	0.4830
Eur Financ Manag **17**, 145 (2011)	The Return of the Size Anomaly: Evidence from the German Stock Market	A. Amel-zadeh	0	0.4736

Seminal papers

These are papers that are cited by communities of cited/citing papers, where these communities share little to no overlap. The intuition behind this is that a "seminal" work has the characteristic that it triggers research activity (typically in a non-trivial way) across multiple and seemingly disparate fields. Given the induced subgraph G' consisting of node i and its nearest incoming-neighbours $\Gamma_{in}(i)$, then the *network constraint* of node i on G is (Burt, 1995):

$$c(i) = \sum_{j \in \Gamma_{in}(i)} \left(p_{ij} + \sum_{q \neq i \neq j} p_{iq} p_{qj} \right)^2 \tag{A.4}$$

where

$$p_{ij} = \frac{(A_{ij} + A_{ji})}{\sum_{k \in \Gamma_{in}(i)} (A_{ik} + A_{ki})} \tag{A.5}$$

The idea here is that a seminal paper i spans a *structural hole* within the literature:

$$\begin{aligned} \text{SeminalScore}(i) &= c(i)^{-1} \\ &= \frac{1}{\sum_{j \in \Gamma_{in}(i)} \left(p_{ij} + \sum_{q \neq i \neq j} p_{iq} p_{qj} \right)^2} \end{aligned} \tag{A.6}$$

This intuition is as depicted in Figure A.1. The top 10 results are listed in Table A.4.

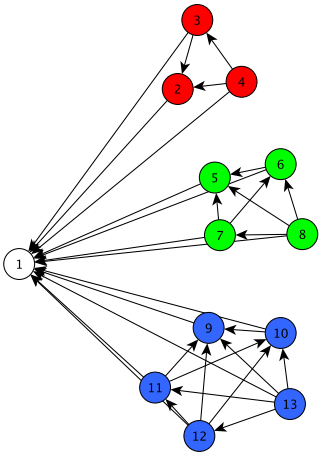

Figure A.1: A seminal paper spans a structural hole in the citation network, i.e., advances work in different groups of densely connected papers (indicated by different colours).

Table A.4: Top 10 papers by seminal score $S(i)$

(b) BUSINESS, FINANCE dataset

Paper	Title	Author(s)	Cites	$S(i)$
J Financ Econ **13**, 187 (1984)	CORPORATE FINANCING AND INVESTMENT DECISIONS WHEN FIRMS HAVE INFORMATION THAT INVESTORS DO NOT HAVE	S.C. Myers and N.S. Majluf	1947	191.95
J Financ Econ **33**, 3 (1993)	COMMON RISK-FACTORS IN THE RETURNS ON STOCKS AND BONDS	E.F. Fama and K.R. French	2000	182.03
Rev Financ Stud **22**, 435 (2009)	Estimating Standard Errors in Finance Panel Data Sets: Comparing Approaches	M.A. Petersen	482	181.56
J Financ **52**, 57 (1997)	On persistence in mutual fund performance	M.M. Carhart	952	158.62
J Financ Econ **43**, 153 (1997)	Industry costs of equity	E.F. Fama and K.R. French	600	149.81
J Financ Econ **14**, 3 (1985)	USING DAILY STOCK RETURNS - THE CASE OF EVENT STUDIES	S.J. Brown and J.B. Warner	945	145.95
J Financ **47**, 427 (1992)	THE CROSS-SECTION OF EXPECTED STOCK RETURNS	E.F. Fama and K.R. French	1561	122.62
J Financ Econ **14**, 71 (1985)	BID, ASK AND TRANSACTION PRICES IN A SPECIALIST MARKET WITH	L.R. Glosten and P.R. Milgrom	785	121.27
J Financ Econ **32**, 263 (1992)	THE INVESTMENT OPPORTUNITY SET AND CORPORATE FINANCING, DIVIDEND, AND COMPENSATION POLICIES	C.W. Smith and R.L. Watts	593	109.08
J Monetary Econ **15**, 145 (1985)	THE EQUITY PREMIUM - A PUZZLE	R. Mehra and E.C. Prescott	1103	104.48

Integrative papers

These are papers that cite works that themselves share little or no overlap. Given that node i has outgoing-neighbours $\Gamma_{out}(i)$ totalling $n = |\Gamma_{out}(i)|$ nodes, we can express the completeness of ties between outgoing-neighbours of i by the local clustering coefficient (Watts & Strogatz, 1998):

$$C(i) = \frac{1}{n(n-1)} \sum_{j \neq k \in \Gamma_{out}(i)} A_{jk} \tag{A.7}$$

The adjacency matrix element A_{jk} encodes the presence ($= 1$) or absence ($= 0$) of a link between node j and k. The value of $C(i)$ is equal to zero when there is zero transitivity among all neighbours, that is, the induced subgraph G' consisting of node i and its nearest outgoing-neighbours forms a star structure. This intuition is as depicted in Figure A.2. Consequently, the integrative score of a paper i can be computed as:

$$\text{IntegrativeScore}(i) = 1 - C(i) \tag{A.8}$$

provided that $0 \leq C(i) \leq 1$. The top 10 results are as listed in Table A.5.

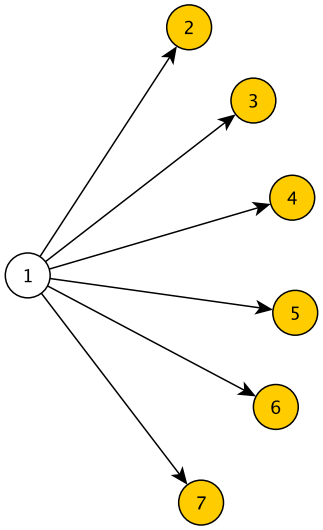

Figure A.2: An integrative paper cites a set of papers that themselves do not cite each other.

Table A.5: Top cited papers by decreasing integrative score $I(i)$. These papers have at least 10 cited references to other ISI papers within the dataset.

"BUSINESS, FINANCE" dataset				
Paper	Title	Author(s)	Cites	$I(i)$
Rev Financ Stud **22**, 435 (2009)	Estimating Standard Errors in Finance Panel Data Sets: Comparing Approaches	M.A. Petersen	482	0.9841
J Financ Econ **32**, 263 (1992)	THE INVESTMENT OPPORTUNITY SET AND CORPORATE FINANCING, DIVIDEND, AND COMPENSATION POLICIES	C.W. Smith and R.L. Watts	593	0.9559
J Financ **52**, 737 (1997)	A survey of corporate governance	A. Shleifer and R.W. Vishny	1372	0.9557
J Financ **48**, 831 (1993)	THE MODERN INDUSTRIAL-REVOLUTION, EXIT, AND THE FAILURE OF INTERNAL CONTROL-SYSTEMS	M.C. Jensen	968	0.9538
J Financ Econ **14**, 3 (1985)	USING DAILY STOCK RETURNS - THE CASE OF EVENT STUDIES	S.J. Brown and J.B. Warner	945	0.9359
J Financ **50**, 23 (1995)	THE NEW ISSUES PUZZLE	T. Loughran and J.R. Ritter	479	0.9307
J Financ Econ **33**, 3 (1993)	COMMON RISK-FACTORS IN THE RETURNS ON STOCKS AND BONDS	E.F. Fama and K.R. French	2000	0.9007
J Financ **48**, 65 (1993)	RETURNS TO BUYING WINNERS AND SELLING LOSERS - IMPLICATIONS FOR STOCK-MARKET EFFICIENCY	N. Jegadeesh and S. Titman	857	0.9000
J Financ **42**, 483 (1987)	A SIMPLE-MODEL OF CAPITAL-MARKET EQUILIBRIUM WITH INCOMPLETE INFORMATION	R.C. Merton	544	0.8939
J Financ Econ **17**, 223 (1986)	ASSET PRICING AND THE BID ASK SPREAD	Y. Amihud and H. Mendelson	564	0.8897

APPENDIX B

PUBLICATIONS

Scientometrics
DOI 10.1007/s11192-010-0254-0

The structure of collaboration in the Journal of Finance

Choong Kwai Fatt · Ephrance Abu Ujum · Kuru Ratnavelu

Received: 27 January 2010

Abstract This paper studies the structure of collaboration in the *Journal of Finance* for the period 1980–2009 using publication data from the *Social Sciences Citation Index* (*SSCI*). There are 3,840 publications within this period, out of which 58% are collaborations. These collaborations form 405 components, with the giant component capturing approximately 54% of total coauthors (it is estimated that the upper limit of distinct *JF* coauthors is 2,536, obtained from the total number of distinct author keywords found within the study period). In comparison, the second largest component has only 13 members. The giant component has mean degree 3 and average distance 8.2. It exhibits power-law scaling with exponent $\alpha = 3.5$ for vertices with degree ≥ 5. Based on the giant component, the degree, closeness and betweenness centralization score, as well as the hubs/authorities score is determined. The findings indicate that the most important vertex on the giant component coincides with Sheridan Titman based on his top ten ranking on all four scores.

Keywords Co-authorship · Collaboration · Network structure

Introduction

A co-authorship network is a mapping of collaborative ties or communication between coauthors within a research community. Two coauthors are connected and assumed to be in communication if they have previously coauthored a paper together. Studies on such social networks provide insight into the social structure of the research community, thus revealing which coauthors are central to communication processes on the network. The first empirical studies on social networks were documented in Milgram (1967). The earliest documented study on co-authorship networks can perhaps be traced back to the

C. K. Fatt (✉)
Faculty of Business and Accountancy, University of Malaya, 50603 Kuala Lumpur, Malaysia
e-mail: kwaifatt@um.edu.my

E. A. Ujum · K. Ratnavelu
Institute of Mathematical Sciences, University of Malaya, 50603 Kuala Lumpur, Malaysia

Published online: 10 June 2010

mathematics community in 1969 through the concept of Erdös number, i.e. collaboration distance to the late famous mathematician, Paul Erdös (Grossman 1996; Hoffman 1998). Decades later, Newman constructed and studied co-authorship networks based on papers published in MEDLINE, the Los Alamos Preprint Archive, SPIRES and NCSTRL (Newman 2001a, b, c). This work was then extended in Newman (2004a) and Newman (2004b). A similar study on the RePEc (Research Papers in Economics) database was covered in Krichel and Bakkalbasi (2006). There are several theories regarding how co-authorship networks are structured and formed (Barabási et al. 1999, 2002; Pennock et al. 2002). Since then, empirical studies to test these theories have been conducted primarily in the sciences. This paper is an attempt to fill the gap for the field of finance. The *Journal of Finance* is used as a case study as it is one of the core journals in financial research (Borokhovich et al. 1994).

Data and methodology

Publication data for the *Journal of Finance* (*JF*) was sourced from the *Social Sciences Citation Index* (*SSCI*) database. This consists of a total of 3480 publications within 1980–2009 corresponding to 3,082 distinct author keywords. Roughly 42% of the total publications are single authorship papers. The remaining 58% are collaborations: 1,365 dual authorship papers, 568 triple authorship papers, seventy-three 4-author papers, six 5-author papers and one 7-author paper. The single authorship papers are published under 1,050 distinct author keywords, 546 (approximately 52%) of which do not appear in any of the collaboration papers. This implies that a maximum of 546 authors contributing to *JF* are not connected to any of the other contributing authors within the study period. The mean number of coauthors is 1.79 ± 0.80 (median $= 2.00$). Co-authorship ties were deduced from the author field of the *SSCI* data. The binary network model (Krichel and Bakkalbasi 2006) was then used to map co-authorship ties between *JF* researchers (represented by vertices on the network with collaborative ties between them marked by edges). Two researchers who have co-authored a paper in the past are connected by an edge with collaboration weight of one to signify the existence of co-authorship. Pairs of researchers who have no history of collaboration throughout the study period are assigned a collaboration weight of zero to indicate that they are unconnected on the *JF* co-authorship network. All calculations on the resulting network were first carried out using the network analysis program Pajek (Batagelj and Mrvar 1998), and then re-computed using the *igraph* package (version 0.5.3) for the GNU R statistical environment (Csardi and Nepusz 2006).

