UNIVERSITY OF SOUTHAMPTON

Understanding institutional collaboration networks: Effects of collaboration on research impact and productivity

by

Jiadi Yao

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

in the
Web and Internet Science Group
Electronics and Computer Science
Faculty of Physical and Applied Sciences

June 2014

UNIVERSITY OF SOUTHAMPTON

Abstract

Web and Internet Science Group
Electronics and Computer Science
Faculty of Physical and Applied Sciences

Doctor of Philosophy

by Jiadi Yao

There is substantial competition among academic institutions. They compete for students, researchers, reputation, and funding. For success, they need not only to excel in teaching, but also their research profile is considered an important factor. Institutions accordingly take actions to improve their research profiles. They encourage researchers to publish frequently and regularly (publish or perish) on the assumption that this generates both more and better research. Collaboration has also been encouraged by institutions and even required by some funding calls.

This thesis examines the empirical evidence on the interrelations among *institutional* research productivity, impact and collaborativity.

It studies article publication data across ACM and Web of Science covering five disciplines – Computer Science, Pharmacology, Materials Science, Psychology and Law. Institutions that publish less seek to publish collaboratively with other institutions. Collaboration boosts productivity for all the disciplines investigated excepted Law; however, the amount of productivity increase resulting from the institutions' attempt to collaborate more is small. The world's most productive institutions publish at least 50% of their papers on their own. Institutions doing more collaborative work are not found to correlate strongly with their impact either. The correlation between collaborativity and individual paper impact or institutional impact is small once productivity has been partialled out. In Computer Science, Pharmacology and Materials Science, no correlation is found. The decisive factor appears to be productivity. Partialling out productivity results in the largest reductions in the remaining correlations. It may be that only better equipped and well-funded institutions can publish without having to rely on external collaborators. These institutions have been publishing most of their output non-collaboratively, and are also of high quality and highly reputable, which may have equipped and funded them in the first place.

Contents

List of Figures

List of Tables

Declaration of Authorship

I, Jiadi Yao, declare that this thesis titled, ‘Understanding institutional collaboration networks: Effects of collaboration on research impact and productivity’ and the work presented in it are my own. I confirm that:

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

Signed: ____

Date: ____

To my wife Yiwen and my parents

For all your love, support and encouragement

Acknowledgements

Thank you to Prof Les Carr for being a fantastic supervisor over the past four years. He has given me many opportunities to present this work at conferences, provide me research exchange opportunities during my study.

Thank you to Prof Stevan Harnad for being an incredible supervisor. He has given me crucial insights and guidance towards the completion of this work. Thank you for looking after me while I was on the research exchange in Montréal. Without any of them, this work would never have gone this far.

Thank you to Yves Gingras and Vincent Larivière for the discussion we had in Montréal and provided me the research data from the Web of Science.

Thank you to Andrew Bell and Sheila Bowen for spending hours to proof read my work, which probably is not interesting to them.

Thank you to everyone who has worked around me in level 3 building 32 for good humour and distracting me from writing this thesis. I've enjoyed having you guys around. Thank you to all members of WAIS group, past and present. It's been an honour to know and work with you all.

Abbreviations

There are three main institutional variables – productivity, impact and collaboration, each is represented by the abbreviations below and carries the meaning as following:

P Institutional productivity, includes the paper count.

Q Institutional impact, includes all three impact variables (total citation per institution, pagerank weighted citation and average citation per paper).

C Institutional collaboration, includes all three collaboration variables (collaborative paper count, size-weighted collaboration and percent collaboration).

'Quality' has been used to indicate different aspects of the research activity in the literature. The specific meaning of quality discussed in this thesis is as follows:

Institutional Quality The perceived quality of an institution. (e.g. historical & current reputation, research output, research output impact, funding and infrastructures etc.) Institutions generally cover entities such as universities, research centres and company research laboratories that produce research output. We focus mainly on universities in this study.

Research Quality The quality of the entire research cycle. (e.g. methodology quality, report quality etc.)

Paper Quality The quality of a published paper. Citation impact and its variants are often used as proxies in the literature.

The abbreviation for the sub-variables are as follows.

The original and raw variables are represented in bold and italic:

PUBTOT	Total institutional paper output
CITTOT	Total citations per institution.
CITTOTw	PageRanked citations per institution. Incoming citations weighted by citation weight of citing institution.
CITAV	Average citations per paper.
WRANK	Institutional Webometrics Rank, July 2010 version.
PUBCOLL	Number of collaborative papers.
PUBCOLLw	Number of collaborative papers, with collaboration size-weighted.
PUBCOLL%	Percentage of collaborative papers over total papers per institution.

This study uses partial correlation to remove the effect of the third variable in order to find the true relationship. The partialled states of the variables are represented in italics-only:

PUBTOT	Total institutional paper output with impact or collaboration controlled.
CITTOT	Citations per Institution with productivity or collaboration controlled.
CITTOTw	PageRanked citations per institution with productivity or collaboration controlled.
CITAV	Citations Per Paper with productivity or collaboration controlled.
WRANK	Institutional Webometrics Rank with productivity or collaboration controlled
PUBCOLL	Number of Collaborative Papers with productivity or impact controlled.
PUBCOLLw	Size-weighted Collaboration with productivity or impact controlled.
PUBCOLL%	Percent Collaboration with productivity or impact controlled.

Chapter 1

Introduction

Scientific and scholarly research is a systematic investigation of data in order to establish facts and reach new conclusions. It is mostly conducted by scientists working in universities, research institutes and companies' research laboratories.

A typical research work that carried out by the above establishments generally has 4 stages:

1. Identification of the knowledge gap
2. Creation of knowledge
3. Quality assurance
4. Dissemination

At the gap identification stage, researchers make their observation of the world, review previous publications and then propose research questions. The knowledge creation stage is where the practical investigation happens. They design experiments, collect data and then analyse the data. They then document this process, so that anyone can replicate the procedure to obtain the same outcome. Their interpretation of the result is also presented. Peer reviews are conducted afterwards for quality assurance, and these are carried out by experts in the field to make sure the research is sound. Finally, the research is disseminated, so that other researchers can use it as basis for further research.

1.1 Research Collaboration

Traditionally, research was carried out by a single scientist or their colleagues in the same laboratory. They would conduct the required experiments themselves, even though they may not initially have the necessary skills or equipment. If they could not conduct an experiment themselves, they would reach out to a potential collaborator to get help, in return for including them as co-authors. This kind of ad-hoc collaboration is heavily dependent on the researcher's personal connections. The first paper with multiple authors listed was published in 1665[122].

The problems and questions scientists try to resolve are getting more and more complex and increasingly multi-disciplinary. For example, the problem of global warming and climate change; research of the Web and the need to understand its reciprocal effects on human society. These topics require experts from multiple disciplines and the mode of ad-hoc collaboration can no longer sustain the needs of these researches. As a result, collaborative activities, both from bottom up (researchers collaborate with each other) and top down (funders funding collaborative research; institutional collaboration) have increased dramatically in the recent years.

Collaborative research has also been strongly encouraged by institutions on the assumption that researchers working together – especially across institutions and across national borders – generate research that is higher in both quantity and impact. Institutions are not alone in encouraging collaboration. Funding bodies, such as JISC and the European Framework Programme (FP), often have specific requirements for individual, inter-institutional and international collaboration in their funding calls[8, 55].

1.2 Funding Research

The funding structure differs from country to country. In the UK, most of the research is publicly funded. A percentage of tax payers' money is allocated to funding councils and research councils. England, Scotland, Wales and Northern Ireland each has a corresponding higher education funding council to allocate funds to institutions. In 2014, they have jointly conducted the Research Excellence Framework (REF 2014) to assess the research output of the UK institutions. The REF outcome will be used as a reference for allocating funds between institutions. There are also different research councils focusing on specific fields, e.g, Engineering and Physical Science Research Council (EPSRC), Economics and Social Science Research Council (ESRC). These research councils steer the direction of the research by releasing call for bids. Scientists then put forward research proposals to compete for grants. Charities in the UK are also major funders, especially in medical research (e.g. Cancer Research UK is a charity focuses on funding cancer related research; British Heart Foundation focuses on funding heart related research). Some research is also funded by industry and private companies.

1.3 The Role of Universities in Research

The role of universities in producing research output varies depending on the countries' development and economic status. Research involves a high investment, which does not generate direct return on the investments. Universities in less developed countries, due to their economic status and government strategies, may not allocate much funding in conducting research. In these countries, universities tend to focus on education.

In the UK, US, Canada and Australia, universities are one of the biggest output of research publications[105]. In the UK, a group of 24 research focused universities form the Russell Group. This group represents two-thirds of all UK research grant spending. In the US, universities are also evaluated and ranked depending on their research activities, resulting in high pressure for them to do well in research. The US alone represents 28% of all world research as recorded by WoS, most of which is university output[200].

1.4 Competitions between Universities and Scientists in Research

There is substantial competition among academic institutions. They compete for students, scientists, reputation, and funding. For success, they need to excel in both teaching and research. Universities accordingly have policies that encourage researchers to publish frequently on the assumption that this generates both more and better research.

Public perception on quality of a university often includes, but not limited to students' experience, students' prospects, and research output performance. In these respects, university rankings play an important role in communicating universities' quality to the public. There are more than a dozen well established university quality rankings. They are used by the general public, potential students, researchers and donors to assess the university's overall quality. Almost all of these rankings' evaluations are based on some aspects of research activities, *e.g.*, some count the university's number of publications and some use citations to the papers. Many universities treat these rankings seriously and work hard to improve them, even though experts may have their own views regarding the validity of these rankings[22].

As a scientist, they have their own incentives to make a stronger impact to the community. To produce more publications is one effective approach. In academia there is a well known phrase – "publish or perish" [13] that describes the pressure to rapidly and regularly publish academic work. Publishing more is important for a scientist's career advancement. In order to secure a research position, to sustain or to further one's career, and to succeed in applying for grants, scientists need to publish frequently. They are often judged by the number of papers they have published, the qualities of the journals they have published in and the number of citations they have received for their papers. Producing more publications apparently makes some of these measurements look better.

1.5 Research Contributions

The literature of studying the relationship among citation impact, collaborativity and productivity is extensive and comprehensive. But most of them focus on the aggregation level of individual researcher, journal, discipline or country. This study look at these relationships specifically at the institution level using large scale data covering multiple disciplines.

Many studies consider variables separately. As we have demonstrated in chapter 4, citation impact, collaborativity and productivity are circularly correlated. This means that there might be a common factor presented in each variable which lead to the high pairwise correlation. The current study uses partial correlation to remove the effect of the third variable in the circular correlation before correlate the remaining pair of variable, giving evidence that increased citation is more associated with productivity than collaborativity at the institution level in the disciplines studied.

This also gives additional evidence on the disciplinary differences towards research collaboration and citation. Institutions are recommended to develop discipline-specific strategy in encouraging research collaboration and publication. Top cited institutions do not publish engineering and nature science papers collaboratively, while they do for social science and humanity papers.

Finally, this study processed tens of thousands of lines of free text representing universities, and created a lookup table of university's alternative names that maybe useful in other studies.