Results

By limiting the focus to collaborating authors within the *Journal of Finance*, a network of 2,538 coauthors connected by 3,038 collaborative ties can be constructed. This network is fragmented into 405 components with mean degree of 2.4. In all social networks, there exists the possibility of a percolation transition (Barabási et al. 1999). In networks with very small number of connections, all individuals belong to small, isolated components (no path exists to connect one component to the next). However, as the total number of connections increases, there comes a point at which a giant component forms—a large group of individuals who are all connected to one another by paths of intermediate acquaintances. Newman (2001a) reported that the collaboration networks for MEDLINE,

Los Alamos Preprint Archive, SPIRES and NCSTRL possess giant components that capture roughly 80–90 percent of all authors: almost everyone in the community is connected to almost everyone else by some path of intermediate coauthors. Furthermore, Krichel and Bakkalbasi (2006) reported that the giant component of the RePEc network encompasses 83% of its total authors. The present work finds that the giant component for the *JF* co-authorship network (Fig. 1a) captures only 54% (1,362 vertices) of its total coauthors. Thus, the *JF* network is quite fragmented in comparison to the networks previously studied by Newman or Kirchel and Bakkalbasi. It must be pointed out however that these electronic database networks are more extensive since they were constructed from publication data sourced from a large number of journals.

The giant component may signify the core of mainstream research activity (other components may be specialized clusters or sub-communities). This is the case if the network growth mechanism is governed by a cumulative advantage process (Simon 1955; Price 1976)—also known in the literature as preferential attachment (Barabási et al. 1999) or rich-get-richer process—that is, coauthors with many collaborations in the past, tend to gain more collaborations in the future. The signature of such a process in network structures is the existence of a power-law or heavy tailed degree distribution (e.g. Yule-Simon distribution). Such a degree distribution is found on the giant component of the *JF* network (Fig. 1b) i.e. the tail of the cumulative degree distribution can be approximated by a power law with exponent $\alpha = 3.5$ for vertices with degree ≥ 5 (Kolmogorov–Smirnov D-statistic $= 0.0423$). The mean degree on the giant component is 3 (median $= 2$).

The diameter of a network is given as the maximum separation between the pairs of authors on the network. For the giant component, the most distant pair of vertices is ANGEL, JJ and PIRIE, WL separated by 21 edges (red path in Fig. 1a). The average distance between coauthors in the giant component is 8.2. In comparison, the average distance is 4.4 for MEDLINE, 4.0 for SPIRES, 9.7 for NCSTRL and 5.9 in Los Alamos Preprint Archive (Newman 2001a). The average distance gives a measure of the "connectedness" of the network (Kretschmer 2004). Small distances give rise to what is called the "small world effect" on networks, whereby it is possible to connect any two strangers

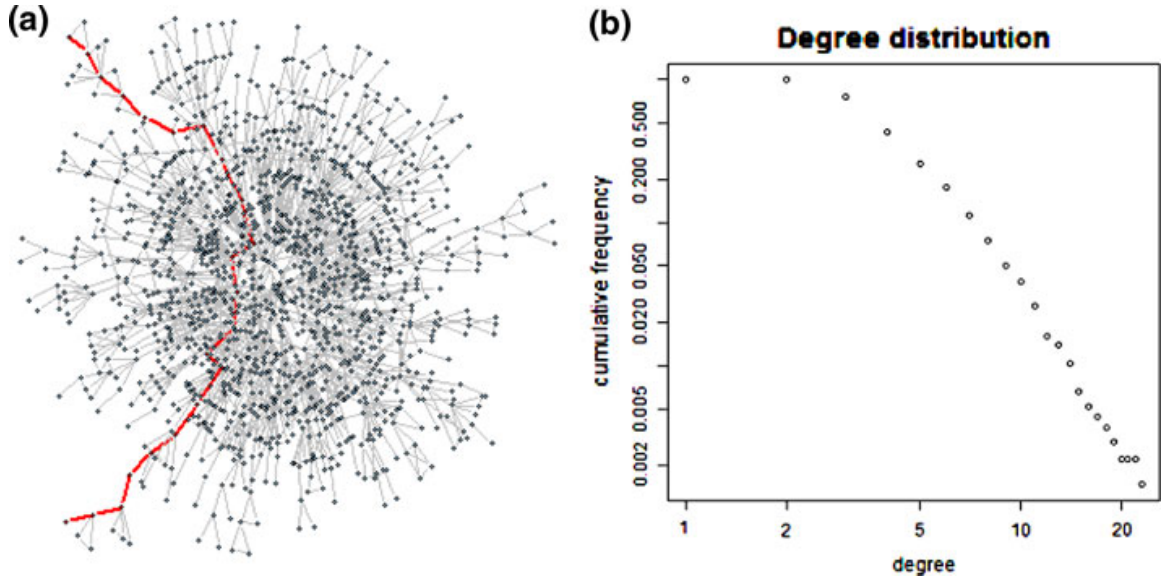

Fig. 1 **a** Giant component in *Journal of Finance* co-authorship network (1980–2009). It consists of 1,362 vertices connected by 2,044 edges. The *black* path marks the diameter of the network. **b** Cumulative degree distribution for the giant component

in the network through a small number of intermediate acquaintances. In theory, a small world community allows people to coordinate their actions towards mutually beneficial goals.

As observed from the *JF* network, some coauthors are positioned at the core while others are at the periphery of the network. In either case, coauthors that are central to the network can be identified in the sense that they connect different parts of the network together (Lu and Feng 2009). As is typically the case with social networks, one shall find that some coauthors are more central than others. The present work covers four measures that quantify centrality in this paper: degree, closeness, betweenness and hubs/authorities score. Degree centrality of a vertex is measured by the number of edges connected to it (see Appendix I). A coauthor with high degree centrality is directly connected to many coauthors, from which he/she can (presumably) pool useful knowledge or skill sets. The more collaborators one has, the larger the pool of knowledge and skill sets that he/she can directly tap into. According to (de Nooy et al. 2004), "[the] degree centralization of a network is the variation in the degrees of vertices divided by the maximum degree variation which is possible in a [star] network of the same size". The degree centralization for the giant component has an arithmetic mean of 0.0022053 and median 0.0014695, with values ranging between 0.0007348 and 0.0161646. The top 10 coauthors with the highest degree centrality and degree centralization score are presented in Table 1. The highest ranked coauthors based on this measure are Josef Lakonishok (lakonishok_j) of the College of Business, University of Illinois and Sheridan Titman (titman_s) of the Graduate School of Business, University of Texas at Austin. The two have co-authored *JF* papers with a

Table 1 Top 10 ranked coauthors by degree centralization

Degree rank	Author keyword	Degree	Degree centralization
1	titman_s	22	0.0161645849
1	lakonishok_j	22	0.0161645849
2	mcconnell_jj	21	0.0154298310
3	michaely_r	18	0.0132255694
4	john_k	17	0.0124908156
5	longstaff_fa	16	0.0117560617
6	stulz_rm	15	0.0110213079
7	schwartz_es	14	0.0102865540
7	travlos_ng	14	0.0102865540
8	shleifer_a	13	0.0095518001
8	lang_lhp	13	0.0095518001
8	ross_sa	13	0.0095518001
8	saunders_a	13	0.0095518001
8	thakor_av	13	0.0095518001
9	brennan_mj	12	0.0088170463
9	whaley_re	12	0.0088170463
9	megginson_wl	12	0.0088170463
9	dumas_b	12	0.0088170463
9	starks_lt	12	0.0088170463
10	hirshleifer_d	11	0.0080822924
10	lee_cmc	11	0.0080822924
10	senbet_lw	11	0.0080822924

Table 2 Top 10 ranked coauthors by closeness centralization

Closeness centralization rank	Author keyword	Closeness centralization
1	schwartz_es	0.191798196
2	longstaff_fa	0.187853692
3	mcconnell_jj	0.183596385
4	subrahmanyam_a	0.183447904
5	chan_kc	0.182856375
6	titman_s	0.182000535
7	brennan_mj	0.180121758
8	roll_r	0.179646251
9	grinblatt_m	0.178749672
10	karolyi_ga	0.177955021

total of 22 collaborators. To date, Lakonishok and Titman have not co-authored a *JF* paper together but are indirectly connected by Narasimhan Jegadeesh (jegadeesh_n) of the Goizueta Business School, Emory University (previously affiliated with University of Illinois and UCLA).

Within the production of a joint publication, collaborators may function as social resources to each other by catalyzing the formation of new collaborative ties—they may directly know someone, or know someone who knows someone (and so on) with the crucial knowledge or skill set to complete the project. This notion is more accurately captured by closeness centrality whereby a person in a social network is considered more central if on average, he/she is reachable from everyone else in the network through a short chain of acquaintances. The closeness centrality of a vertex is based on the total distance between one vertex and all other vertices, where larger distances yield lower closeness centrality scores (see Appendix I). The closer a vertex is to all other vertices, the easier information may reach it, the higher its centrality. According to (de de Nooy et al. 2004), "[the] closeness centralization is the variation in the closeness centrality of vertices divided by the maximum variation in closeness centrality scores possible in a [star] network of the same size". The closeness centralization for the giant component has an arithmetic mean of 0.125236 and median 0.1259719, with values ranging between 0.0709224 and 0.1917982. The top 10 coauthors with the highest closeness centralization scores are presented in Table 2. The highest ranked coauthor according to this measure is Eduardo Schwartz (schwartz_es) of the UCLA Anderson School of Management. The data suggests that Schwartz has the shortest average distance from any other coauthor in the network.

Similarly, one may also consider network centrality in terms of who frequently plays the role of "go-between", i.e. vertices that frequently mediate the transfer of information on the network (Rousseau and Zhang 2008). In the context of this paper, the more likely a coauthor appears on geodesics (shortest path on the network between any pair of vertices), the higher the "betweenness" centrality (see *Appendix I*). Here it is assumed that geodesics are the optimal channels of communication used between any pair of coauthors. According to (de Nooy et al. 2004), "The betweenness centrality of a vertex is the proportion of all geodesics between pairs of other vertices that include this vertex. Betweenness centralization is the variation in the betweenness centrality of vertices divided by the maximum variation in betweenness centrality scores possible in a star network of the same size". A coauthor with high betweenness centrality is crucial to the flow of information on the

Table 3 Top 10 ranked coauthors by betweenness centralization

Betweenness rank	Author keyword	Betweenness centralization
1	mcconnell_jj	0.155275774
2	schwartz_es	0.145687965
3	chan_kc	0.125473581
4	titman_s	0.117428506
5	brennan_mj	0.116206381
6	senbet_lw	0.113209806
7	michaely_r	0.108469312
8	longstaff_fa	0.106276877
9	lakonishok_j	0.095553181
10	jegadeesh_n	0.083703176

co-authorship network. The betweenness centralization for the giant component has an arithmetic mean of 0.0053202 with values ranging between 0 (which correspond to dangling vertices at the periphery of the giant component) and 0.1552758. A list of the Top 10 scorers in terms of this measure is presented in Table 3. The highest ranked coauthor according to this measure is John J. McConnell (mcconnell_jj) of the Krannert School of Management, Purdue University. The data suggests that McConnell is in a good position to play the role of intermediary for most coauthors in the *JF* co-authorship network.