1.6 Thesis Structure

In chapter 2, we start with studying the meaning of institutional quality, institutional collaboration and productivity, and what methods were used to measure these factors. The attempts to establish the relationship between these three activities were also reviewed.

Chapter 3 describes the resources, datasets and methods to be used in the further chapters. In particular, it describes the necessary pre-processing done to the dataset in order to apply the correlation methods. The assignment of the papers, citations and collaborations to the universities are presented. Based on the dataset and previous studies, I make decisions on what metrics to use to approximate the three factors. These decisions were justified and the metrics were explained. The statistical methods – correlation, normalisation, partial correlation that help understand the relationship between the metrics are described. A network visualisation of the institutional collaboration network is presented, showing evidence of correlation expected.

Chapter 4 investigates and presents the results of the pairwise correlation between the three factors. These pairwise correlations used the unfiltered variables that directly come from the statistics of the universities. Unfiltered variables were commonly used by previous studies; comparison is made when they are comparable with previous studies.

Chapter 5 takes chapter 4 one step further by partialling out the unwanted variables between the pairwise correlations. The results are presented and compared.

Finally, the conclusions of the correlation analysis, answers to the research questions and recommendations on research practice for universities and scientists are presented in chapter 8.

Chapter 2

Measuring Research Activities and their Relationships

2.1 Research Collaboration

The concept of collaboration has been taken for granted in most of the literature but there have been few attempts to address the meaning of collaboration. Collaboration means that individuals work together to achieve a common goal. Research collaboration, therefore, means that individuals work together to advance scientific knowledge. But this immediately raises a question: how closely should they be working to qualify as collaboration? In one extreme case, as Subramanyam [178] suggested in his paper, the entire scientific community is like a big collaboration because they work collaboratively to advance science: scholars learn from each other from their publications, make comments to each other, suggest hypotheses to test, exchange ideas and share techniques.

In another extreme case, for example, if research collaborators are only those who have contributed directly and frequently to all tasks in a piece of research, then almost all collaborating researchers will be excluded, even those working closely together who have published scientific papers together. This is because one of the main reasons for researchers to collaborate is that not everyone knows and does everything, so during the scientific collaboration, it is not common for every collaborator to be involved equally intensively in every aspect of the research.

Collaboration can have different meanings in different research areas and contexts. A technician operating a specific piece of machinery can be counted as a collaborator in one discipline (*e.g.* in physics, operators are listed as co-author), but may not in others (*e.g.* in medical research, the operator of radioactive medical equipment is not listed as a co-author).

2.1.1 Co-authorship as Collaboration Measurement

We have seen that it is not possible to strictly measure the strength of research collaboration by the interactions due to its complexity. So scholars have also tried quantitative measures.

Attempts to quantify research collaboration date back to the 1950s. Psychologist Smith [172] observing the increase of multi-authored publications was the first to suggest using multi-authorship as a proxy to quantify collaborative activities between researchers. This is often referred as co-authorship. Co-authorship is a research practice, where the primary author includes other researchers as co-authors in the publication to recognise their contributions in the work.

Price and Beaver [159] were among the first to use co-authorship as the measure of collaboration strength between authors. Since then, co-authorship has been used widely as a metric of research collaboration.

Collaboration strength between a pair of authors is often approximated by counting the number of papers that lists the pair's names. A pair of authors who have co-authored more papers is more collaborative than than a pair who have co-authored fewer papers. By connecting authors who have co-authored papers, a network of authors can be constructed. The network methods are discussed in chapter 6. More advanced algorithms than simply counting the number of co-authored papers were presented and compared by Rousseau [163].

The suitability of co-authorship as a measure of collaboration has been investigated by Subramanyam [178]. He noted that the nature and magnitude of the collaboration changes during the course of the collaboration, so the precise nature and magnitude of collaboration cannot be determined using methods of interviews or questionnaires, let alone co-authorship. A popular example is that a casual conversation during a coffee break could be more valuable than week-long laboratory work towards the success of the project.

Using co-authorship relies on the assumption that the goal of collaboration between the authors is publishing articles. That is, a co-authored paper is indeed a result of collaboration, not something else. However, Katz [104, 105] and He *et al.*[93] demonstrated several scenarios, where either collaborative work does not yield a co-authored paper, or a co-authored paper does not involve any collaboration between the listed co-authors. In addition, social pressures come into play in determining whose name should be included on a published article [97]. Indeed, some higher position scientists (*e.g.* the leader of a research group) are often listed as co-authors despite they did not contribute. Furthermore, Bozeman and Lee [33] identified two specific types of co-authorship that often appear but are less relevant to collaboration: 1. data provider listed as co-author; 2. equipment provider listed as co-author.

As a result, we should be aware of the limitation in using co-authorships. While the social aspects of the collaborations may not be quantifiable using conventional methods, co-authorship offers an unique quantitative approach to measure the strength of the collaboration between authors. Here is a list of advantages using co-authorship to measure collaboration:

- invariant and verifiable. Anyone with the same dataset, using the same method can reproduce the result.

- inexpensive and practical. Counting co-authorship in datasets is computationally cheap. The programming is often simple and straight-forward. The dataset is readily available.

- large sample size, statistically more informative than case-study or qualitative method.
- non-reactive. It will not lead authors to co-author papers in the short term to increase their collaboration. However, wide adoption of this measurement potentially may affect the collaboration process in the longer term [178].

2.1.2 Other Approaches to Measure Collaboration

Beyond co-authorship, other measuring methods were also used. Rigby and Edler [162] used the proportion of co-subprojects (people who worked on the same sub-project) to measure the intensity of collaboration within a project. They constructed a network based on the sub-projects having bi-lateral and multi-lateral cooperation in their project reports. The proportion of the number of nodes and the edges in the network (i.e.the network density) represents the intensity of the collaborations within the project. They used this measurement to compare multiple projects for their intensity of collaboration. An interesting collaboration measurement model based on citation was proposed by Katz and Hicks [103]. In their research, they found that where the papers represent a home or domestic institution collaboration, the citation count increases 0.75 on average compared to a non-collaborative paper. However, if the paper involves international institutions, the citation count increases 1.6 on average. As a result, the different collaboration types can be quantified by the increased citation number, thus perhaps also reflecting the *quality* of different categories of collaboration.

2.1.3 Collaboration Model of Institutions

In academia, a model of institutional collaboration is based on projects that involve multiple institutions. Starting from the application of a grant, researchers from multiple institutions jointly submit applications for the project. If the application is successful, the project then becomes a point of collaboration for these researchers. The papers come out of the project is commonly affiliated to participating institutions. (Figure 2.1).

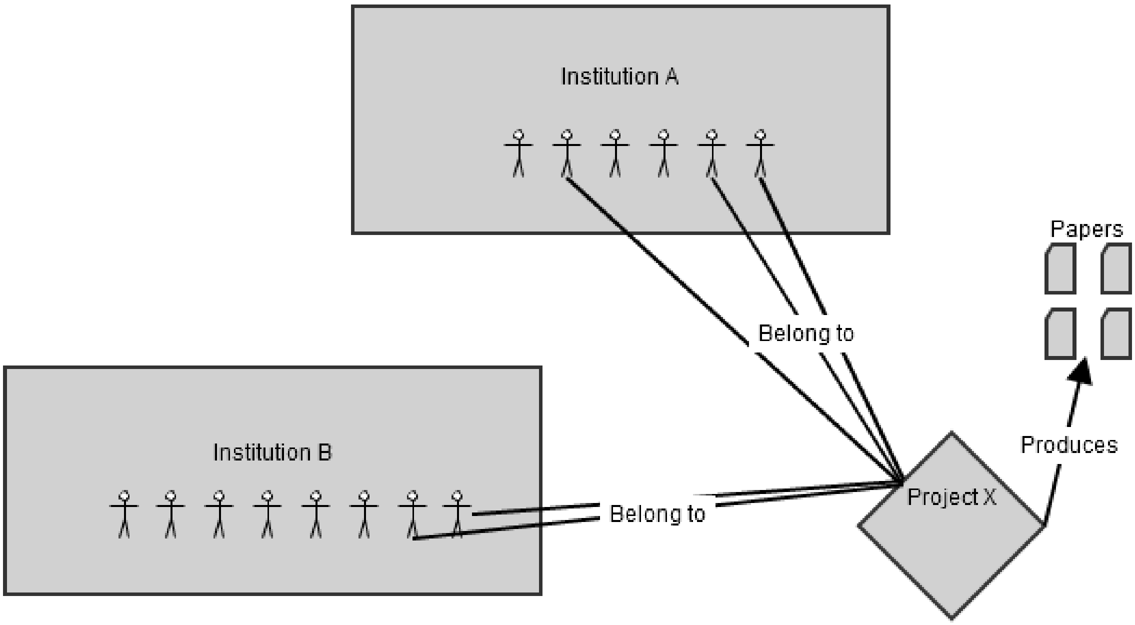


Figure 2.1: Institutional Collaboration Model. Researchers from two institutions form a project, which then produces papers signed by both institutions.

Based on this model, we propose a definition of institutional collaboration used in this thesis:

> **Definition** Two institutions are said to have collaborated if there is a published paper signed by both institution's researchers.

Using measures like co-authorship as indicators of degree of collaborativity is not new. Davidson Frame and Carpenter [52] used multi-country affiliations (two or more authors affiliated with institutions in different countries) as a metric for international collaborativity; [105] used multi-institutional affiliations (two or more authors affiliating with different institutions) as a metric for inter-institutional collaborativity. In Katz's data analysis, he realised that international collaboration, and inter-institutional collaboration need not be based on individuals. Nearly 12% of articles in his Web of Science (WoS) dataset list more institutions than authors, which indicates that researchers hold positions in multiple institutions. He took this to be evidence for top-down inter-institutional collaborations between the institutions that share the same researcher. About 2% of the papers in our WoS dataset contain more institutions than authors.

The measurements discussed so far are all at a macro level, which measures how many *times* the collaboration happens between the authors and institution. The more frequent the collaboration, the stronger the collaboration between the entities. These macro level methods assume that every collaboration is equal. This is not always the case. In recent years, there has been a growing realisation that collaboration is not well understood at the individual level, nor how it affects macro level collaboration.

2.1.4 Collaboration at the Micro Level

One line of work considers the collaborator's role and its impact on collaboration. Heffner [94] suggested the concept of 'sub-authorship', where sub-authors perform two primary kinds of roles: technical aid and theoretical aid. Technical aid provides technical support such as collecting data, operating machinery *etc.*. Theoretical aid provides assistance such as reading, editing or commenting on the research paper prior to publication. Each collaborator takes a different role in the collaboration, such that the research could not be completed without some of those roles. The strength of these kinds of collaboration cannot be easily quantified. Melin [129] had different views on the way researchers collaborated. He conducted surveys, interviewed researchers and concluded that there were two modes of collaboration. One mode is that the work is coordinated in all aspects and there are clear divisions of labour, which supports Heffner's finding; the other mode carries out discussion and idea exchange between the collaborators over many research questions, followed by an iterative process of research and re-writing, such that individual contributions are no longer identifiable in the final product.

Jeffrey [98] studied an inter-disciplinary research collaboration and found there were more problems than benefits. Communication is one of the biggest obstacles due to the different backgrounds, specialised words and so on. He also found that the group had to communicate using metaphors, story-lines and had to choose their words very carefully in order to be understood. This work suggests that at a micro level, not all collaborations are positive and bring benefit to the end result. There can be collaborations that bring more problems than they solve. Pravdić and Oluić-Vuković[155] suggested something similar in analysing collaboration at the individual level as well as the group level in

Chemistry. They found that productivity is affected by whom you collaborate with: collaboration with high productivity authors increases an author's personal productivity; however, collaborating with low productivity authors decreases it. This result confirms that not all collaborations are positive.

2.1.5 Summary of measuring research collaboration

Measuring degree of research collaboration is not a solved problem. Researchers have attempted novel ways to quantify collaboration, but due to the nature of human interaction, true collaboration strength can never be accurately captured. At the micro level, collaboration is a complex social phenomenon, researchers have different reasons to collaborate, and they have different roles in collaboration, so types of collaborations are best treated differently. At the macro level, when hundreds and thousands of collaborations take place between individuals and institutions, collaboration frequency gives us a measure of collaboration. Co-authorship has been widely used as a measure of strength of collaboration in the literature. Its advantages are listed and discussed along with its limitations. Building on the co-authorship, the strength of the higher aggregation level, *e.g.* departmental collaboration, inter-disciplinary collaboration, inter-institutional collaboration and international collaboration can be modelled.

2.2 Research Productivity

A well recognised definition of productivity (within the domain of management), according to Swiss [179], is the ratio between output and input of a system. The system can be an individual or an institution where productivity can be assessed. This implies that to accurately measure productivity in conducting scientific research, one has to identify and measure all the input which went into the research and all the outcomes resulting from the research.

Potential inputs for conducting research have been explored in the bibliometric field. These include, but are not limited to: expenditure, number of researchers, person-hours, length of time, and efficiency of researchers [70, 78]. However, many of these data are difficult to obtain, either because they are difficult to measure (person-hours, efficiency of researcher), or the data are generally not available (expenditure, number of researchers, person-hour). Without the data to measure input to research, previous studies tried to estimate using various indicators, based on the assumption that they are correlated with the input. These indicators include number of journal articles, books and citations [70, 71, 78]. The range of outputs of a research activity includes journal articles, conference papers, proceedings, patents, books, book chapters, book reviews, comments, prizes, licences, lectures and technical reports. Since the variables used to estimate the output overlap heavily with the output variables, the ratio between the two does not yield meaningful productivity data.

With the current data availability, precise calculation of research productivity by using the ratio between output and input is not possible in the present study, as was also the case in many prior bibliometric studies. Instead of using this ratio, research productivity is often measured as publication productivity [70]. A strong correlation had been demonstrated between the two [101, 121, 159].

However, publication productivity is not the same as research productivity, it only includes what is written, which is only a subset of research productivity (research conducted). The literature also points out its problems as a metric. Katz [101] found that different research fields put different emphasis on different publication types. For instance, social sciences pay more attention to publishing books while natural sciences mostly publish in peer reviewed journals. So to count only one type and not the other makes disciplines incomparable and comparisons inequivalent. More recently, Kyvik and Teigen [107] took an interesting perspective and argued that a good quality article should be treated as more productive (weighted more) than a lower quality one: where two researchers are both publishing ten papers in a year, the more productive researcher should be considered the one who has published in better journals (higher impact factor or other measure of journal quality) or has been cited more often. These limitations must be take into account when using publication productivity as a metric for research productivity.

2.2.1 Publication Attribution

The research process is complex, especially with multiple collaborators involved. The contribution of each collaborator is difficult to measure, and the actual contributions can vary significantly. On one hand, a collaboration can be the result of an individual doing most of the work, contributing the most, and so he should be weighted more for the credit; on the other hand, researchers can take different roles within collaborations, and any one of these roles may be important to the success of the piece of research, which sometimes justifies an equal weighting among these collaborators. The way to credit a co-authored paper to its authors is also an area of interest to the bibliometric research community. There are three established counting methods [33, 114, 183] depending on the author's affiliation:

Straight Count

Only the first author gets the credit and all the rest are ignored. Cronin and Kara [51] have shown that the straight count gives disproportionate credits to senior researchers, whose name often lists first while the junior researchers are completely ignored. This counting was used widely in the early years of bibliometric research, when electronic citation databases were not available. Using this counting method would mean that only the first author's institution gets the credit.

Full Count

All authors listed on the publication get 1 credit. Their institutions also get at least one credit[1]. The drawback of this counting method is that it exaggerates the contribution made by the heavily co-authored publication and weights the heavily co-authored publications more. For example, a publication with 10 co-authors gives 1 credit to each of the authors, giving a total credit of 10; while a paper with 2 co-authors only carries 2 credits.

Partial count

[1] The exact counting process in the literature is far from obvious. Some literature may have credited the same institutions with multiple credits due to multiple authors come from the same institution.

Sometimes referred to as am adjusted count, normalised count or fractional count, where the 1 credit is equally divided amongst all authors, so each institution gets a fraction of the one credit. The full count and partial count are frequently used in citation based performance research. The full count is mathematically simpler than the partial count, but it has the problem of double counting credit. The partial count method does not have this problem, the sum of all the credits equals the total number of publications in the dataset.

Different counting methods can give quite different results in subsequent analyses. Gauffriau *et al.*[79] compared the country research scores obtained by using different counting methods, and found score reduction as high as 72% when using partial count instead of full count. The full count particularly favours those countries that frequently participate in collaboration.

Publication counts as the productivity metric have also been applied at the institutional and country level. Katz [105] used the number of published papers as the university research output. Egghe *et al.* [61] used publication count to evaluate individual, department and institutional performance.

2.2.2 Summary of measuring research productivity

Due to the current limitation of accessing and measuring the inputs that went into the research, research productivity is difficult to measure precisely based on conventional productivity definitions. Alternatively, publication productivity is widely used as the approximation for the productivity of individuals and institutions. It should also be noted that at the institution level, using publication productivity alone without factoring out the input (such as number of researchers) gives an unfair advantage to larger institutions. Assigning co-authored publications to the contributing authors is not straight-forward. Bias may occur if an inappropriate counting method is used. Three publication attribution methods were discussed and their advantages and implications were presented.

Institutions conducting research do not only make quantitative contributions by publishing more; the quality of their research should also be accounted for. I will discuss how a piece of research is assessed for quality in the following section.

2.3 Research impact

The quality of a piece of research is multi-dimensional. It can be viewed from different angles: *e.g.* a quality piece of research may be the one that advances scientific understanding; or it may be one that makes the impact on the scientific community, generating a lot of debate and discussion; or it may be one that makes impact in the industry; or it may also be one that experts in the field recognise as such. There are broadly four perspectives in the literatures which refer to measuring the quality of a piece of research. They are: 1. methodological quality (how good is the idea and process), 2. reporting quality (how good is the write up), 3. quality based on peer review (how well it is recognised by the experts), and 4. bibliometric as a proxy (how well it is recognised by the community). This is often referred to as research impact. These perspectives are not mutually exclusive, *e.g.* an expert doing peer review may consider factors involving reporting quality. I give a brief introduction to 1-3, then I discuss research impact using bibliometric methods in detail.

Methodological quality

The methodological quality of a piece of research measures research by how well it follows the following research processes [82, 88] : significance of the research question, coverage of the literature, design of the experiment and whether the design of the experiment can in fact address the research question. But since these processes vary across disciplines, methodological quality measurement are more frequently used in certain disciplines, *e.g.* health [120], education [82] than others. There is no consensus on a specific set of standards that ensures research of high quality, and different disciplines may have their own definition for high quality research that can be very different from others.

Reporting quality

In published research, the report write-up is often evaluated for quality. Poorly produced reports that lack essential details to replicate the experiment often demonstrates a low quality research. In medical research, the degrees of freedom and P-values are sometimes absent[75]. When such core details are missing, the credibility of the result comes into question and the entire research, whatever interesting findings it may contain, becomes of little value and hence is considered lower quality research. To facilitate better reporting, several 'checklists' have been developed by various consortia for general and specific research design, *e.g.*,Consolidated Standards for Reporting Trials (CONSORT)[2], Quality of Reporting of Meta-Analyses (QUOROM)[3]. These checklists help authors to decide what needs to be included in the report, but they are not an evaluation instrument.

Peer review

Peer review is the process of asking experts' opinion on the quality of a piece of research. The report of the research is sent to the experts to read, and they give a rating and comment on the research. Peer review is an established method and is widely used in research journals and research conference article evaluation. Peer review is able to provide immediate quality indicators for a piece of research, unlike quantitative metrics (*e.g.* citations) that may take months or years to accumulate. Peer review is often referred to in the literature as the preferred process of evaluating research when possible[34, 89].

However, to conduct peer review is expensive. Firstly, groups of experts covering the entire subject area must be employed; secondly, reading and reviewing articles is a very time-consuming process; finally, an article needs to be reviewed by several experts in order to calculate a fair rating, adding a greater work load. Peer review is a common practice for journals and conferences to rate and select quality papers. It is because they are in very specific subject area and there is no problem to find experts (peer reviewers tend to be the participants themselves).

Peer review can be subjective and contain personal opinions on the quality. Experts, especially rivals, may have strong disagreement on certain subject, which can potentially

[2] http://www.consort-statement.org/statement/revisedstatement.htm

[3] http://www.consort-statement.org/QUOROM.pdf

lead to unfairness and abuse in the peer review process. As a result, articles' quality rating based on peer review can become non-subjective and non-reproducible.

2.3.1 Research Impact using Bibliometric Methods

I have discussed three of the four perspectives evaluating the quality of a piece of research, ranging from methodological quality, (evaluating whether the *process* of the research is up to standard); reporting quality (a specific peer review that focus on evaluating whether the content of the research report is up to standard; and expert review (experts in the area evaluating the quality of piece of research of a piece of research for publication or research performance review). However, these are qualitative methods and are very expensive to execute. In this section, I introduce bibliometric method, which is based on quantitative bibliometric data, such as citation data to indicate the impact of research. These data are recorded in the scientific research process and access to these data is becoming easier.

Authors cite the research they use and discuss, and these citations can be counted. In the 1960s, Garfield [76] introduced an electronic citation database, where the paper citations were harvested from each paper and indexed. Using his database, citation has been studied widely. Although Garfield warned against it, citations have also been used as quality metrics. The rationale is that at large scale, researchers' citations are most likely to be positive responses to previous work. Compared to the other three methods, bibliometric methods have data recorded and collected in a public way, so that anyone with access to the bibliometric data is able to reproduce the same measurement, making the metric more objective than subjective evaluations.

Citation counts are often taken as a proxy for the impact of earlier research on later research [30, 126]; citation counts are hence referred to as citation impact. Citation impact and citation frequency have been used as an indicator of quality in many previous studies [76, 112, 113].

However, Goldfinch [87] argues that an increased number of citations can also be the result of an author's having a larger social network, which in turn increases his visibility and his citations, so it not necessarily reflects impact. Citation has been compared with peer review and generated intensive debates. I discuss them in 2.3.2. With the potential limitations of citations in mind, I present next the various uses of citations in measuring impact.