Hub and authorities weight is based on the Hyperlink-Induced Topic Search (HITS) algorithm (see *Appendix I*). Similar to Google's PageRank (Brin and Page 1998), it is an iterative link analysis algorithm based on eigenvector centrality (Bonacich 1972)—the centrality of a vertex is formulated as a linear combination of scores of other vertices (Correa et al. 2009). For social networks, HITS allows us to gauge the importance (prestige) of a person by the importance of the company they keep. For undirected networks, the hub and authorities scores are nearly identical (Shafer et al. 2006). The hubs score for the giant component has an arithmetic mean of 0.003581 and median 0.0017400, with values ranging between 0 (for dangling vertices) and 0.3595926. The Top 10 ranked coauthors in terms of this score is presented in Table 4.

Only Sheridan Titman (titman_s), from the University of Texas appears in the Top 10 rank of all four measures considered. He has the highest hubs/authorities score and degree centrality rank, stands at 4th highest ranked author for closeness centralization and 6th in betweenness centralization. The numbers suggest that Titman is the central hub to the *Journal of Finance* co-authorship network with 22 collaborators throughout his publication history under *JF*. The high closeness centralization score suggests that Titman is one of the most closely connected coauthors in the network (separated by a small distance to other *JF* coauthors, on average). The high betweenness centralization score suggests that Titman is in a good position to influence the flow of information on that network. Incidentally, Titman along with Kent Daniel (i.e. daniel_k; ranked 6th according to hubs/authorities score) from Northwestern University received the 1997 Smith-Breeden First Prize for their paper entitled "Evidence on the Characteristics of Cross Sectional Variation in Stock Returns". This award is given out annually to the top three research papers published in *Journal of Finance*. Daniel has also been awarded the 1999 Smith-Breeden First Prize along with David Hirshleifer (hirshleifer_d; ranked 10th in degree centralization, and 7th in hubs/authorities score) from University of California, Irvine and Avanidhar Subrahmanyam (subrahmanyam_a; ranked 4th in closeness centralization and hubs/authorities score)

Table 4 Top 10 ranked coauthors by hub and authorities weight (computed with Pajek)

Hub rank	Author keyword	Hub score	Authority score	ISI-HC
1	titman_s[1997]	0.359592609	0.359573124	Yes
2	lakonishok_j[1995]	0.278483387	0.278459457	Yes
3	jegadeesh_n	0.208685496	0.208672459	–
4	subrahmanyam_a[1999]	0.207324975	0.207319259	–
5	grinblatt_m[2001]	0.187684851	0.187677954	–
6	daniel_k[1997,1999]	0.184238279	0.184229344	–
7	hirshleifer_d[1999]	0.178694679	0.178685814	–
8	shleifer_a[1995,1999]	0.172889075	0.172881002	Yes
9	schwartz_es	0.155392638	0.155398207	Yes
10	longstaff_fa	0.151856337	0.151866768	–

The ISI-HC column indicates whether the coauthor is listed under the Economics/Business category of ISIHighlyCited.com. The superscript indicates the year(s) in which that coauthor received the Smith-Breeden prize

from University of California, Los Angeles for their joint work entitled "Investor Psychology and Security Market Under- and Overreaction".

The 1995 Smith-Breeden Distinguished Paper prize was also awarded to Josef Lakonishok, Andrei Shleifer and Robert W. Vishny for their paper entitled "Contrarian Investment, Extrapolation and Risk". Since Josef Lakonishok (lakonishok_j) ranks first in degree centrality (with Sheridan Titman), 9th in betweenness rank and 2nd in hubs/authorities score, this suggests that Lakonishok is a structurally important vertex in relation with other individuals on the *Journal of Finance* co-authorship network. According to Table 4, Andrei Shleifer, i.e. shleifer_a, ranks 8th in degree centrality as well as hubs/authorities score. Like Kent Daniel, Shleifer was awarded once more in 1999 along with Rafael La Porta of Dartmouth College and Florencio Lopez-de-Silanes of University of Amsterdam for their joint work entitled "Corporate Ownership Around the World". As of March 2010, RePEc lists Shleifer as the 2nd highest ranked economist in the world, after the 2001 Nobel Memorial Prize in Economic Sciences recipient, Joseph Stiglitz (http://ideas.repec.org/top/top.person.all.html).

Lastly, Mark Grinblatt (grinblatt_m) of University of California, Los Angeles received the 2001 Smith-Breeden Distinguished Paper award along with Matti Keloharju of Helsinki School of Economics for their paper entitled "What Makes Investors Trade?"

Concluding remarks

Two important idealizations were made in the construction of the network studied in this paper. Firstly, vertices can enter the network at any time but once they do, they never exit (this corresponds to a pure birth process). This is unrealistic in the sense that the individuals who make up the network have finite lifetimes, beyond which communication is no longer possible. For the *JF* giant component, 638 of its 1,362 members have not published in *JF* for the past ten years. Such cases are dubbed "ghost vertices" to reflect the ambiguity or uncertainty tied to their structural presence. However, since the *JF* network appears to exhibit scale-free behavior, the overall network structure should be robust under the

removal of these ghost vertices unless they coincide with major hubs (Albert et al. 2000) or possess high betweenness centrality (connect different sub-communities on the network). On that note, only senbet_lw (ranked sixth in betweenness centrality), has no publications beyond 1996. Perhaps, a more refined approach is to study the *JF* network using a 1-year sliding window across the same study period. This should enable one to accurately resolve temporal variations in the communication links between coauthors.

The second idealization was made through the usage of the binary network model, which by construction, assigns equal strength to collaborative ties between connected pairs of authors thus obscuring strong and weak ties between them. Intuition suggests that strong ties are evidenced by more frequent collaboration, while weak ties can be attributed to collaborations that occur only once or occasionally (Krichel and Bakkalbasi 2006). We can account for this by using multiple edges to signify multiple collaboration events between two coauthors. If coauthor *A* and *B* have collaborated twice in the past, then we connect the two by two undirected edges instead of one. This construction directly affects the degree distribution and hence affects the resulting ranking of coauthors by degree centrality as well as hubs/authorities score. In the case of the *Journal of Finance*, the difference is quite remarkable. For degree centrality, the top three positions are occupied by Andrei Shleifer, Josef Lakonishok and Sheridan Titman with 38, 35 and 34 total collaborations respectively. For hubs/authorities score, the resulting ranking is as listed in Table 5. It is interesting to see that the analysis of the *JF* co-authorship network with multiple edges picks out more entries on ISIHighlyCited.com than the single edges case.

Another reason for concern is that the boundary of the network is artificial: the network studied in this paper is only a partial mapping of collaborative ties within the finance research community. Two coauthors that have previously collaborated in the *Journal of*

Table 5 Comparison between top 10 ranked coauthors by hub and authorities weight for single edges case and multiple edges case (computed with *igraph*; score values differ with Pajek by a constant multiple ~ 0.36)

Single edges				Multiple edges			
Rank	Coauthor	Hubs/authorities score	ISI-HC	Rank	Coauthor	Hubs/authorities score	ISI-HC
1	titman_s[1997]	1.000000000	Yes	1	shleifer_a[1995,1999]	1.000000000	Yes
2	lakonishok_j[1995]	0.774777301	Yes	2	vishny_rw[1995]	0.788534523	Yes
3	jegadeesh_n	0.580390260	–	3	la_porta_r[1999]	0.773629109	Yes
4	subrahmanyam_a[1999]	0.576311248	–	4	lopez-de-silanes_f[1999]	0.689170697	Yes
5	grinblatt_m[2001]	0.521787585	–	5	lakonishok_j[1995]	0.312280383	Yes
6	daniel_k[1997,1999]	0.512307556	–	6	summers_lh	0.134852615	Yes
7	hirshleifer_d[1999]	0.496902673	–	7	delong_jb	0.134509363	Yes
8	shleifer_a[1995,1999]	0.480683506	Yes	7	waldmann_rj	0.134509363	–
9	schwartz_es	0.431511718	Yes	8	lee_cmc	0.128694841	–
10	longstaff_fa	0.421464429	–	9	thaler_rh	0.127412919	Yes
–	–	–	–	10	chan_lkc	0.104625474	–

The ISI-HC column indicates whether the coauthor is listed under the Economics/Business category of ISIHighlyCited.com. The superscript indicates the year(s) in which that coauthor received the Smith-Breeden prize

Finance may have collaborated elsewhere in the past, or may choose to exclusively collaborate elsewhere in the future (e.g. *Journal of Financial Economics* or *Review of Financial Studies*). It is also possible that authors without collaborations in *JF* may actually have a history of collaboration in other journals. This affects the degree distribution, as well as the distance and centrality measures. In order to get a full mapping, one would need to extend the source data to include journals outside of the *Journal of Finance*.

In summary, the present work used network centrality measures (degree, closeness and betweenness centralization score as well as hubs/authority score) to find the most structurally important vertices on the co-authorship network of the *Journal of Finance* covering the period 1980–2009. These important vertices coincide with key players within the co-authorship network. It is assumed that co-authorship networks are communication networks where the members (coauthors) tap into the knowledge and expertise of their nearest neighbors or their nearest neighbors' neighbors. In this context, "key" players refer to coauthors that are crucial to communication or information flow on the *JF* co-authorship network. As a closing remark, the authors emphasize that this work is not intended to rank financial researchers by their importance; rather, the authors draw interest in the fact that there are important coauthors, and that their proportion is consistent with predictions from the scale-free model (assuming one can neglect vertices on the lower end of the degree distribution). However, everyday experience tells us that simply forming connections is not the whole story; a certain amount of maintenance is required to manage strong and weak ties (i.e. social ties vary in quality and are not symmetrical in general, as opposed to what was implicitly assumed in this paper). In order to deduce the mechanisms responsible for the fine structure it seems that we need to know more about the coauthors beyond their structural contribution to the network. In this respect, knowing the identity of structurally important coauthors could provide useful clues in that direction. How this information can be encapsulated into a working model is left as a challenge for future works.

Appendix I

Consider a graph $G = (V, E)$ where E is the set of edges connecting vertices defined in vertex set V. The construction of a binary network model (Krichel and Bakkalbasi 2006) based on G requires that each $e_{ij} \in E$ encodes the presence or absence of a connection between vertex i and j. For the case of a directed graph: we set the edge weight $e_{ij} = 1$ if a link exists from vertex i to j, and $e_{ij} = 0$ if i and j are unconnected ($i \neq j$). For the case of an undirected graph: $e_{ij} = e_{ji} = 1$ if vertex i and j are connected ($i \neq j$), while $e_{ij} = e_{ji} = 0$ if unconnected. For both directed and undirected graphs, we set $e_{ii} = 0$ so that G does not contain any loops.

Degree centrality

The degree centrality of vertex v is simply given by the number of edges incident upon it. Suppose that there are n vertices in vertex set V, then the degree centralization is defined by the following formula (Freeman 1979):

$$C_D(v) = \frac{deg(v)}{n-1}, \text{ where } deg(v) = \text{ degree of vertex } v. \quad (1)$$

Closeness centrality

The closeness centrality of vertex v is defined as the average number of steps required to reach every other reachable vertex in the graph. Specifically, it is the inverse of the mean geodesic distance (length of shortest paths) to/from all the other vertices in the graph, as defined by the following formula (Freeman 1979):

$$C_C(v) = \frac{n-1}{\sum_{i \neq j} d(i,j)}, \tag{2}$$

where $d(i, j) =$ distance between vertex i and j.