2.3.1.1 Citation Count

Citations count means the total citations an article has received to date. This is one of the most widely used article impact measures due to its availability and simplicity. Various databases publish data with citation counts included. However, there are problems. Firstly, due to the difference in the year of publication, the longer the articles have been published, the more time they have to accumulate citations, so unfair advantage is given to articles which were published earlier in any year. The other problem is that this count treats all citations equally. A citation from a highly cited paper should be worth more than a citation from a never-been-cited article. The impact of each individual citation has been omitted. It is also important to note that across disciplines, due to differences in citation practices, some disciplines can have short publication cycles resulting in a faster rate of citation, accumulating more citations than other disciplines. Citation measurements cannot be directly compared without normalisation.

Author Self-citation Frequently, authors cite their own work. According to an analysis by Arsnes [6], in a three-year citation window, author self-citation in certain disciplines can reach as high as 36% of total citations. Baldi and Hargens [14] showed that author self-citation is relatively infrequent in general, but quite frequent in certain disciplines.

Authors cite their own work for different reasons, for example, due to the cumulative nature of research; the need for personal gratification; or as a struggle for visibility and scientific authority in the community [27, 68].

Self-citation can inflate the metrics based on citation. Fassoulaki *et al.*[66] report that self-citation can significantly increase a journal's average citation count per article.

Schreiber [166] showed that self-citation increases h-index[4]. As a result, it is a common practice to remove self-citations in the source data before analysis begin.

2.3.1.2 Windowed Citation Count

Citation windows are used to address the first problem. Instead of counting all the citations an article has received to date, a window-period after an article's publication is used. Only citations received within this window period are counted for the article concerned. WoS impact factor calculation adopts a two-year window. Wider window sizes were also investigated. He *et al.*[93] compared a two-year window with a three-year window for estimating article impact and found no difference; Campanario [38] compared journal impact factor based on a two-year citation window with a longer five-year window and found the two behave very similarly. Wang [192] conducted a larger analysis by using 30-years worth of WoS citation data in multiple disciplines that the more frequently cited the less the correlation between the windowed citation and the eventual citation count. Windowed citation would not be a good estimation for those highly cited articles.

2.3.1.3 PageRank Weighted Citation

As it was mentioned in the citation count's problem, a refinement of the citation count is to consider the impact of the citers. In addition to the number of citations a paper has received, the citing paper's citation should also been taken into account. This idea of the iterative measurement is commonly exercised in the Web. The ranking algorithm – PageRank – used by Google search engine is specifically designed to calculate this iterative score of the linking documents. The Web can be viewed as a set of documents with links pointing to other document. Page *et al.*[151] found that a popular web page would be the one not only linked to by many other web pages, but also linked to by heavily linked web pages. An iterative algorithm was proposed by them that would work on a directed-graph like the web, where each link between the web pages has a direction of source and destination. A score is calculated by the algorithm for each web

[4] A weighted joint indicator of citations and productivity for measuring researcher quality [95].

page, which determines the quality of the web page. Like web-pages linking each other, citations link citing articles to the cited articles. By connecting these articles using the citation link, a directed-graph like the web can be constructed, therefore, the PageRank algorithm can be applied [24, 127]. With PageRanked citation, the articles' measured impact is not always the more citations the better, when the citer's citation is high, it will be counted more towards the impact of the article.

2.3.1.4 Network-based Centrality Measures

Another variation of using citation as impact measurement adopts social network techniques. These methods treat citations as links in a network, similar to the PageRank method; but instead of considering each citation as a weighted 'vote' to an article, citations are considered as information flow in a network [23]. If article A cites article B, then information flowed from B to A. The *centrality* of the articles is calculated based on their position in the network. This includes *degree centrality*, which is how many articles have cited it (this coincidentally has the same meaning as the original citation count); and *closeness centrality*, which calculates the mean distance to all of the rest of the articles in terms of number of citations. The article with the greatest closeness centrality is the one that has the *shortest mean distance* to all other articles. Such an article can be a good starting point for a literature research in the domain. Finally, *betweenness centrality* measures article importance in terms of information flow. The article with the greatest betweenness centrality would be on the shortest path between the most pairs of articles. This may indicate that this article channels through the most information in the citation network.

Applying social network analysis techniques in citation metrics is interesting and opens a new perspective on viewing citation, but it lacks the theoretical support and empirical evidence of the robustness of the methodology, so I will not use them as a way to approximate an article's impact in this study.

2.3.1.5 Journal Impact as an Indicator of Article Impact

A journal's impact is often used as a proxy to the impact of its articles. In particular, WoS's impact factor is sometimes directly used to express the impact of the articles [93]. This may at first sound odd and has obvious flaws, but the argument is that if an article is accepted by a journal, the article is at least up to the standard of the journal. This is in fact the same logic used when graduates are judged by the university they graduated from rather than by the level of achievement they graduated with. Still, this practice is frequently criticised and discouraged. The obvious problem is that it treats all the articles in the journal as having the same impact, which is certainly false. In addition, several studies [3, 39] show that the citation distribution of articles published in the same journal follows a power law. That is, the top few articles in a journal receive the majority of the citations, while the rest of the articles receive the remaining few citations. Using an arithmetic mean citation of a journal adopted by WoS impact factor to describe all the articles is inaccurate. Journals evolve over time, their impact factor varies year by year, taking a snapshot of this continuous changing impact factor to describe the articles is also inappropriate. WoS impact factor is also well known for its inability to prevent manipulation. Many previous studies [4, 11, 133] have investigated and listed strategies some journal publishers use to increase their impact factor dramatically (e.g., requiring authors to prior research published in the same journal).

2.3.1.6 Reading Statistics and Download Logs

With the wider adoption of institutional repositories and the availability of website article download data logs from these repositories, researchers have studied whether the paper access data recorded by repositories offer any new information for paper impact estimation. The primary difference with these log data is that they are generated by potential readers (people that have accessed the paper's abstract page and/or downloaded the fulltext) rather than actual readers. This data is one stage ahead of the citation data. People citing the article must have already downloaded the article, while people who have downloaded the article may not necessarily cite the article. As a result, these data are gathered much earlier in the scientific publication cycle and are hence faster compared to citation. Generally, citations only become available after at least one year has elapsed since the article's publication due to the long publication cycle. By contrast download and reading data are generated and collected as soon as the article becomes available. Bollen *et al.* [23, 25] studied the effect of reader generated statistics and showed how it relates to citation count. Brody and Harnad [35] investigated how much variation is indicated by the early download statistics to the citation count. They found a significant ($r = 0.4$) correlation between the download counts and the eventual citation counts already in 6 months of download data.

2.3.2 Debate between peer-review based metrics and bibliometrics based metrics

Two categories of measurements are frequently discussed in the literature. Maier [125] tested the correlation between experts' opinions (based on survey) and journal impact factors in regional sciences, she found that the expert opinion is more close to the true reputation of the journal than the impact factor, and the impact factors did not correlate with experts' opinion and sometimes even correlated negatively. Brinn *et al.*[34] conducted a survey and found that senior researchers consider peer reviewing more important in measuring the scientific performance than metrics based indicators.

van Raan [187] used 147 university chemistry research groups in the Netherlands and found a significant correlation between peer review and citation based metrics (they used *h*-index and crown indicator[5] as the citation based metrics.). Opthof and Leydesdorff [148] referred to Van Raan's work and raised concerns for their crown indicator, which was a starting point for a series of debates in the literature. The contributors to this debate include Waltman *et al.* [190, 191], Opthof and Leydesdorff [149], Bornmann [31], Gingras and Larivière [83], Moed [130] and Taylor [181] arguing whether the data really showed significant correlations between the peer-review and citation based metrics; whether the normalisation method has affected the result; whether an alternative metrics (such as crown indicator) is appropriate in scientometrics.

[5] They define the crown indicator as the ratio between the citations per paper and the field based average citation

Harnad [91] on the other hand found significant correlations between peer review (based on the UK RAE 2008 research assessment exercises) and citation metrics as well as new indicators such as downloads. He pointed out that these measurements for research performance quality are not "face-valid", because none of them are direct measure of research quality. Correlations only reveal that the pair have underlying common factors.

The main difference between the two types of measurements is the way the impact indicator is based. Citation metrics are based on quantitative statistical data – how many times the work has been cited, how frequently it has been cited and so on. These citation data are recorded in journals, repositories and databases. Thus the calculated impact indicators are reproducible given the same dataset. However, it does require years for the citations to accumulate before the data are useful.

The expert-review based metrics is a qualitative method. Compared to citations, expert-review can be done as soon as the work has been published; there is no need to wait. However, it is based on a few experts' opinions and their personal experience in the domain, so it is more likely the result is biased and none reproducible in the future study with a different group of experts reviewing.

So far, we have looked at how a piece of research can be judged for its quality. Using articles as the building block, we would like to measure an aggregated research impact. In the following sections, we explore how WoS Journal Impact Factor is calculated; how a conference quality can be estimated based on a set of papers. We also review National Research Evaluation Exercises carried out in two difference countries and learn what factors these exercise have been taken into account to quantify research quality for the institutions.

2.3.3 National Research Evaluation Exercises

The UK and Australian governments have conducted research assessment nation-wide in the past. Both of their evaluation methodologies are based on expert review of the published work that the institutions submit for assessment.

2.3.3.1 UK

The UK research assessment exercise (RAE) was conducted to evaluate the overall quality of UK research output across a 6-year interval. Its aim was to measure the country's research performance by institution as well as by discipline, hence justifying the investments in research and to serve as the basis for future research funding. It had happened four times in the past 20 years. The latest published assessment result is RAE2008. The next iteration will be called Research Excellence Framework (REF) and the result will be published in 2014.

Method: Institutions are invited to submit a profile for each Unit of Assessment (UOA) they would like to be assessed in. There are in total 67 UOAs attempting to cover all government funded research. The submission includes the research profile of the institution, as well as up to four items of research output per researcher submitted to the UOA. These submissions are evaluated by a panel of experts, the research outputs are then rated according to the rules. There are five rating categories, ranging from "the works make significant or substantial contribution to knowledge" to "the work falls below the standard of nationally recognised work"[152]. The scores of each output are combined to produce an institution's overall score for the discipline. The combined percentage of the measurements is as follows: 70% based on citations, 20% based on research environment and infrastructure, 10% based on esteem and impact (measured by the recognitions of the researchers and membership of funding bodies *etc.*). The RAE2008 submission and results can be downloaded from its website (http://www.rae.ac.uk/).

2.3.3.2 Australia

The Excellence in Research for Australia (ERA) is an initiative taken by the Australian research councils to evaluate their higher education research performance and to allocate funding. They started ERA in 2007[10], replacing the Research Quality Framework (RQF) used previously between 2003-2006. The latest published result is for 2012 at the time of writing.

Method: The evaluation process involves data collection and evaluation. Data collection is done by the Higher Education Research Data Collection (HERDC) scheme, which collects data on Australian universities' research output and income annually. The research data is split into eight disciplines for each university, and each piece of research outcome is measured by panels of experts on the following four aspects: 1. research quality, which is based on citation, peer review and research income; 2. research volume, which is based on research output and research income; 3. research applications (*e.g.* commercialisation income) 4. research recognition. The scores for each of the eight disciplines are listed separately for the universities and the ERA specifically advises that it is not possible to add or average scores from the ERA report to derive overall rankings for universities.