Betweenness centrality

The betweenness centrality of vertex v is defined as the number of geodesics (shortest paths) on the graph that pass through it. Its value can be computed by the following formula (Freeman 1979):

$$C_B(v) = \sum_{\substack{i \neq v \neq j \in V \\ i \neq j}} \frac{\sigma_{ij}(v)}{\sigma_{ij}}, \tag{3}$$

where $\sigma_{ij}(v)$ is the number of shortest paths from vertex i to j that pass through v, while σ_{ij} is the number of shortest paths from vertex i to j. The betweenness centralization is given by the betweenness centrality divided by $(n-1)(n-2)$ for directed graphs and $\frac{1}{2}(n-1)(n-2)$ for undirected graphs.

HITS algorithm: hubs/authorities score

Hyperlink-Induced Topic Search, or HITS (Kleinberg 1998), is a link analysis algorithm originally designed to rank webpages by using the method of eigenvector centrality (Bonacich 1972). HITS assigns two scores to each vertex on graph G: a hub score y_i and an authority scorex_i. The underlying logic behind the method is that a good authority is cited by many good hubs, while a good hub cites many good authorities. This mutual reinforcement between authority and hub vertices can be represented by two operations I and O. The I operation updates the x-weights (authorities score) as follows.

$$x_i \leftarrow \sum_{j:(j,i) \in E} y_j \tag{4}$$

The O operation updates the y-weights (hubs score) as follows.

$$y_i \leftarrow \sum_{j:(i,j) \in E} x_j \tag{5}$$

In matrix representation, these two operations can be written succinctly as:

$$I(\cdot) = L^T, O(\cdot) = L. \tag{6}$$

By recursively updating the x- and y-weights, the authority and hub scores of each vertex eventually converge at their final values. At the tth iteration, we obtain the following expressions:

$$\begin{aligned} x^{(t+1)} &= I\left(O\left(x^{(t)}\right)\right) = L^T L x^{(t)} \\ y^{(t+1)} &= O\left(I\left(y^{(t)}\right)\right) = L L^T y^{(t)} \end{aligned} \quad (7)$$

The final solutions x^*, y^* are the principal eigenvectors of $L^T L$ (authority matrix) and LL^T (hub matrix), which are the singular decomposition of L (Ding et al. 2002). For undirected graphs, L is symmetric and therefore $L^T L = LL^T = L^2$ (Shafer et al. 2006).

References

Albert, R., Jeong, H., & Barabási, A.-L. (2000). Error and attack tolerance of complex networks. *Nature, 406*, 378–482.

Barabási, A.-L., Albert, R., & Jeong, H. (1999). Mean-field theory for scale-free random networks. *Physica A, 272*, 173–187.

Barabási, A. L., Jeong, H., Neda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A, 311*, 590–614.

Batagelj, V., & Mrvar, A. (1998). Pajek—Program for large network analysis. *Connections, 21*, 47–57.

Bonacich, P. (1972). Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology, 2*, 113–120.

Borokhovich, K. A., Bricker, R. J., & Simkins, B. J. (1994). Journal communication and influence in financial research. *The Journal of Finance, 49*, 713–725.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems, 30*, 107–117.

Correa, C., Crnovrsanin, T., Kwan, L. M., Keeton, K. (2009). *The derivatives of centrality and their applications in visualizing social networks*. Technical Report, UC Davis Department of Computer Science. Available online: http://www.cs.ucdavis.edu/research/tech-reports/2009/CSE-2009-5.pdf.

Csardi, G., Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal Complex Systems* 1695. http://igraph.sf.net.

de Nooy, W., Mrvar, A., & Batagelj, V. (2004). *Exploratory network analysis with Pajek.* Cambridge: Cambridge University Press.

Ding, C., Xiaofeng, H., Husbands, P., Hongyuan, Z., & Simon, H. D. (2002). PageRank, HITS and a unified framework for link analysis. In *Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval*, 11–15 Aug 2002, Tampere, Finland.

Freeman, L. C. (1979). Centrality in social networks I: Conceptual clarification. *Social Networks, 1*, 215–239.

Grossman, J. W. (1996). The Erdős Number Project. http://www.oakland.edu/enp/.

Hoffman, P. (1998). *The man who loved only numbers: The story of Paul Erdős and the search for mathematical truth*. New York: Hyperion.

Kleinberg, J. (1998). Authoritative sources in a hyperlinked environment. In *Proceedings of the 9th ACM-SIAM symposium on discrete Algorithms*.

Kretschmer, H. (2004). Author productivity and geodesic distance in bibliographic co-authorship networks, and visibility on the Web. *Scientometrics, 60*, 409–420.

Krichel, T., & Bakkalbasi, N. (2006), A social network analysis of research collaboration in the economics community. In *International workshop on webometrics, informetrics and scientometrics & seventh COLLNET meeting, 10–12 May 2006, Nancy, France*.

Lu, H., & Feng, Y. (2009). A measure of authors' centrality in co-authorship networks based on the distribution of collaborative relationships. *Scientometrics, 81*, 499–511.

Milgram, S. (1967). The small world problem. *Psychology Today, 2*, 60–67.

Newman, M. E. J. (2001a). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences USA, 98*, 404–409.

Newman, M. E. J. (2001b). Scientific collaboration networks: I. Network construction and fundamental results. *Physical Review E, 64*, 016131.

Newman, M. E. J. (2001c). Scientific collaboration networks: II. Shortest paths, weighted networks, and centrality. *Physical Review E, 64*, 016132.

Newman, M. E. J. (2004a). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences USA, 101*, 5200–5205.

Newman, M. E. J. (2004b). Who is the best connected scientist? A study of scientific coauthorship networks. In E. Ben-Naim, H. Frauenfelder, & Z. Toroczkai (Eds.), *Complex networks* (pp. 337–370). Berlin: Springer.

Pennock, D., Flake, G. W., Lawrence, S., Glover, E. J., & Giles, C. L. (2002). Winners don't take all: Characterizing the competition for links on the web. *Proceedings of the National Academy of Sciences USA, 99*, 5207–5211.

Price, D. J. D. (1976). General theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science, 27*, 292–306.

Rousseau, R., & Zhang, L. (2008). Betweenness centrality and Q-measures in directed valued networks. *Scientometrics, 75*, 575–590.

Shafer, P., Isganitis, T., & Yona, G. (2006). Hubs of knowledge: Using the functional link structure in Biozon to mine for biologically significant entities. *BMC Bioinformatics, 7*, 71.

Simon, H. A. (1955). On a class of skew distribution functions. *Biometrika, 42*, 425–440.

Inter-journal Citations and Ranking Top Journals in Financial Research

Kuru Ratnavelu[1], Choong Kwai Fatt[2], and Ephrance Abu Ujum[1]

[1]Institute of Mathematical Sciences,
University of Malaya, Kuala Lumpur 50603, Malaysia
[2]Faculty of Business and Accountancy,
University of Malaya, Kuala Lumpur 50603, Malaysia

Abstract. This article uses the methodology developed by Borokhovich, Bricker and Simkins (1994) to determine the relative influence of seven prominent finance journals. The original analysis is expanded on a longitudinal basis for the years 1990 to 2006 inclusive. It is found that the relative influence rank produces some stable ordering over the study period with the Journal of Finance and Journal of Financial Economics occupying top spots. A change in the ordering of the relative influence rank indicates a shift in inter-journal communication trends.

Keywords: Bibliometrics, Finance literature, Interjournal communication, Influence ranking.

1 Introduction

There are a number of accepted measures to rank and measure the quality of journals. Nevertheless, the challenge of measuring journal influence is fraught with pitfalls. The common convention is to use the Journal Citation Reports (JCR) impact factor score which measures a journal's average number of citations per article within a specific time window (Garfield [1]; Adler, Ewing and Taylor [2]). It has become the convention to compute the impact factor over a period of two (or five) years. Shorter time windows give greater weight to rapidly changing fields. On the other hand, longer time windows take into account a larger number of citations and/or sources, but results in a less current measure of impact. This measure has received considerable attention, most notably because citation rates vary from one field to the next and therefore a standardized two-year time window across all fields may exaggerate the impact of some journals (especially multidisciplinary ones) while under estimating others. Clearly, a field specific treatment is required.

In 1994, Borokhovich, Bricker and Simkin [3] (hereon referred to as BBS) had presented a case study of inter-journal communication and influence between eight of the most prominent mainstream journals in finance during 1990-1991. The eight journals are: Financial Management (FM), Journal of Banking and Finance (JBF),

P. Balasubramaniam and R. Uthayakumar (Eds.): ICMMSC 2012, CCIS 283, pp. 505–513, 2012.

Journal of Business (JBUS), Journal of Finance (JF), Journal of Financial Economics (JFE), Journal of Financial Research (JFR), Journal of Financial and Quantitative Analysis (JFQA), and the Review of Financial Studies (RFS). In their analysis, the concept of "self-citation index" was developed, defined as a measure of how frequently a journal cites itself compared to how frequently its articles are cited by other journals. Self-citations are instances where an article cites another article published within the same journal. According to BBS, there are a number of possible explanations for differences in self-citation rates across journals. On one hand, it might be argued that a journal, in order to promote itself, encourages self-citations. Then, differences in these self-citation rates reflect the extent of self-promotion. Alternatively, journals that more frequently publish important studies tend to be cited more frequently. In this case, the differences in self-citation rates reflect the relative importance of the journals' articles. Finally, it is reasonable to assert that journals publishing in narrower or more specialized research areas tend to cite themselves more frequently, simply because they are the principal source of knowledge in that area."

Following the BBS approach, this paper uses synchronous citation data to explore the inter-journal citation patterns between journals, using the field of financial research as a case study. We believe that this study will provide a further insight into the inter-journal communication and influence on the use of a larger set of data (1990-2006) from these core journals. We organize this paper in a similar manner: Section 1 describes the data used in this study, while Section 2 covers the analysis of inter-journal citations. We begin Section 2 by first reconstructing Table 1 in BBS to benchmark the results of our methods. We then analyze the time series for self-citation rates and self-citation index.

2 Data and Methodology

We select the following seven of the eight mainstream finance journals, as identified by BBS: FM, JBF, JBUS, JF, JFE, JFQA, and RFS. Publication and citation data for the eight journals were obtained from the Social Science Citation Index (SSCI), covering the period 1990-2006. The JFR, which was originally a part of the BBS dataset, was omitted from this study as it had been dropped from the SSCI from 1995 onwards. We also deliberately chose to end our study period at 2006 as the JBUS ceased publication after November 2006. Extra care was expended to handle typographical irregularities in the cited references of the journal articles sampled; e.g. "j finan" and "j fiance", as opposed to the correct abbreviated form for the JF, "j financ".

Table 1. Summary information for the publication data (1990-2006) used in this study

Journal	Source Publications	Number of Citations	Mean Citations per Publication
Financial Management (*FM*)	586	14,040	23.96
Journal of Banking and Finance (*JBF*)	1,695	43,485	25.65
Journal of Business (*JBUS*)	554	17,726	32.00
Journal of Finance (*JF*)	1,967	47,292	24.04
Journal of Financial Economics (*JFE*)	916	29,640	32.36
Journal of Financial and Quantitative Analysis (*JFQA*)	549	15,834	28.84
Review of Financial Studies (*RFS*)	627	21,131	33.70
Total	6,894	189,148	27.44

In order to put the present work into context, we have benchmarked the results of our methods with those obtained by Borokhovich, Bricker and Simkins in [2]. Discrepancies are used to identify possible errors. Errors resulting from the present computer codes are debugged accordingly. The numbers presented in the following Table 2 represent the number of times articles published in each of the eight finance journals cited articles in these journals during 1990 and 1991. The eight journals are FM, JBF, JBUS, JF, JFE, JFR, JFQA, and RFS. Additional entries are for the Journal of Political Economy (JPE), the American Economic Review (AER), Econometrica (ECMA), and an aggregate of other journals and nonjournals.