2.3.4 Journal Impact

Publication is an important channel through which research results are disseminated. The main choice of many of the disciplines is to publish in peer-reviewed research journals. In the following sections, we review a few widely used methods for evaluating journals' impact.

2.3.4.1 Impact Factor

The journal impact factor is one of the established ways to measure the importance of a journal. For journals that are indexed by the Journal Citation Report (JCR), the impact factor is published annually. Due to citation practices and size of the field, the impact factor of journals is only comparable within the same field. Normalisation methods exist to enable cross discipline impact factor comparison [192].

The use of impact factor is subject to various criticisms. Firstly, due to the distribution type of the citations to articles in a journal, there is no average number of citations to the articles. So calculating an algorithmic mean as the average citation number is not accurate. Secondly, the citation accumulation pattern for original research papers and review papers is very different. Review papers tend to attract more citations than research papers[187]. Mixing them without normalisation would skew the measurement of impact. Finally, the journals included in the citation index used by the calculation are primarily in English. It poses a language bias for impact.

Impact factor is an application of the citation-based metric. Impact factor is calculated by averaging the citations received in the measuring year for the articles published in the previous two years. In the next section, I review the two journal quality lists published by UK and Australian governments that use a hybrid of expert review and citation based metric.

2.3.4.2 Acceptance Rate

When deciding in which journal to publish results, researchers do not just choose any journal. In order to give their research the maximum coverage, they tend towards a high impact journal. Every researcher would like to publish in a high impact journal, but the space in an issue of a printed magazine is fixed. So the more submission a journal receives, the more they have to reject, to make room for those that can attract the most readership and citations, resulting a low acceptance rate (or high rejection rate). A potential metric of measuring journal impact would be the journal's submission and rejection ratio. Some journals publish their rejection ratios, but many do not. The bigger problem with this metric is that it can be easily manipulated by the publisher.

2.3.4.3 ABS's Journal Quality Guide

The Association of Business Schools (ABS) in the UK publishes a quality guide on business and management journals. The purpose is to give researchers authoritative information. The guide categorises journals into four quality levels and the latest version 3 was published in March 2009. This guide includes more than 800 journals, and evaluates them based on factors including expert opinion, impact factors and citation based metrics. The coverage of this guide is wider than the citation indices (as many business journals are not presented in any of the citation databases). The method used and the detailed score for each journal is unfortunately not published.

2.3.4.4 ABDC Journal List

The Australian Business Deans Council (ABDC) publishes a journal ratings list that serves a purpose similar to that of ABS's guide. It reviews over 400 journals covering 17 disciplines. The criteria the journals are judged on are:

- Relative standing of the journal in other recognized lists (such as the Association of Business Schools)
- Citation metrics
- International standing of the editorial board
- Quality of peer-review processes
- Track record of publishing influential papers
- Sustained reputation
- Influence of publications in the journal in relation to hiring, tenure and promotion decisions.

The output of this list also categorises journals into four quality levels. However, as in the ABS's list, the process of how each of the criteria is judged is not available.

2.3.5 Conference Quality

One of the other important channels to disseminate research results is conferences. This is especially the case in fast developing disciplines that require rapid communication of research outcome, such as Computer Science. In these disciplines, top conference publications often carry as much influence as the journal publications.

Depending on the timing of the quality calculated – before or after the conference has taken place, research splits into pre-conference and post-conference. Less data is available for a pre-conference quality prediction than a post-conference quality measurement. For instance, the citation count of papers will only available after a conference has taken place. Pre-conference quality prediction is becoming more useful as more conferences are being organised year on year, researchers would like to know the quality of the conference before attending it. Zhuang *et al.*[203] took an approach by mining and analysing program committee members of the conference. They proposed a number of heuristics to identify reputable conferences, for example, the number of program committees, the committee's average publication, the committee's average co-authors and centrality measures of the committee. By combining these factors, they were able to identify top conferences and the extremely low-quality ones before the conference took place. Taking it one step further, Souto *et al.*[175] developed a classification model for Computer Science conferences. Using a conference ranking, the system was able to support semi-automatic evaluation of the quality of Computer Science conferences.

As for the post-conference quality measurements, Martins *et al.*[127] used article citations as the basis for their conference assessment. Starting from the citation analysis of articles published in conferences, their metric also takes into consideration conference specific characteristics such as longevity, popularity, prestige and periodicity. Yan and Lee [199] proposed an approach to evaluate conferences based on the assumption that top papers in a discipline are often recognised. Using these top papers, the co-authors of these top papers are identified. To determine the quality of a conference, they then look for the conferences that these authors regularly attend and publish papers at. However, the top papers in a discipline are often very limited and confined, so they do not give a good coverage of conferences.

2.3.6 University Rankings

Universities are the largest producers of research output across the globe. More than half of the research output recorded in WoS and ACM is from universities. Although in the present study research quality is the prime concern, the overall quality of a university is often covers teaching, reputation and many other factors. Below is a list of potential factors that affect a university's perceived quality.

- Historic reputation

- Publications
- Citations
- Visibility
- Teaching
- Student feedback
- Student prospects
- Resources, funding, infrastructure

Various ranking and evaluation exercises can only consider a subset of the above list due to the data availability. Data availability is also a limiting to the current study.

The economic development of a country has a strong impact on the role of the higher education system in the country. As a result, the word "university" may not mean the same type of institution in all countries. Some universities may be set up solely to do teaching, while others put more weight on research. Research-oriented institutions and teaching-oriented institutions should be treated differently, if they can be distinguished.

Based on the Webometrics ranking of world universities, there are currently more than 10000 university level higher education establishments in the world. The country with the most is US, which has 2,830, followed by China(702) with only a fraction compared to the US. The strong presence of US institutions in higher education put it far ahead of other countries in terms of research output.

van Raan's work [185] forms the basis for many recent ranking methodologies. He has raised and discussed the problems of making measurements (which include statistical issues), issues in indicator choices, language bias, timeliness and variations between different research systems. Moed [131] applied bibliometric analysis in ranking universities. He used the articles published in the Web of Science database, calculated indicators that measure the universities' quality, including the article output rate, the percentage of internationally co-authored articles and the percentage of industrial co-authored papers.

The advantage of using publication data is that the resultant ranking is free of personal opinion. In this respect, van Raan [186] firmly supports metric based methods. He has criticised expert-review based ranking, showing that no significant correlation can be found between expert opinion and bibliometric outcome. In more recent papers [188, 189], he showed that papers published in non-English language would decrease the ranking of the university. This reveals that the world research arena is predominantly in English.

In recent years, ranking universities has become increasingly popular. There are a few university rankings published every year by various organisations across the world. I am going to review four of the most commonly used lists amongst students and scholars: the Webometrics ranking of the world universities, the Guardian university ranking, Academic Ranking of World Universities (ARWU) published by Jiao Tong University and non-profit organisation 4icu's world university ranking.

2.3.6.1 Webometrics Ranking of World Universities

The Webometrics Ranking of World Universities (Webometrics) is an initiative of the Cybermetrics Lab, a research group belonging to the largest public research body in Spain (CSIC). The aim of the ranking is to promote Web publication, support Open Access initiatives, and support electronic access to scientific publications and other academic material. These are strongly reflected in its methodology. We introduce the methodology of the 2010 version, which we have the data for, the method has changed in the current version (as of 2015).

Methodology Webometrics ranking (2010 version) [5] uses data indexed by popular search engines as the basis of ranking. Four indicators are used and their weightings are as follows:

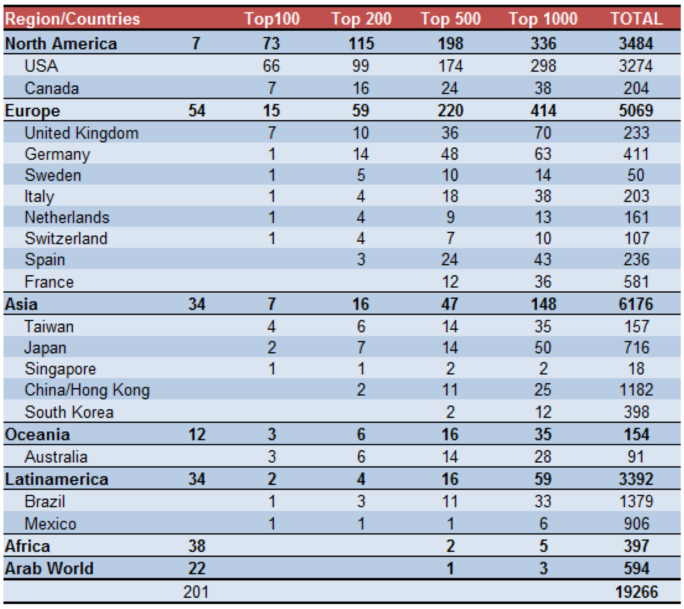
- Size. Number of pages recovered from four engines: Google, Yahoo, Live Search and Exalead. (20%)

- Visibility. The total number of unique external links received (inlinks) by a site (Yahoo Search only). (50%)
- Rich Files. The number of academic related files retrieved from the domain. Files include: Adobe Acrobat (.pdf), Adobe PostScript (.ps), Microsoft Word (.doc) and Microsoft Powerpoint (.ppt). The data was extracted using Google, Yahoo Search, Live Search and Exalead. (15%)
- Scholar. Number of papers and citations reported by Google Scholar for each academic domain. (15%)

Coverage The design and weighting of the indicators used by this ranking are entirely based on the Web. Therefore, any university or college with web presence is analysed and ranked. This immediately overcomes the university identification problem [131, 186], which causes difficulties in ranking universities across countries, enabling impartial analysis and ranking of universities from purely the web perspective.

The Webometrics ranking covers almost all universities in the world. The data is a very useful resource to learn about the world universities in less well known countries. Table 2.1 shows the coverage of the ranking by continent and country. The web domain name for each university is included. The domain name was used to obtain the online resources (number of webpages, academic files etc). It was also used as the university identifier in this study.

Shortcoming Because universities are identified by the web domain, those with more than one web domain, or those with domain name changed (*e.g.* Imperial College London), the ranking is based on each individual domain only, no aggregation was performed to concatenate multiple domains. Therefore, a lower than expected ranking is expected for these universities.



Region/Countries		Top100	Top 200	Top 500	Top 1000	TOTAL
North America	**7**	**73**	**115**	**198**	**336**	**3484**
USA		66	99	174	298	3274
Canada		7	16	24	38	204
Europe	**54**	**15**	**59**	**220**	**414**	**5069**
United Kingdom		7	10	36	70	233
Germany		1	14	48	63	411
Sweden		1	5	10	14	50
Italy		1	4	18	38	203
Netherlands		1	4	9	13	161
Switzerland		1	4	7	10	107
Spain			3	24	43	236
France				12	36	581
Asia	**34**	**7**	**16**	**47**	**148**	**6176**
Taiwan		4	6	14	35	157
Japan		2	7	14	50	716
Singapore		1	1	2	2	18
China/Hong Kong			2	11	25	1182
South Korea				2	12	398
Oceania	**12**	**3**	**6**	**16**	**35**	**154**
Australia		3	6	14	28	91
Latinamerica	**34**	**2**	**4**	**16**	**59**	**3392**
Brazil		1	3	11	33	1379
Mexico		1	1	1	6	906
Africa	**38**			**2**	**5**	**397**
Arab World	**22**			**1**	**3**	**594**
	201					**19266**

Table 2.1: Coverage of Webometrics Ranking

2.3.6.2 Guardian Ranking

The Guardian university ranking is published annually by The Guardian, a respected news agency in the UK. The primary aim of the rankings is to inform potential applicants about UK universities.