Table 2. A summary of the publication data (1990-1991) used in the study. Items in brackets indicate BBS values

Journal	Source Publications	Number of Citations	Mean Citations per Publication
Financial Management (*FM*)	99 (62)	1,681	16.98
Journal of Banking and Finance (*JBF*)	140 (130)	2,925	20.89
Journal of Business (*JBUS*)	58 (54)	1,250	21.55
Journal of Finance (*JF*)	210 (173)	4,590	21.86
Journal of Financial Economics (*JFE*)	74 (74)	1,998	27.00
Journal of Financial Research (*JFR*)	62 (62)	788	12.71
Journal of Financial and Quantitative Analysis (*JFQA*)	73 (73)	1,549	21.22
Review of Financial Studies (*RFS*)	63 (57)	1,601	25.41
Total	779 (685)	16,382	21.03

The self-citation rate is the percentage of a journal's citations attributable to its own articles. The self-citation index is a measure of how frequently a journal cites itself compared to how frequently the articles are cited by other journals. We had examined inter-journal communications by measuring the journal citation patterns within and outside the eight-journal set.

In summary, all journal datasets are within 10% of the BBS values except for JFQA. The reason for the extremely large discrepancy with JFQA is unknown at this point. We point out that the original table in BBS contained one typographical error, i.e. the value of citations from JBF to RFS is 15 and not 145 if the row sum is to equal 1,003 as indicated. It may be possible that there are further typographical errors yet to be identified. For this reason, we choose to include the present analyses for JFQA in case these values turn out to be more accurate estimates of journal citation patterns for these eight journals.

3 Inter-journal Citation Patterns: 1990-2006

A research article typically makes cited references to other research articles, thus creating a network of papers and journals that are connected through citation linkages. If one group cited the references of source articles by their respective source journals, the journal-to-journal citations can be split into two types: those that are directed internally and externally. The former corresponds to journal self-citation while the latter represents inter-journal communication. Since the total cited references made by a journal are proportional to its source article volume (total number of publications), and because the latter generally fluctuate from year to year, it is perhaps more appropriate to talk about inter-journal citation patterns through percentage of contributions.

3.1 Financial Management

From the present analysis, we find that the mean total citing frequency is 828, with a standard deviation of roughly 200 cited references. The total citing frequency ranges between 640 and 1,417 cited references. On average, *FM* cites *JF* the most (17.68 ± 3.59%), followed by *JFE* (16.55 ± 2.74%), *FM* (8.35 ± 3.75%), *JFQA* (2.78 ± 0.69%), *JBUS* (2.39 ± 0.98%), *RFS* (1.87 ± 1.01%) and *JBF* (1.05 ± 0.54%). This suggests that *FM* is primarily influenced by works in *JF*, *JFE* and *FM* itself. The self-citation rate for *FM* spikes considerably in 1997 at 0.1794 from 0.1076 in 1996 (See Table 3). This is largely due to a significant drop in citing frequency during that year for *JF* (by half) and *JFE* (by nearly a third), while *FM* experiences a considerable increase (22.3%).

3.2 Journal of Banking and Finance

For *JBF*, we find that the mean total citing frequency is 2572, with values each year ranging between 1,349 and 5,187 cited references, on the rise with annual publication

volume. *JBF* contributes the most citations to *JF* (11.62 ± 2.11%), *JFE* (7.41 ± 1.71%), *JBF* (6.57 ± 1.35%), *JFQA* (2.48 ± 0.64%), *JBUS* (1.94 ± 0.56%), *RFS* (1.81 ± 0.85%) and *FM* (0.78 ± 0.28%). The self-citation rate for *JBF* swings between 0.0319 and 0.0831.

3.3 Journal of Business

For *JBUS*, we find that the mean total citing frequency during 1990-2003 is roughly 646, with a standard deviation of 108 cited references. The total citing frequency then doubled with publication volume in 2004 to 1,747 cited references. In 2005 and 2006, that value soared to 3,013 and 3,991 cited references, respectively. On average, *JBUS* cites *JF* the most (13.75 ± 3.57%), followed by *JFE* (10.66 ± 2.73%), *JBUS* (4.52 ± 1.79%), *RFS* (2.79 ± 1.44%), *JFQA* (1.66 ± 0.62%), *JBF* (0.72 ± 0.56%) and *FM* (0.56 ± 0.35%). The self-citation rate for *JBUS* swings between 0.0238 and 0.0833.

3.4 Journal of Finance

For *JF*, we find that the mean total citing frequency is 2,793. On average, *JF* cites itself the most (19.30 ± 1.95%), followed by *JFE* (14.59 ± 2.06%), *RFS* (3.97 ± 1.30%), *JBUS* (2.53 ± 0.45%), *JFQA* (2.16 ± 0.45%), *FM* (0.69 ± 0.20%) and *JBF* (0.63 ± 0.24%). The self-citation rate for *JF* swings between 0.1518 and 0.2228.

3.5 Journal of Financial Economics

For *JFE*, we find that the mean total citing frequency is 1,755. On average, *JFE* cites itself the most (20.22 ± 4.18%), followed by *JF* (16.61 ± 2.85%), *RFS* (3.31 ± 1.19%), *JBUS* (2.13 ± 0.56%), *JFQA* (1.98 ± 0.60%), *FM* (0.95 ± 0.46%) and *JBF* (0.66 ± 0.30%). The self-citation rate for *JFE* swings between 0.1389 and 0.2910.

3.6 Journal of Financial and Quantitative Analysis

For *JFQA*, we find that the mean total citing frequency is 935. On average, *JFQA* cites *JF* the most (19.97 ± 2.98%), followed by *JFE* (16.45 ± 2.47%), *JFQA* (4.93 ± 1.26%), *RFS* (4.21 ± 1.92%), *JBUS* (2.68 ± 0.89%), *FM* (1.04 ± 0.54%) and *JBF* (0.92 ± 0.45%). The self-citation rate for *JFQA* swings between 0.0315 and 0.0806.

3.7 Review of Financial Studies

For *RFS*, we find that the mean total citing frequency is 1,246. On average, *RFS* cites *JF* the most (16.60 ± 3.47%), followed by JFE (11.97 ± 2.22%), RFS (6.79 ± 1.16%), JFQA (2.33 ± 0.52%), JBUS (2.41 ± 0.73%), JBF (0.63 ± 0.35%) and FM (0.40 ± 0.20%). The self-citation rate for *RFS* swings between 0.0272 and 0.0811.

The self-citation index for each journal in year Y is computed as the (self-citation rate in year Y × 100) ÷ (normalized average citations from other journals in year Y). From Fig. 1, the two journals with the highest self-citation index are FM and JBF with values rising and dipping below 1.00 throughout 1990-2006. Fig. 2 excludes the FM and JBF plots to resolve annual variations for the other five journals. Evidently, the other five journals possess a self-citation index below 1.00 throughout the study period. Furthermore, RFS appears to be experiencing a decreasing growth trend, while the other four journals fluctuate more or less around their mean values.

Table 3. Self-citation rates (1990-2006) for the seven journals studied

	Self-citation rate						
Year	FM	JBF	JBUS	JF	JFE	JFQA	RFS
1990	0.0685	0.0319	0.0833	0.1518	0.2451	0.0679	0.0272
1991	0.0529	0.0831	0.0480	0.1814	0.2126	0.0806	0.0692
1992	0.0471	0.0576	0.0474	0.1751	0.2488	0.0488	0.0637
1993	0.0543	0.0691	0.0464	0.1576	0.2910	0.0540	0.0747
1994	0.0815	0.0562	0.0782	0.1946	0.1988	0.0556	0.0723
1995	0.0672	0.0738	0.0317	0.1796	0.2182	0.0633	0.0627
1996	0.1076	0.0539	0.0511	0.1954	0.2147	0.0379	0.0700
1997	0.1794	0.0759	0.0375	0.1826	0.2410	0.0510	0.0671
1998	0.1292	0.0746	0.0552	0.2141	0.2270	0.0427	0.0748
1999	0.1282	0.0830	0.0238	0.1971	0.2149	0.0405	0.0633
2000	0.1139	0.0606	0.0255	0.2117	0.1673	0.0353	0.0703
2001	0.0810	0.0814	0.0517	0.2042	0.1710	0.0492	0.0704
2002	0.0909	0.0557	0.0653	0.2109	0.1442	0.0478	0.0811
2003	0.0765	0.0742	0.0364	0.2029	0.1807	0.0478	0.0657
2004	0.0531	0.0707	0.0280	0.2056	0.1652	0.0315	0.0750
2005	0.0437	0.0572	0.0266	0.1941	0.1584	0.0361	0.0765
2006	0.0450	0.0575	0.0323	0.2228	0.1389	0.0483	0.0704

Panel B: Basic data descriptors

Min.	0.0437	0.0319	0.0238	0.1518	0.1389	0.0315	0.0272
Median	0.0765	0.0691	0.0464	0.1954	0.2126	0.0483	0.0703
Mean	0.0835	0.0657	0.0452	0.1930	0.2022	0.0493	0.0679
Max.	0.1794	0.0831	0.0833	0.2228	0.2910	0.0806	0.0811
Range	0.1357	0.0512	0.0596	0.0710	0.1521	0.0491	0.0539
Std Dev	0.0375	0.0135	0.0179	0.0195	0.0418	0.0126	0.0116

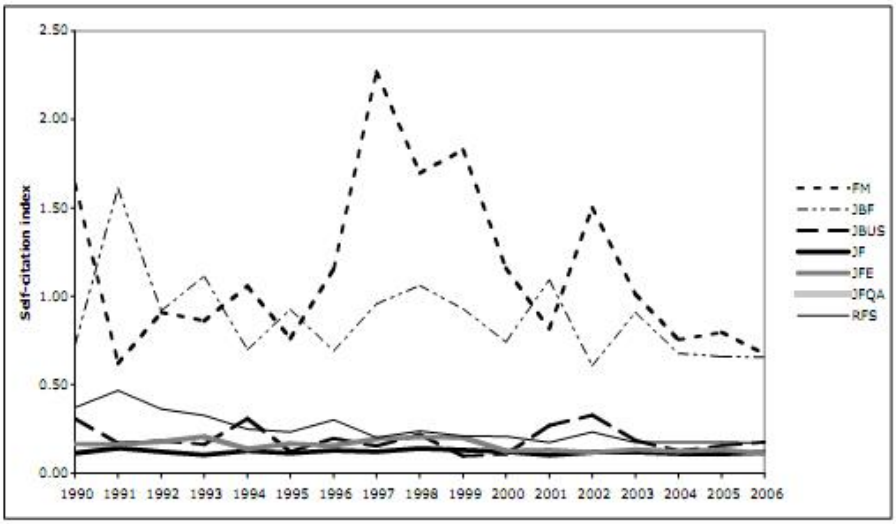

Fig. 1. Time evolution of self-citation index for the journals in the study dataset