Methodology The Guardian's ranking uses eight indicators, which all focus around the student. The indicators and their weightings as follows:

- Overall quality - National Student Survey (NSS) overall results (5%)
- Teaching quality - NSS feedback section rated by graduates of the course (10%)
- Feedback - NSS feedback section rated by graduates of the course (10%)
- Spending per student (15%)
- Staff/student ratio (15%)
- Job prospects - proportion of graduates who find graduate-level employment or study full-time within six months of graduation.(15%)
- Value added score - comparison of students' degree results and their entry qualifications (15%)

- Entry score (15%)

Unlike many other rankings, this ranking does not include a measure of research output.

Coverage This ranking considers all listed UK universities and ranks 110 (70%) of them that they have data for. This is a domestic ranking, no universities outside the UK are ranked. 46 major subject areas are separately ranked. The subject ranking uses different weightings on the indicators. In addition, the size of the department for the subject is considered before including the universities in the subject ranking.

Shortcoming This ranking is based the taught student data, no research performance is taken into account. It is an UK only ranking, no international universities are included.

2.3.6.3 Academic Ranking of World Universities (ARWU)

Academic Ranking of World Universities (ARWU) is an annual ranking since 2003, published by Shanghai Jiao Tong University, China. The original goal of this ranking was for the Chinese education authorities to learn the academic research positions of Chinese universities. It since became one of the widely used international university rankings. It is also a very controversial ranking due to the metrics considered. Many papers are published specifically to criticise them [22, 67].

Methodology The ARWU ranking considers four aspects of academic and research performance, including:

- Quality of Education. Alumni of an institution winning Nobel Prizes and Fields Medals. (10%)
- Quality of Faculty. Staff of an institution winning Nobel Prizes and Fields Medals Award. (20%)
- Quality of Faculty. Highly cited researchers in 21 broad subject categories. (20%)

- Research Output. Papers published in Nature and Science. (20%)
- Research Output. Papers indexed in Science Citation Index-expanded and Social Science Citation Index. (20%)
- Per Capita Performance. The weighted scores of the above five indicators divided by the number of full-time equivalent academic staff. (10%)

Coverage The ARWU ranking includes all universities that have any Nobel Laureates, fields medallists, highly cited researchers, or papers published in Nature or Science. In addition, universities with a significant number of papers indexed by Science Citation Index-Expanded (SCIE) and Social Science Citation Index (SSCI) are included. The coverage of the university is limited, with only 500 universities ranked across the globe. The top 100 universities are ordered, the remaining universities are put into 101-150, 151-200, 201-300,301-400 and 401-500 five ranking groups, where universities are not ranked within each group.

Shortcoming There are many critiques in the literature, the main ones are the following:

- The use of Nobel prizes and Fields medals winners – The researchers awarded the prize don't necessarily conduct the work in the current university.
- The choice of the 21 subject domains is biased towards medicine and biology.
- The weighting of the authorships on the Nature and Science papers is debatable, where 100% goes to corresponding author, 50% for the first author, 25% for the next author affiliation, and 10% for other author affiliations.
- Limited to two citation databases only(SCIE and SSCI), and there is no evidence to show that they have understood the impact of such limitation.

Additionally, the coverage of the university is very small, and is therefore unsuitable for use in our study.

2.3.6.4 World University Ranking By 4 International Colleges & Universities

4 International Colleges & Universities (4ICU) is a not-for-profit organisation reviewing accredited Universities and Colleges in the world. The aim of the website is to provide an approximate popularity ranking of world Universities based upon the popularity of their websites. They intend to help international students and academic staff to understand how popular a specific University is in a foreign country. The ranking is published on the website every 6 months in January and July.

Methodology The ranking is entirely based on the data gathered from the web. Three independent search engine metrics are used: Google Page Rank, Yahoo Inbound Links, Alexa Traffic Rank. The unique inbound links and traffic to the university domains are counted and reviewed separately for the three search engines. The result is then combined to give a final value for ranking. However, the exact process and formula is kept secret due to copyright and to minimise manipulation attempts.

Coverage 4icu.org claims to include around 10000 Colleges and Universities in 200 countries, but they only publish the top 200 universities in the world. Continent and country ranking lists are also available. They list the top 100 universities.

Shortcoming This ranking is not based on actual academic or research performance, it merely ranks a university website's popularity. Only the top universities are listed, which is too few to be used in this study.

2.3.6.5 Compare and Contrast

These four rankings vary largely in the data used to rank the institutions. The Guardian's ranking is purely based on the data collected regarding undergraduate student experiences; no research performance is considered; Webometrics analyses the documents and traffic within the institution's web domain, assuming that all evidence of teaching and

	Webometrics	Guardian	ARWU	4ICU
Webometrics		r=0.58 (p<0.01)	r=0.32 (p<0.01)	r=0.02 (p<0.39)
Guardian			sample too small	r=0.61 (p<0.01)
ARWU				r=0.52 (p<0.01)

Table 2.2: Pearson Correlations between the four rankings. r is the Pearson Correlation, p is probability significance for the correlation.

researching would be demonstrated on the web. ARWU ranking, although receiving a lot of criticism, is the one that has considered the largest number of factors ranging from teaching, size, research, prestige and impact of the university. 4ICU ranking is an institution's website activity ranking.

Table 2.2 shows the correlations between each of the reviewed rankings in 2010. The Guardian ranking shows a significant and middling correlation with both web-based Webometrics and 4ICU rankings. But the two web-based rankings – Webometrics and 4ICU – do not show any correlations between one another. The ARWU ranking shows a small but significant correlation with both Webometrics and 4ICU. The correlations between ARWU and the Guardian ranking are not calculated due to the small number of overlapping institutions. The Guardian only ranks UK universities, and apparently, according to ARWU ranking, there are only 4 UK universities in the top 100, resulting in a small overlapping sample size between the two.

The correlation between the rankings is significant, but not very high. A middle correlation coefficient (0.3-0.6) between the two rankings means that there are factors one ranking considers, but the other does not.

2.3.7 Summary of measuring research impact

This section explores various established methodologies used to measure research impact using research publications. The two common methods – expert-review based and citation based – were discussed and compared. Higher aggregation levels, *e.g.* journal level, institution level and country level adopt different approaches to measure impact. These approaches were presented and discussed. In addition, four popular university rankings were presented, their methodology as well as their ranking were compared, contrasted and correlated.

With the measurements of research collaboration, research productivity and research impact established, I will move on to investigate the relationship between these three research activities.

2.4 The Relationship Between Research Collaboration and Productivity

Understanding research collaboration and its impact on productivity has raised much interest in the past. Although collaboration metrics and productivity metrics vary from one study to another, findings indicate that research collaboration positively correlates with research productivity.

In 1966 Price and Beaver [159], in one of the earliest studies on this topic, found that the number of collaborators was positively correlated with the number of articles published by the author. By qualitative analysis, they found that the most prolific researcher also tends to be the most collaborative, and that 3 out of 4 of the next most prolific persons are amongst the next most frequently collaborating persons. No causal analysis was performed.

A year later, Zuckerman [205] interviewed 41 Nobel Laureates in science disciplines, and identified a strong relationship between collaboration and productivity. She found that laureates published more papers and were more willing to collaborate than a matched sample of scientists.

Pravdić and Oluić-Vuković [155] used research data collected from Chemistry, and found that the number of papers published is dependent on the frequency of collaboration among the authors. After they had interviewed a sample of the authors, they also learnt that collaboration with highly productivity authors increases personal productivity while collaborating with less productivity decreases it.

Glänzel and Schubert [85] considered collaborations from three aggregation levels: individual level co-authorship, cross-country co-authorship and multi-country co-authorship. In all three levels, co-authorship is positively correlated with collaboration. The same positive correlation was found by Adams *et al.* [2] between the size of the collaboration groups and scientific productivity.

Lee and Bozeman appear to have found something more subtle. In their 2003 research report [33], they used a regression model to determine whether the explanatory power of collaboration was diminished by factors such as job satisfaction, rank, age, gender *etc.* They surveyed and interviewed 443 academics to obtain their data and then completed the regression. They then concluded that despite the extra variables they had included, the number of collaborators remained the strongest predictor of the number of publications. However, in a later paper by the same authors [114], they extended the journal paper and book counting method to include partial count, in addition to the full count they have used before. The full count method counts a collaborated item as many times as the number of co-authors listed, while the partial count split a collaborated item by the number of co-authors. While they still found the number of journal papers strongly and significantly correlate with the number of collaborators, they could not find the same correlation using 'partial count'. Different counting method can lead to different result, it is important that the appropriate counting method is used.

More recently, Defazio *et al.* [55], using the EU framework programme to study similar variables in Chemistry, found that researchers tend to collaborate just to secure funding, the impact of funding on productivity is positive, but the impact of collaboration on productivity is weak. By splitting the period into pre-funding, during-funding and post-funding periods, they found that collaboration during the funding period does not correlate with productivity; in post-funding period, although the collaboration count decreases compared to the other two periods, however, it has a strong positive correlation with productivity. So it appears that the connections that the researchers established pre-funding and during-funding went on to have a positive effect on subsequent research output.

At a higher aggregation level - institution level, Abramo *et al.* [1] studied collaboration intensity and productivity (normalised by number of staffs). While the correlation varied substantially among different research areas, a strong correlation was found in information engineering.

The research discussed so far was all based on cross-sectional data, making cause-effect inferences un-testable. He *et al.* [93] constructed a longitudinal dataset of 65 New Zealand researchers for 14 years. Among other findings, they claimed that international collaborations are positively related with future research output. Although they could not find any significant correlation with future output for within-university collaboration and domestic collaboration.

The positive relationship between collaboration and productivity has been confirmed previously at the individual researcher level. I will expand these studies and report the results of testing the same correlation at the institution level.

2.5 The Relationship Between Research Collaboration and Impact

The relationship between collaboration and research impact is more studied than the other two pairs. The effects of collaboration on impact have been studied from various angles, for example, different types of collaboration (e.g domestic or international collaborations) have different effect on impact; time factors of the collaboration (how the impact of the collaboration changes over time) and regional variations of the collaboration (how the relationship vary depending on country and region).

Collaboration is often measured using co-authorship. Presser [156] splits papers into co-authored papers and singly authored papers (non-collaborative papers), and compares editorial decisions with respect to these two categories of papers. Statistical analysis based on more than 200 papers submitted to a journal showed that collaborative papers were considered “less bad” than the non-collaborative papers. Beaver [19] reached a very similar conclusion analysing multiple authored papers by citation counts: singly authored papers are slightly more likely never to be cited than collaborative ones.

2.5.1 Types of collaboration

Different types of collaboration have different relationships with the research impact. The most studied types include international collaboration, domestic collaboration, inter-institutional collaboration and intra-institutional collaboration. Narin *et. al.*[137] found that internationally collaborated papers were cited twice as heavily as papers authored by a single scientist in a single country. 65 Biomedical scientists in a New Zealand university were closely investigated by He *et. al.* [93]. They found intra-institutional collaboration and international collaboration significantly and positively correlated with article's citations. However, they could not find the same correlation with domestic collaborations. In a different study, Didegah and Thelwall [56, 57] attempted to find the determinants of high impact research. Amongst a range of other factors including impact factor of the publishing journal, document properties (abstract readability, abstract length, document length, keyword count etc), cited reference's impact factor, they found individual and international collaboration give a citation advantage in Biology and Biochemistry and Chemistry, but inter-institutional collaboration is not important in any of the subject areas they studied, reaching the same conclusion as He *et. al.*

Positive associations between international collaboration and impact have also been found for New Zealand sciences (Goldfinch *et al.*[87]), Italian sciences [72] collaborations of South American institutions (Sooryamoorthy [173]), and a case study of Harvard university research (Gazni and Didegah [80]).

2.5.2 Time factors of the collaboration

The impact of the collaboration has changed over time. Levitt and Thelwall[116] used nearly 30 years of WoS data in Information Science & Library Science subject category to learn how collaboration and citation change over time. Breaking the article citation into five strata, they found that collaboration in the highest four citation strata increased over time, whereas collaboration in the un-cited articles remained low. In other words, collaborative research is becoming increasingly significant and influential whereas non-collaborative work is becoming more and more difficult to be influential in IS&LS.

2.5.3 Regional and subject variations

Levitt and Thelwall [117] studied regional variations of the effects between the collaboration and higher citation. Using dataset from Social Science Citation Index (SSCI), they compared 18 countries, 17 US states. While they confirmed that in all regions they studied, the mean citation level of the collaborative articles was at least as high as that for the non-collaborative research, they noted that five of the US states had at least one citations indicator showing higher citation for non-collaborative articles. Leimu and Koricheva [115] studied articles from an ecology journal between 1998 and 2000, and compared US ecologist with the European counterparts. They claimed that the collaboration in ecology had a minor effect on the impact of the resulting publications, as measured by citation rates. Comparing the citation rates of the article by European authors and US authors, US ecologists benefited from collaboration more than their European colleagues.

2.5.4 Alternative collaboration measurements

The positive correlation in a different collaboration setup is also confirmed. Rigby and Edler [162] studied the collaboration intensity within projects and the quality of the research groups conducting these projects. 22 sets of research data were examined from Austria, and it was found that increased levels of collaboration within projects are associated with lower levels of variability of quality within each dataset. In other words, when collaboration is frequently conducted within the project, the quality of the output is more stable.

The main message from the literature is that collaboration is, to certain extend, related to the impact of the research paper produced. Between individuals, low impact research is improved by collaboration; in collaboration between countries, collaborative research is cited more frequently than singly published research, with exceptions in certain discipline (e.g. ecology). Extra attention must also paid to the types of collaboration and regional differences.

2.6 The Relationship Between Research Productivity and Impact

Compared to research collaboration, the study of the relationship between research productivity and research impact is rather sparse. There are very few studies directly investigating this relationship.

h Lanjouw and Schankerman [108] investigated relationships between companies' productivity and impact of patents. Patents, like article publications, present novel ideas (inventions) and are sometimes cited by other patents. Their productivity was measured by the ratio between the number of patents filed by the company and the resources spent on them. Patent impact was measured by the amount of revenue generated from it. They found a negative correlation between productivity and patent impact. That is, the higher the productivity (more patents filed for the unit amount of resource), the less the revenue generated by those patents. This result showed that the impact of a patent as measured by the revenue generated is dependent on how much resource has been spent on it. For the same amount of investment into producing the patent, the company with a lower number of patents filed actually generates more revenue.

Indeed, higher impact research takes more input because of the extra effort in conducting a thorough background review, careful design of experiments and the final presentation. It follows that the relationship between research productivity and research impact could be inversely proportional, although no previous research has confirmed this yet.

2.7 Network Models

Correlation analysis is a good mathematical tool for revealing relationships between selected variables. It tells us precisely how much correlations between the variables and the result can be used for predicting variables. However, with the growing amount of data available to us in the recent years, discovering the existence of relationships in the hundreds of variables has become equally important and challenging.

Network models abstract the data into graphs. These graphs can be visualised and analysed effectively. The recent advancement in personal computing power, it is possible to visualise and conduct graph analysis using quite sophisticated algorithms, and hence discover visually the relationships between variables in large scale.

In this chapter, we first present the mathematical background of the network models and network methods, then we review previous works in modelling and analysing networks that constructed based on the research publications.

2.7.1 Graph

The mathematical graph provides a solid foundation for network analysis. The type of network this study is concerned with can be modelled by a graph. **Graph Definition** A mathematical graph is defined as a pair of sets $G = \{V, E\}$, where V is a set of vertices (or nodes) $v_1, v_2 ... v_n$ and E is a set of edges (or links) that connect two vertices. The edges can also have values attached, so the graph becomes a valued graph.

Modelling real networks using graphs was recorded as early as the 18th century. Leonardo Euler tackled the famous Königsberg's Seven Bridges problem by modelling the islands and bridge connection as a graph. Since then, the study of the network models took off.

2.7.2 Random Network Models

This is a class of network models which includes the original Erdős and Rényi random graph model [63, 64] and relevant variations. The original model defines a very simple graph: a graph $G_{n,p}$ is defined as n nodes that connects each pair of nodes with probability p. In fact, graph $G_{n,p}$ is not a single graph, but a collection of graphs with n nodes and all the possible ways of connecting the nodes together with probability p.

There are two crucial aspects that the Erdős and Rényi graph model cannot model in social networks:

1. Degree distribution[6]. Due to the random nature of this model, the degree distribution follows Poisson distribution. This is very different from many real networks, such as social networks and citation networks, which follow the power-Law degree distribution [143, 158].
2. Clustering. Social networks have high clustering [84], indicating a locally well connected structure. However, the random network model cannot produce this local structure due to its random nature.

As a result, the random network model is not a very suitable network model to be used to study social networks. There are variations of the Erdős and Rényi network model to address these problems. The configuration model [132] and Chung and Lu's model [47, 48] specifically targeted the degree distribution of random networks. Holland and Leinhardt [96] and Strauss [73, 176] proposed models to address the clustering. But a common problem is that they become too complex to be useful in many studies.

Researchers started asking: are we starting from the correct foundation for modelling social networks?

2.7.3 Small-World Phenomenon

First we need to introduce a metric that measures the connectedness of a network – the Average Path Length (APL). The APL in a network is the average of the shortest path between all pairs of nodes. For instance, a network with APL of 3 tells us that on average, the path length between *any* pair of nodes is 3.

[6] Degree of a node is the number of edges connected to that node.

The Small-World Phenomenon is the observation that large networks – with millions or even billions of nodes – only have a small APL. This phenomenon is often found in many real networks. The human acquaintanceship network [184] consists of billions of people and its APL is only 6; the co-authorship network with 250,000 researchers [59, 144] has an APL of only 7. More recently, the small-world phenomenon has taken a precise meaning[9, 106]: networks are said to show small-world phenomenon if the APL of the network scales logarithmically (or slower) with the network size for a fixed average degree.

One of the important features of this class of networks is that the information transmission is much faster than, for example, a network that has APL in the order of thousands or millions. So this small-world effect is desirable in networks like the scientific collaboration network and the World Wide Web (WWW), where information and knowledge can channel through quickly, but is not so desirable in situations like disease transmission or the spread of rumours in social networks.

Another property of the small-world networks is that the nodes are locally clustered [59, 193]. This means that if node A is connected to node B and C, then B and C are very likely to be connected too. One demonstration in social networks is two close friends of someone are very likely friends themselves too.

The Erdős and Rényi random network model and its variations reproduce the small-world effect well [17, 26, 58, 74]. But as we have already discussed in section 2.7.2, the random network model cannot produce high clustering.

Watts and Strogatz [194] proposed a simple model that caters for both the small-world effect and high clustering (Figure 2.2). The model starts with a ring of nodes, then each node is connected to the nearest neighbour on both sides. A randomness parameter p is introduced, such that the amount of edges in the model is randomly rewired according to p. When $p = 0$, no edge is rewired and the graph remains a lattice; while when $p = 1$ the graph is completely rewired and becomes a random graph. By varying the randomness p, there is a sizeable region, as shown by Watts and Strogatz using numerical simulation, where the model has small-world phenomenon and is highly clustered.

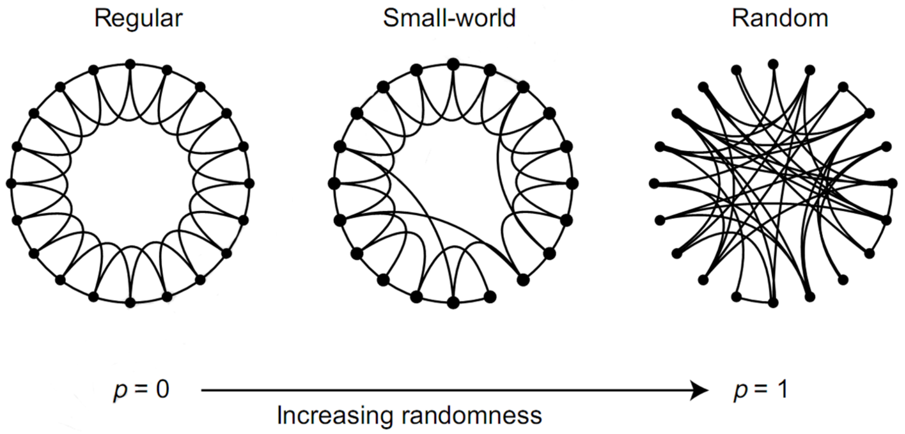


Figure 2.2: Watts and Strogatz Model. Left: the regular ring lattice with no randomness; middle, some randomness introduced when connecting neighbours, the network became small-world; right, a complete random graph. Figure reproduced from [194].

This model demonstrated observations in many social systems, where most people are friends with people they are geographically close to – colleagues, house neighbours, classmates – and the lattice represents these connections. Many people also have a few friends that live a long way away – friends living in other cities or other countries – the randomness adds long distance connections to the network.

Analysis of this model shows a surprising result: in order to convert a lattice network into a small-world network, only a tiny fraction of rewiring is required. What this means is that the small-world phenomenon found in social networks is stable and is not on the edge of collapse.

There are many variations of the Watts and Strogatz model. A much studied variant was proposed by Newman and Watt [145], which randomly adds edges to the graph but does not remove edges from the regular lattice. This prevents isolated clusters forming, which made the network easier to analyse. Models with higher dimensions have also been proposed and studied [54, 135, 146, 150], and the results are qualitatively similar to the one-dimensional case.