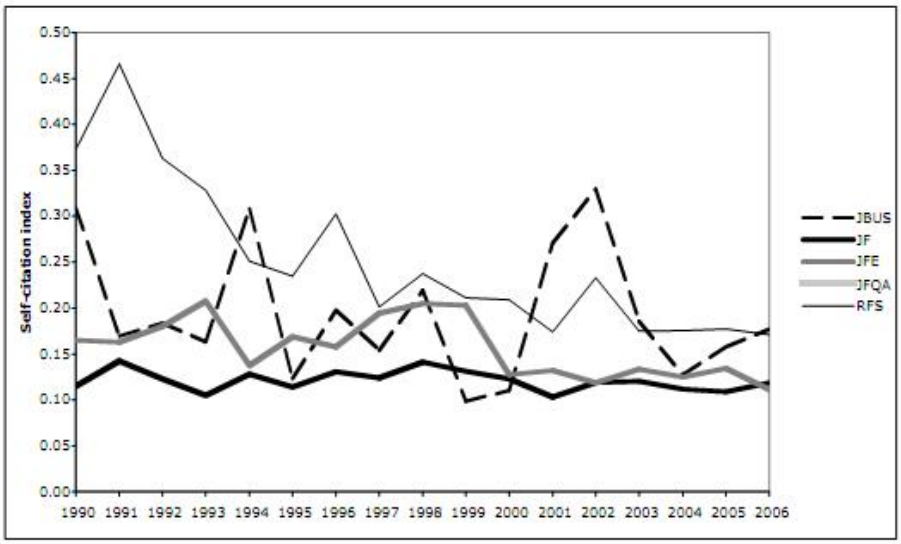

Fig. 2. Time evolution of self-citation index for JBUS, JF, JFE, JFQA and RFS

By sorting the self-citation index values in ascending order, we obtain the annual relative influence rank as shown in Table 4. This ranking reflects inter-journal citation patterns, with the highest rank representing the journal with either the smallest self-citation rate, and/or the largest annual citations contributed from other journals. This gives a practical and simple measure to gauge relative influence between journals, i.e. a journal is more influential to the development of other journals if it is cited more externally than internally. Although some permutations in the ordering occur

throughout the study period, a few patterns are visible: (1) JF always appears at rank 1 or 2; (2) JFE maintains its position in the top 4 ranks; (3) JBF and FM are positioned at ranks 6 and 7. Accordingly, a change in the ordering of journals by relative influence rank indicates a shift in inter-journal communication trends. This can be seen with the RFS, which has shown a gradual decline in self-citation index, corresponding to an upward shift in relative influence ranking. This suggests that RFS is becoming more prominent in the finance literature. On the other hand, the Journal of Business can be seen shuffling around ranks 1 (1999-2000) to 5 (1994, 2001-2002). Drops in ranking occur when the self-citation rate increases or when there is a decrease in the normalized average citations from other journals (indicative of reduced external influence).

Table 4. The relative Influence Rank obtained by sorting self-citation index values in ascending order

	Relative Influence Rank						
Year	1	2	3	4	5	6	7
1990	JF	JFE	JBUS	JFQA	RFS	JBF	FM
1991	JF	JFE	JBUS	JFQA	RFS	FM	JBF
1992	JF	JFE	JBUS	JFQA	RFS	FM	JBF
1993	JF	JBUS	JFE	JFQA	RFS	FM	JBF
1994	JF	JFE	JFQA	RFS	JBUS	JBF	FM
1995	JF	JBUS	JFE	RFS	JFQA	FM	JBF
1996	JF	JFQA	JFE	JBUS	RFS	JBF	FM
1997	JF	JBUS	JFE	RFS	JFQA	JBF	FM
1998	JF	JFQA	JFE	JBUS	RFS	JBF	FM
1999	JBUS	JF	JFQA	JFE	RFS	JBF	FM
2000	JBUS	JF	JFE	JFQA	RFS	JBF	FM
2001	JF	JFE	RFS	JFQA	JBUS	FM	JBF
2002	JFE	JF	JFQA	RFS	JBUS	JBF	FM
2003	JF	JFE	RFS	JBUS	JFQA	JBF	FM
2004	JF	JFE	JBUS	JFQA	RFS	JBF	FM
2005	JF	JFE	JBUS	RFS	JFQA	JBF	FM
2006	JFE	JF	RFS	JBUS	JFQA	JBF	FM

4 Concluding Remarks

In this paper, we have quantified the inter-journal citation patterns for seven prominent finance journals. Our analysis suggests that these journals have a particular ordering in terms of relative influence rank, with JF and JFE occupying top positions throughout the period of study. This could be indicative of a significant number of influential works located within the two journals that are current (relevant) to the development of other works elsewhere. Incidentally, this creates a bias for older, more

established journals which have a larger pool of works to cite from. This could explain the low rank for the younger, comparatively less self-citing, yet highly cited RFS. Despite beginning publication only in 1988, RFS averages 33.70 citations per publication during the period 1990-2006, the highest among the seven journals studied (see Table 1). To address this issue, one could tally citations to journals within a fixed time window, but this is exactly the approach utilized by the impact factor which we are trying to avoid. A more promising approach is to conduct a centrality analysis of the citation network for business/finance journals, from which a number of prominence scores can be constructed (reflecting different aspects of a journal's relative position within a structure of citation ties). This will be explored in future works.

Acknowledgements. The authors would like to acknowledge the UM-High Impact Research Grant No: F000013-21001 as well as the Fundamental Research Grant Scheme No: RG298-11HNE for support of this research work.

References

1. Garfield, E.: Agony and the ecstasy - the history and meaning of the journal impact factor. In: International Congress on Peer Review and Bibliomedical Publication, Chicago, September 16 (2005)
2. Adler, R., Ewing, J., Taylor, P.: Citation Statistics - A report from the International Mathematical Union (IMU) in cooperation with the International Council of Industrial and Applied Mathematics (ICIAM) and the Institute of Mathematical Statistics (IMS). Technical report, Joint Committee on Quantitative Assessment of Research (2008)
3. Borokhovich, K.A., Bricker, R.J., Simkins, B.J.: Journal Communication and Influence in Financial Research. J. Finance 49(2), 713–725 (1994)

REFERENCES

Acharya, V. V., Bharath, S. T., & Srinivasan, A. (2007). Does industry-wide distress affect defaulted firms? evidence from creditor recoveries. *Journal of Financial Economics*, *85*(3), 787–821.

Acharya, V. V., John, K., & Sundaram, R. K. (2000). On the optimality of resetting executive stock options. *Journal of Financial Economics*, *57*(1), 65–101.

Acharya, V. V., & Pedersen, L. H. (2005). Asset pricing with liquidity risk. *Journal of Financial Economics*, *77*(2), 375–410.

Adler, R., Ewing, J., & Taylor, P. (2009). Citation Statistics. *Statistical Science*, *24*(1), 1–28.

Albert, R., Jeong, H., & Barabási, A.-L. (2000). Error and attack tolerance of complex networks. *Nature*, *406*(6794), 378–382.

Almeida, H., & Philippon, T. (2007). The risk-adjusted cost of financial distress. *The Journal of Finance*, *62*(6), 2557–2586.

Almeida, H., & Wolfenzon, D. (2005). The effect of external finance on the equilibrium allocation of capital. *Journal of Financial Economics*, *75*(1), 133–164.

Althouse, B. M., West, J. D., Bergstrom, C. T., & Bergstrom, T. (2009). Differences in Impact Factor Across Fields and Over Time. *Journal of the American Society for Information Science and Technology*, *60*(1), 27–34.

Amaral, L., Scala, A., Barthélémy, M., & Stanley, H. (2000). Classes of small-world networks. *Proceedings of the National Academy of Sciences*, *97*(21), 11149–11152.

Archambault, É., & Larivière, V. (2009). History of the journal impact factor: Contingencies and consequences. *Scientometrics*, *79*(3), 635–649.

Baker, M., & Wurgler, J. (2002). Market timing and capital structure. *The Journal of Finance*, *57*(1), 1–32.

Banks, D. L., & Carley, K. M. (1996). Models for network evolution. *Journal of Mathematical Sociology*, *21*(1-2), 173–196.

Bansal, R., & Yaron, A. (2004). Risks for the long run: A potential resolution of asset pricing puzzles. *The Journal of Finance*, *59*(4), 1481–1509.

Barabási, A., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, *311*(3), 590–614.

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An Open Source Software for Exploring and Manipulating Networks.. Retrieved from `http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154`

Baumgartner, H., & Pieters, R. (2003). The Structural Influence of Marketing Journals: A Citation Analysis of the Discipline and Its Subareas over Time. *Journal of Marketing*, *67*(2), 123–139.

Beck, T., Levine, R., & Levkov, A. (2010). Big bad banks? the winners and losers from bank deregulation in the united states. *The Journal of Finance*, *65*(5), 1637–1667.

Bergstrom, C. (2007). Eigenfactor – Measuring the value and prestige of scholarly journals. *Coll Res Libr News*, *68*(5), 314–316.

Bertsimas, D., Brynjolfsson, E., Reichman, S., & Silberholz, J. M. (2014). *Moneyball for Academics: Network Analysis for Predicting Research Impact.* Available at SSRN 2374581.

Bilke, S., & Peterson, C. (2001). Topological Properties of Citation and Metabolic Networks. *Physical Review E*, *64*(3), 036106.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, *3*, 993–1022.

Blondel, V., Guillaume, J., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008.

Bornmann, L. (2008). Scientific Peer Review: An Analysis of the Peer Review Process from the Perspective of Sociology of Science Theories. *Human Architecture: Journal of the Sociology of Self-Knowledge*, *6*(2), 23–38.

Bornmann, L., & Daniel, H.-D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, *64*(1), 45–80.

Bornmann, L., Mutz, R., Neuhaus, C., & Daniel, H.-D. (2008). Citation counts for research evaluation: standards of good practice for analyzing bibliometric data and presenting and interpreting results. *Ethics in Science and Environmental Politics*, *8*(1), 93–102.

Borokhovich, K. A., Bricker, R. J., & Simkins, B. J. (1994). Journal Communication and Influence in Financial Research. *The Journal of Finance*, *49*(2), 713–725.

Bradford, S. C. (1985). Sources of information on specific subjects. *Journal of Information Science*, *10*(4), 173–180.

Brav, A., Graham, J. R., Harvey, C. R., & Michaely, R. (2005). Payout policy in the 21st century. *Journal of Financial Economics*, *77*(3), 483–527.

Breslin, J. G., Bojars, U., Aleman-Meza, B., Boley, H., Mochol, M., Nixon, L. J., ... Zhdanova, A. V. (2007). Finding experts using Internet-based discussions in online communities and associated social networks. In *First International ExpertFinder Workshop* (Vol. 11, pp. 38–47).

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

Burt, R. S. (1993). The Social Structure of Competition. In R. Swedberg (Ed.), (pp. 65–103). New York, USA: Russell Sage Foundation.

Burt, R. S. (1995). *Structural Holes: The Social Structure of Competition*. Harvard University Press, USA.

Burt, R. S. (2005). *Brokerage and closure: An Introduction to Social Capital*. Oxford University Press, USA.

Burt, R. S. (2010). *Neighbor Networks: Competitive Advantage Local and Personal: Competitive Advantage Local and Personal*. Oxford University Press, USA.

Cameron, B. D. (2005). Trends in the Usage of ISI Bibliometric Data: Uses, Abuses, and Implications. *portal: Libraries and the Academy*, *5*(1), 105–125.

Campbell, J. Y., Lettau, M., Malkiel, B. G., & Xu, Y. (2001). Have individual stocks become more volatile? an empirical exploration of idiosyncratic risk. *The Journal of Finance*, *56*(1), 1–43.

Carl, P., Peterson, B. G., Boudt, K., & Zivot, E. (2009). *PerformanceAnalytics: Econometric tools for performance and risk analysis.* Retrieved July 1, 2014, from `http://cran.r-project.org/web/packages/PerformanceAnalytics/index.html`

Chen, J., Hong, H., & Stein, J. (2001). Forecasting crashes: Trading volume, past returns, and conditional skewness in stock prices. *Journal of Financial Economics*, *61*(3), 345–381.

Chen, J., Hong, H., & Stein, J. (2002). Breadth of ownership and stock returns. *Journal of Financial*

Economics, *66*(2), 171–205.

Chen, P., & Redner, S. (2010). Community Structure of the Physical Review Citation Network. *Journal of Informetrics*, *4*(3), 278–290.

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

Chi, M. T. (2006). Two approaches to the study of experts' characteristics. In K. Ericsson, N. Charness, P. Feltovich, & R. Hoffman (Eds.), *Cambridge Handbook of Expertise and Expert Performance* (pp. 121–130). Cambridge University Press.

Choe, H., Lee, D. H., Seo, I. W., & Kim, H. D. (2013). Patent citation network analysis for the domain of organic photovoltaic cells: Country, institution, and technology field. *Renewable and Sustainable Energy Reviews*, *26*(0), 492 - 505. Retrieved from http://www.sciencedirect.com/science/article/pii/S1364032113003407 doi: http://dx.doi.org/10.1016/j.rser.2013.05.037

Clauset, A., Shalizi, C. R., & Newman, M. E. (2009). Power-law distributions in empirical data. *SIAM review*, *51*(4), 661–703.

Cole, S. (1970). Professional Standing and the Reception of Scientific Discoveries. *American Journal of Sociology*, 286–306.

Crane, D., & Kaplan, N. (1973). Invisible colleges: Diffusion of knowledge in scientific communities. *Physics Today*, *26*, 72.

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*, *1695*, 38.

Cukierski, W., Hamner, B., & Yang, B. (2011, July). Graph-based features for supervised link prediction. In *Neural Networks (IJCNN), The 2011 International Joint Conference on* (p. 1237-1244). doi: 10.1109/IJCNN.2011.6033365

Daud, A., Abbasi, R., & Muhammad, F. (2013). Finding rising stars in social networks. In *Database Systems for Advanced Applications* (pp. 13–24).

Ding, C., He, X., Husbands, P., Zha, H., & Simon, H. D. (2002). PageRank, HITS and a Unified Framework for Link Analysis. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 353–354). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/564376.564440 doi: 10.1145/564376.564440

Ding, Y. (2011a). Applying weighted PageRank to author citation networks. *Journal of the American Society for Information Science and Technology*, *62*(2), 236–245.

Ding, Y. (2011b). Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks. *Journal of Informetrics*, *5*(1), 187–203.

Dong, P., Loh, M., & Mondry, A. (2005). The "impact factor" revisited. *Biomedical Digital Libraries*, *2*(7), 1–8.

Easley, D., Hvidkjaer, S., & O'Hara, M. (2002). Is information risk a determinant of asset returns? *The Journal of Finance*, *57*(5), 2185–2221.

Egghe, L., & Rousseau, R. (1990). *Introduction to Informetrics: Quantitative Methods in Library, Documentation and Information Science*. Elsevier Science Publishers.

Ellis, K., Michaely, R., & O'Hara, M. (2000). When the underwriter is the market maker: An examination of trading in the ipo aftermarket. *The Journal of Finance*, *55*(3), 1039–1074.

Erdös, P., & Rényi, A. (1959). On random graphs I. *Publ. Math. Debrecen*, *6*, 290–297.

Erdös, P., & Rényi, A. (1960). On the evolution of random graphs. *Magyar Tud. Akad. Mat. Kutató Int. Közl*, *5*, 17–61.

Fafchamps, M., Leij, M., & Goyal, S. (2006). Scientific networks and co-authorship. *Department of Economics Discussion Paper Series*.

Fafchamps, M., Leij, M., & Goyal, S. (2010). Matching and network effects. *Journal of the European Economic Association*, *8*(1), 203–231.

Falagas, M. E., & Alexiou, V. G. (2008). The top-ten in journal impact factor manipulation. *Archivum Immunologiae et Therapiae Experimentalis*, *56*(4), 223–226.

Fama, E. F. (1965). The behavior of stock-market prices. *The Journal of Business*, *38*(1), 34–105.

Fatt, C. K., Ujum, E. A., & Ratnavelu, K. (2010). The structure of collaboration in the Journal of Finance. *Scientometrics*, *85*(3), 849–860.

Fehr, E., & Schneider, F. (2007). Implicit reputation cues and strong reciprocity. *unpublished paper, Institute for Empirical Research in Economics, University of Zürich*.

Feld, S. (1981). The focused organization of social ties. *American Journal of Sociology*, *86*(5), 1015–1035.

Festiger, L. (1949). The analysis of sociograms using matrix algebra. *Human Relations*, *2*(2), 153–158.

Fiske, A. P. (1991). *Structures of social life: The four elementary forms of human relations: Communal sharing, authority ranking, equality matching, market pricing*. New York, USA: Free Press.

Forsyth, E., & Katz, L. (1946). A matrix approach to the analysis of sociometric data: preliminary report. *Sociometry*, *9*(November), 340–347.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, *486*(3), 75–174.

Franceschet, M. (2010). Ten good reasons to use the Eigenfactor™ metrics. *Information Processing & Management*, *46*(5), 555–558.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, *40*(1), 35–41.

Fry, B. (2007). *Visualizing data: Exploring and explaining data with the processing environment*. California, USA: O'Reilly Media, Inc.

Fulda, J. S. (2008). The ethics of self-citation: Good reasons and bad to cite oneself. *Journal of Information Ethics*, *17*(1), 8–11.

Gaeta, T. J. (1999). Authorship: "Law" and Order. *Academic Emergency Medicine*, *6*(4), 297–301.

García-Pérez, M. A. (2009). A multidimensional extension to hirsch's h-index. *Scientometrics*, *81*(3), 779–785.

Garfield, E. (1970). Citation indexing for studying science. *Nature*, *227*(5259), 669–671.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science*, *178*(4060), 471–479.

Garfield, E. (1996). How can impact factors be improved? *BMJ: British Medical Journal*, *313*(7054), 411–413.

Garfield, E. (1999). Journal impact factor: a brief review. *Canadian Medical Association Journal*, *161*(8), 979–980.

Garfield, E. (2006). The history and meaning of the journal impact factor. *JAMA: The Journal of the American Medical Association*, *295*(1), 90–93.

Garfield, E., Sher, I. H., & Torpie, R. J. (1964). *The use of citation data in writing the history of science* (Tech. Rep.). Institute for Scientific Information Inc., Philadelphia, Pennsylvania, USA: DTIC Document.

Gibson, D., Kleinberg, J. M., & Raghavan, P. (1998). Inferring web communities from link topology. In *Proceedings of the Ninth ACM Conference on Hypertext and Hypermedia: Links, Objects, Time and Space* (pp. 225–234).

Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, *99*(12), 7821–7826.

Gladwell, M. (2008). *Outliers: The story of success*. New York, USA: Little, Brown and Company.

Goyal, S. (2009). *Connections: An Introduction to the Economics of Networks*. Princeton University Press.

Graham, J. R., & Harvey, C. R. (2001). The theory and practice of corporate finance: evidence from the field. *Journal of Financial Economics*, *60*(2), 187–243.

Graham, J. R., & Tucker, A. L. (2006). Tax shelters and corporate debt policy. *Journal of Financial Economics*, *81*(3), 563–594.

Greenberg, S. A. (2009). How citation distortions create unfounded authority: analysis of a citation network. *BMJ*, *339*, b2680. doi: 10.1136/bmj.b2680

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, *101*, 5228–5235.

Grossman, J. W. (1996). *The Erdös Number Project*. Retrieved Jan 18, 2014, from http://www.oakland.edu/enp/. Retrieved from `http://www.oakland.edu/enp/`

Hagberg, A. A., Schult, D. A., & Swart, P. J. (2008, August). Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th python in science conference (scipy2008)* (pp. 11–15). Pasadena, CA USA.

Hajra, K. B., & Sen, P. (2005). Aging in citation networks. *Physica A: Statistical Mechanics and its Applications*, *346*(1), 44–48.

Hanney, S., Frame, I., Grant, J., Buxton, M., Young, T., & Lewison, G. (2005). Using categorisations of citations when assessing the outcomes from health research. *Scientometrics*, *65*(3), 357–379.

Harrison, C. (2004). Peer review, politics and pluralism. *Environmental Science & Policy*, *7*(5), 357–368.

Harzing, A.-W. (2010). *The publish or perish book*. Tarma Software Research.

Hawkins, D., & Simon, H. A. (1949). Note: Some conditions of macroeconomic stability. *Econometrica*, *17*(3), 245–248.

Herrera, M., Roberts, D. C., & Gulbahce, N. (2010). Mapping the evolution of scientific fields. *PloS ONE*, *5*(5), e10355.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(46), 16569.

Hirshfield, L. E. (2011). *Authority, expertise, and impression management: Gendered professionalization of chemists in the academy*. Unpublished doctoral dissertation, The University of Michigan. Retrieved from `http://deepblue.lib.umich.edu/bitstream/handle/2027.42/89796/lhirshf_1.pdf?sequence=1`

Hoffman, P. (1998). *The man who loved only numbers: The story of Paul Erdös and the search for mathematical truth*. New York: Hyperion.

Hood, W. W., & Wilson, C. S. (2001). The literature of bibliometrics, scientometrics, and informetrics. *Scientometrics*, *52*(2), 291–314.

Hubbell, C. (1965). An input-output approach to clique identification. *Sociometry*, *28*(4), 377–399.

Jeong, H., Néda, Z., & Barabási, A.-L. (2003). Measuring preferential attachment in evolving networks. *EPL (Europhysics Letters)*, *61*(4), 567–572.

Johnson, J., & Podsakoff, P. (1994). Journal influence in the field of management: An analysis using salancik's index in a dependency network. *Academy of Management Journal*, *37*(5), 1392–1407.

Judge, T. A., Cable, D. M., Colbert, A. E., & Rynes, S. L. (2007). What causes a management article to be cited—article, author, or journal? *Academy of Management Journal*, *50*(3), 491–506.

Kajikawa, Y., Ohno, J., Takeda, Y., Matsushima, K., & Komiyama, H. (2007). Creating an academic landscape of sustainability science: an analysis of the citation network. *Sustainability Science*, *2*(2), 221–231.

Kajikawa, Y., & Takeda, Y. (2009). Citation network analysis of organic LEDs. *Technological Forecasting and Social Change*, *76*(8), 1115–1123.

Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika*, *18*(1), 39–43.

Kleinberg, J. (1998). Authoritative Sources in a Hyperlinked Environment. In *Proceedings of ACM-SIAM Symposium on Discrete Algorithms* (pp. 668–677). ACM.

Kleinberg, J. (1999). Hubs, authorities, and communities. *ACM Computing Surveys (CSUR)*, *31*(4es). (article No. 5)

Kleinberg, J. (2000). The Small-world Phenomenon: An Algorithmic Perspective. In *Proceedings of the Thirty-second Annual ACM Symposium on Theory of Computing* (pp. 163–170). New York, NY, USA: ACM. Retrieved from `http://doi.acm.org/10.1145/335305.335325` doi: 10.1145/335305.335325

Kleinberg, J., & Raghavan, P. (2005). Query incentive networks. In *Foundations of Computer Science, 2005. FOCS 2005. 46th Annual IEEE Symposium on* (pp. 132–141).

Kleinberg, J. M. (1999, September). Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, *46*(5), 604–632. Retrieved from `http://doi.acm.org/10.1145/324133.324140` doi: 10.1145/324133.324140

Kossinets, G., & Watts, D. J. (2009). Origins of Homophily in an Evolving Social Network. *American Journal of Sociology*, *115*(2), 405–450.

Krampen, G., Becker, R., Wahner, U., & Montada, L. (2007). On the validity of citation counting in science evaluation: Content analyses of references and citations in psychological publications. *Scientometrics*, *71*(2), 191–202.

Langville, A. N., & Meyer, C. D. (2006). *Google's PageRank and Beyond: The Science of Search Engine Rankings*. New Jersey, USA: Princeton University Press.

Lawrence, P. A. (2003). The politics of publication. *Nature*, *422*(6929), 259–261.

Lehmann, S., Jackson, A., & Lautrup, B. (2008). A quantitative analysis of indicators of scientific performance. *Scientometrics*, *76*(2), 369–390.

Lehmann, S., Jackson, A. D., & Lautrup, B. E. (2006). Measures for measures. *Nature*, *444*(7122), 1003–1004.

Leontief, W. W. (1941). *The structure of American economy, 1919-1939: an empirical application of equilibrium analysis*. New York, USA: Oxford University Press.

Leslie, D. (2005). Are delays in academic publishing necessary? *The American Economic Review*, *95*(1), 407–413.

Li, C.-L., Su, Y.-C., Lin, T.-W., Tsai, C.-H., Chang, W.-C., Huang, K.-H., ... Lin, C.-J. (2013). Combination of Feature Engineering and Ranking Models for Paper-author Identification in KDD Cup 2013. In *Proceedings of the 2013 KDD Cup 2013 Workshop* (pp. 2:1–2:7). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2517288.2517290 doi: 10.1145/2517288.2517290

Liu, N. C., & Cheng, Y. (2005). The academic ranking of world universities. *Higher education in Europe*, *30*(2), 127–136.

Luce, R. D., & Perry, A. D. (1949). A method of matrix analysis of group structure. *Psychometrika*, *14*(2), 95–116.

MacRoberts, M. H., & MacRoberts, B. R. (1996). Problems of citation analysis. *Scientometrics*, *36*(3), 435–444.

MacRoberts, M. H., & MacRoberts, B. R. (2010). Problems of citation analysis: A study of uncited and seldom-cited influences. *Journal of the American Society for Information Science and Technology*, *61*(1), 1–12.

Malkiel, B. G., & Fama, E. F. (1970). Efficient Capital Markets: A Review Of Theory And Empirical Work*. *The Journal of Finance*, *25*(2), 383–417.

Marlow, C. (2004). Audience, structure and authority in the weblog community. In *International communication association conference.* New Orleans, LA. Retrieved from http://rockngo.org/wp-content/uploads/mt/archives/ICA2004.pdf

Martin, T., Ball, B., Karrer, B., & Newman, M. E. J. (2013, April). Coauthorship and citation in scientific publishing. *arXiv preprint corr/abs-1304-0473*.

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.

McNee, S. M., Albert, I., Cosley, D., Gopalkrishnan, P., Lam, S. K., Rashid, A. M., ... Riedl, J. (2002). On the Recommending of Citations for Research Papers. In *Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work* (pp. 116–125). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/587078.587096 doi: 10.1145/587078.587096

McPherson, M., Smith-Lovin, L., & Cook, J. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, *27*, 415–444.

Medina, C., & Leeuwen, T. (2012). Seed journal citation network maps: A method based on network theory. *Journal of the American Society for Information Science and Technology*, *63*(6), 1226–1234.

Merton, R. K. (1968). The Matthew Effect in Science. *Science*, *159*(3810), 56–63.

Merton, R. K. (1988). The Matthew Effect in Science, II: Cumulative Advantage and the Symbolism of Intellectual Property. *ISIS*, *79*(4), 606–623.

Moed, H. F. (2000). Bibliometric indicators reflect publication and management strategies. *Scientometrics*, *47*(2), 323–346.

Moed, H. F. (2008). UK Research Assessment Exercises: Informed judgments on research quality or quantity? *Scientometrics*, *74*(1), 153–161.

Myers, J. L., Well, A., & Lorch, R. F. (2010). *Research Design and Statistical Analysis* (3rd ed.). New York, USA: Routledge.

Neff, B. D., & Olden, J. D. (2010). Not so fast: inflation in impact factors contributes to apparent improvements in journal quality. *BioScience*, *60*(6), 455–459.

Newman, M. E. J. (2001a). Clustering and preferential attachment in growing networks. *Physical Review E*, *64*(2), 025102.

Newman, M. E. J. (2001b). Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical Review E*, *64*(1), 016132.

Newman, M. E. J. (2001c). Scientific collaboration networks. I. Network construction and fundamental results. *Physical Review E*, *64*(1), 016131.

Newman, M. E. J. (2001d). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, *98*(2), 404–409.

Newman, M. E. J. (2002). Assortative mixing in networks. *Physical Review Letters*, *89*(20), 208701.

Nicolaisen, J., & Hjørland, B. (2007). Practical potentials of bradford's law: A critical examination of the received view. *Journal of Documentation*, *63*(3), 359–377.

O'Hara, M. (2003). Presidential address: Liquidity and price discovery. *The Journal of Finance*, *58*(4), 1335–1354.

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.

Pontille, D., & Torny, D. (2010). The controversial policies of journal ratings: Evaluating social sciences and humanities. *Research Evaluation*, *19*(5), 347–360.

Prathap, G. (2010). Is there a place for a mock h-index? *Scientometrics*, *84*(1), 153–165.

Pritchard, A. (1969). Statistical bibliography or bibliometrics. *Journal of Documentation*, *25*(4), 348–349.

Radicchi, F., Fortunato, S., Markines, B., & Vespignani, A. (2009). Diffusion of scientific credits and the ranking of scientists. *Physical Review E*, *80*(5), 056103.

Rapoport, A. (1953). Spread of information through a population with socio-structural bias: I. Assumption of transitivity. *The Bulletin of Mathematical Biophysics*, *15*(4), 523–533.

Ratnavelu, K., Fatt, C. K., & Ujum, E. A. (2012). Inter-journal Citations and Ranking Top Journals in Financial Research. In P. Balasubramaniam & R. Uthayakumar (Eds.), *Mathematical Modelling and Scientific Computation* (Vol. 283, pp. 505–513). Springer-Verlag Berlin Heidelberg.

Redner, S. (1998). How popular is your paper? an empirical study of the citation distribution. *The European Physical Journal B-Condensed Matter and Complex Systems*, *4*(2), 131–134.

Redner, S. (2005). Citation statistics from 110 years of Physical Review. *Physics Today*, *58*, 49–54.

Redner, S. (2010). On the meaning of the h-index. *Journal of Statistical Mechanics: Theory and Experiment*, *2010*(03), L03005.

Rogers, L. F. (2002). Impact Factor: The Numbers Game. *American Journal of Roentgenology*, *178*(3), 541–542.

Rosvall, M., Axelsson, D., & Bergstrom, C. T. (2009). The map equation. *The European Physical Journal Special Topics*, *178*(1), 13–23.

Rosvall, M., & Bergstrom, C. T. (2010). Mapping change in large networks. *PloS ONE*, *5*(1), e8694.

Rosvall, M., Grönlund, A., Minnhagen, P., & Sneppen, K. (2005). Searchability of networks. *Physical Review E*, *72*(4), 046117.

Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika*, *31*(4), 581–603.

Salancik, G. (1986). An index of subgroup influence in dependency networks. *Administrative Science Quarterly*, *31*(2), 194–211.

Scully, C., & Lodge, H. (2005). Impact factors and their significance; overrated or misused? *British Dental Journal*, *198*(7), 391–393.

Seglen, P. O. (1997). Why the impact factor of journals should not be used for evaluating research. *BMJ: British Medical Journal*, *314*(7079), 498.

Shafer, P., Isganitis, T., & Yona, G. (2006). Hubs of knowledge: using the functional link structure in Biozon to mine for biologically significant entities. *BMC Bioinformatics*, *7*(1), 71.

Shimbo, M., Ito, T., & Matsumoto, Y. (2007). Evaluation of kernel-based link analysis measures on research paper recommendation. In *Proceedings of the 7th acm/ieee-cs joint conference on digital libraries* (pp. 354–355). New York, NY, USA: ACM. Retrieved from `http://doi.acm.org/10.1145/1255175.1255245` doi: 10.1145/1255175.1255245

Shockley, W. (1957). On the statistics of individual variations of productivity in research laboratories. *Proceedings of the IRE*, *45*(3), 279–290.

Sidiropoulos, A., Katsaros, D., & Manolopoulos, Y. (2007). Generalized hirsch h-index for disclosing latent facts in citation networks. *Scientometrics*, *72*(2), 253–280.

Silen, W. (1971). Publish or perish. *Archives of Surgery*, *103*(1), 1.

Silverman, B. W. (2009). Comment: Bibliometrics in the context of the UK research assessment exercise. *Statistical Science*, *24*(1), 15–16.

Simkin, M. V., & Roychowdhury, V. P. (2005). Stochastic modeling of citation slips. *Scientometrics*, *62*(3), 367–384.

Smith, S. D. (2004). Is an Article in a Top Journal a Top Article? *Financial Management*, *33*(4), 133–149.

Tague-Sutcliffe, J. (1992). An Introduction to Informetrics. *Information Processing & Management*, *28*(1), 1–3.

Tort, A. B., Targino, Z. H., & Amaral, O. B. (2012). Rising Publication Delays Inflate Journal Impact Factors. *PloS ONE*, *7*(12), e53374.

Tscharntke, T., Hochberg, M. E., Rand, T. A., Resh, V. H., & Krauss, J. (2007). Author sequence and credit for contributions in multiauthored publications. *PLoS Biology*, *5*(1), e18.

Usher, A., & Savino, M. (2007). A global survey of university ranking and league tables. *Higher Education in Europe*, *32*(1), 5–15.

Van Raan, A. F. (1997). Scientometrics: State-of-the-art. *Scientometrics*, *38*(1), 205–218.

Vázquez, A. (2001). Statistics of citation networks. *arXiv preprint cond-mat/0105031*.

Wakefield, R. (2008). Networks of accounting research: A citation-based structural and network analysis. *The British Accounting Review*, *40*(3), 228–244.

Walter, G., Bloch, S., Hunt, G., & Fisher, K. (2003). Counting on citations: a flawed way to measure quality. *Medical Journal of Australia*, *178*(6), 280–281.

Watts, D. J., & Strogatz, S. H. (1998, 06 04). Collective dynamics of 'small-world' networks. *Nature*, *393*(6684), 440–442. Retrieved from `http://dx.doi.org/10.1038/30918`

West, J. D., Bergstrom, T. C., & Bergstrom, C. T. (2010). The Eigenfactor Metrics™: A network approach to assessing scholarly journals. *College & Research Libraries*, *71*(3), 236–244.

Williams, G. (2007). Should we ditch impact factors? *BMJ: British Medical Journal*, *334*(7593), 568.

Wright, A. F. (2001). How evidence gets published. *Occasional Paper (Royal College of General Practitioners)*, *80*(July), 7–9.

Yang, Z., Yin, D., & Davison, B. D. (2011). Award Prediction with Temporal Citation Network Analysis. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1203–1204). New York, NY, USA: ACM. Retrieved from `http://doi.acm.org/10.1145/2009916.2010120` doi: 10.1145/2009916.2010120

Zhang, L. (2005). The value premium. *The Journal of Finance*, *60*(1), 67–103.

Zuccala, A. (2005). Modeling the invisible college. *Journal of the American Society for Information Science and Technology*, *57*(2), 152–168.

Życzkowski, K. (2010). Citation graph, weighted impact factors and performance indices. *Scientometrics*, *85*(1), 301–315.