2.7.4 The Scale-Free Network Model

The Erdős and Rényi random network model is one of the simplest yet most studied network models. However, it has major weaknesses in modelling social networks. In 1999, Barabási and Albert [16] presented a new way of modelling networks. They emphasized the growth of networks found in real life. The social network, the citation network and the World Wide Web are evolving networks. They all started with few nodes, then new nodes were created and attached to existing nodes in the graph, which finally resulted in the current network. Barabási and Albert showed that in order to produce a real network's degree distribution, whenever a new node is added to the graph, the node must have a higher chance to connect to nodes that already have many connections. For instance in citation network growth, a new publication has a higher chance to cite one that thousands of other publications cite, than one which only a few dozen publications cite.[7] They call this the *preferential attachment*. The resulting topology is that their degree distribution follows a power law. This means that most nodes have very few links, but the remaining few nodes have all the rest.

The importance of their contribution is not only on a new network model, but also a whole new way of viewing a network – it is a dynamic, evolving structure. Some of the network features are rooted in the evolution of the networks rather than the network's topological characteristics.

2.7.5 The Citation Network

Citation networks are classic knowledge networks. The research papers – the carriers of original ideals and knowledge, cite one another, indicating the path of knowledge evolution. In a typical citation network, the nodes represent papers and the directed links represent citations. Because papers are generally cited after their publication, so citations can only point back in time, *i.e.* only later papers can cite previously published papers. The citation network, unlike the co-authorship network, is a non-cyclic network and the arrows on the links point back in time.

Researchers started studying this network in the 1950s. Price [157] was among the first to investigate the patterns of citations. He found that a small number of papers are cited

[7] This can be simply explained by probability: If thousands of publications cite a paper, then this paper is much easier to be found than one that only a few papers cite.

more frequently than average, while the majority of papers are cited less frequently than average.

Citation is often used as an indicator of scientific performance. Redner [160] studied the relationship between citation count and scientific impact; Cronin [50] investigated the *h*-index[8] and impact ranking of authors; Cole [49] and van Raan [185, 187] evaluated the influence of awards, honours and Nobel laureateships on citations. These studies in general give a positive result for citations measuring scientific activity.

However, some studies have suggested that citations are not a suitable measure for scientific activities [77, 198]. They claim that citation depends on many factors besides scientific impact. These include, for example:

- Time-dependent factor: the more frequently a publication has been cited, the more frequently it will be cited in the future [40, 157];
- Availability of the publication: physical accessibility [174], open access of publications [36, 90] and publishing media influence the probability of citations [171];
- Author-reader dependent factors: results from Mählck and Persson [124] and White [195] showed that citations are affected by social networks, as authors cite personally acquainted authors more often.

Bornmann and Daniel [30] reviewed the citing behaviour of scientists and concluded that at the micro-level, citing is a social and psychological process that is mixed with personal bias and social pressures; but at the macro-level, scientists give credits to colleagues by citing their work. Thus, citations represent an intellectual or cognitive influence on scientific work.

Studies of the citation network enabled us to understand the structure of knowledge and anticipate developments in various domains. The network constructed using citations between published papers is a knowledge network, where papers point towards a source of knowledge. Since papers are produced by researchers and increasingly, co-authored

[8] *h*-index, or Hirsch index is a value used to estimate a researcher's impact, using citation. Original article: [95]

by researchers, networks of researchers can be constructed, both based on the extension of the citation network (co-citation) and co-authorship. In the coming sections, we show how these two types of networks are constructed and what these networks add to the analysis of the scientific activity.

2.7.6 The Co-citation Network

Author cite papers together (*e.g.* in the same sentence or in the same paragraph) when the content of those papers are somewhat relevant. Collectively, the co-citation can represent a group of authors' decision in the content similarities of previous publication. The more frequent the publications have been co-cited, the stronger the similarities between publication, hence stronger the similarities in the researchers' work.

There are two types of co-citation networks which are commonly discussed in the literature: the paper co-citation network and the author co-citation network. These are in fact two different networks. The paper co-citation network builds a network of papers that are frequently cited together, so it is useful for studying knowledge structure and knowledge propagation. The author co-citation network connects the researchers who made the publications. It links researchers together and is useful for studying the potential researcher's relationship and the likelihood of their collaboration. We focus on author co-citation network in this study.

The construction of the author co-citation network can vary depending on availability of data and processing power of the study. A pair of authors can be co-cited in the same sentence, in the same paragraph, or in the same article. The strength weakens as the authors co-cited are further apart. The position of the author in the author list also matters. Some studies only use the first author and ignore all other authors for co-citation, while some use all authors. These studies can give very different result due to the construction of the co-citation network.

The original methodology by White [196] only considers the first author of any given paper and disregards the contributions of other co-authors. This was perhaps due to the limitation in computing technologies. Follow-up studies by Persson [153], Zhao [202] and

Callahan [37] consider all authors listed on a paper and the context where the co-citation occurred, thus helping to identify the domains of authors who are seldom listed as first authors. Su [177] proposed an algorithm to discover authors based on their expertise. The input to her algorithm was a co-citation network.

A few domain co-citation analyses have also been performed in trying to understand the sub-domains and the top active authors. White *et al.* [197] analysed the information science field from 1972 to 1995 using the author co-citation. They generated maps of the top 100 authors in the field and used factor analysis to identify major specialities. They found that information science consists of two major specialities with little overlap.

Chen and Carr [42] used ACM publication data to study the structure of the hypertext literature. Authors cited fewer than 5 times during the period 1989-1998 were filtered, resulting in 367 authors. An author co-citation matrix was constructed and Principal Component Analysis (PCA) was applied. The temporal information of the papers was included in the visualization methods, allowing them to identify emerging research directions in the field.

2.7.7 The Co-authorship Network

Authors form co-authorship relationship if their have co-authored papers and both their names appear on the same paper. In this network, nodes represent authors and links represent co-authorship. Co-authors generally know each other and many of them collaborate with each other. So to a certain degree, the co-authorship network represents researchers' social network and collaboration relationships, hence, it has attracted much research attention in recent years.

2.7.8 The Erdős Number

Calculating the Erdős Number is one of the earliest activities based on the co-authorship. The Erdős Number is a measurement of the number of collaboration steps a researcher co-authored with the famous Mathematician - Paul Erdős. Researchers who co-authored a paper with Paul Erdős have Erdős Number 1; researchers who co-authored a paper with a co-author of Paul Erdős have Erdős Number 2 and so on. Those authors who never co-authored a paper with Paul Erdős don't have an Erdős Number or are said to have an infinite Erdős Number. De Castro and Grossman [53] found that many famous researchers, whatever their research areas, have a finite Erdős Number. Because the famous researchers also are tightly connected within their own research domain, this leads to an entire research community being connected through co-authorship. The implication is that scientific research is a collaborative work rather than individuals making their own discoveries. They also reported that in order to have a smaller Erdős Number, quality is more significant than quantity. The person with whom one has collaborated is more important than the number of collaborators.

2.7.8.1 Domain analyses

Co-authorship analysis is widely used to understand publication and collaborative patterns among researchers in a specific domain.

Newman [139–141, 144] carried out a series of co-authorship analyses in 2001. He answered a wide variety of questions about collaborative patterns by analysing co-authorship networks, such as the number of papers authors write, how many people they write them with and the typical distance between researchers through the network. He compared these attributes across several domains – Biology, Physics, Computer Science and Maths, and made these claims:

- The number of papers written per author is similar across the domains in the study;
- The number of authors per paper and the average number of collaborators vary substantially across domains;
- All of the subject domains have a largest component connecting at least 80% of the researchers;
- The average collaboration distance is small, typically 4 to 6 steps for a network containing millions of nodes;

- The clustering coefficient is much smaller than a random network expected value.

Other domain analysis work using co-authorship include: Glänzel [85] studied scientific networks in general; Moody [134] investigated social science collaboration networks; Liu [118] and Sharma [170] studied the digital library community, including a few others [62, 111, 119, 128, 164, 180]. These studies came to similar conclusions: researchers are mostly connected; the distance between researchers is short and the network is highly clustered; co-authorship networks are small-world networks.

2.7.8.2 Network dynamics and evolution

Studying co-authorship networks has not been limited to understand the snapshot of scientific collaboration in time. Since the published work has a timestamp, which tells us when the co-authorship represented by the publication was added to the network, it is possible to study an evolving network of people.

Barabási and Albert [7, 15] proposed a model based on the co-authorship network that captured the network's time evolution. They realised that the features commonly used to identify a network, such as average degree, diameter, clustering coefficient, are in fact time dependent and no longer suitable to characterise a network. On the other hand, they found that the degree distribution is a stationary measurement for an evolving network, which can potentially be used to characterise a network instead. In addition, they also discovered that the measurements on incomplete data could lead to opposite tendency. For example, the node separation exhibits a decreasing tendency on datasets that only cover certain periods, while their numerical simulation suggest otherwise.

Newman [142] used the evolving social network extracted from co-authorship to predict further collaboration. He analysed what affects researchers' choice over who to collaborate with, given their previous publications, he found that the probability of a pair of researchers collaborating increases with the number of other collaborators they have in common; and the probability of a particular researcher acquiring new collaborators increases with the number of his/her past collaborators. This result demonstrated evidence of *preferential attachment* in co-authorship networks.

Co-authorship has been frequently used as a source data in the network studies. It has facilitated many interesting discoveries in researcher social networks and helped us understanding the disciplinary differences in collaboration. We will be using this data and network visualisation to show the institutional collaboration network.

2.8 Chapter Summary

This chapter reviewed literatures in the area of measuring research activities – doing collaborative research, increasing scientific knowledge production and achieving higher impact. We then investigated how these activities relate to each other. Towards the end, we reviewed network models and network methods in assisting of discovery of relationship between the activities.

The definition of research collaboration is simple, but to accurately measure the complex collaborative interactions between researchers is difficult. Collaboration can vary by closeness, frequency and the respective roles of the researchers in the collaboration. Across disciplines, the roles of collaborators can also vary. While some disciplines consider certain roles as collaboration, others do not as reflected in their inclusion or exclusion among the paper's co-authors.

Despite the complexities of collaboration, using co-authorship as its metric has demonstrated its advantages. The limitations were also discussed. Co-authorship will be used in this thesis as the measure of research collaboration.

The measurements for research productivity were investigated. The widely used definition of productivity takes the ratio between the output and input, *i.e.* the output generated on the unit input. Although the input as well as output of the research activity was identified in the literature, however, the data was generally unavailable, hence this approach was not widely used. The publication productivity, which is an aspect of the research productivity that counts the number of publication, is commonly used in the literature. Despite it is a non-normalised variable, and as a result, it has bias towards larger sized institutions; it is the only variable available to use in a large scale study. Several publication counting methods were reviewed, which may offset the bias.

There were four perspectives on quality commonly discussed in the literature: methodological quality, report quality, expert review based quality and bibliometric impact. In bibliometric methods, citation was one of the most widely used impact indicators in the literature. New metrics, such as download logs and network based methods were showing correlations to the impact of research. Higher aggregation of research entities (*e.g.* journal, country level and university level) have adopted their own unique methodologies to measure their research related qualities. In this respect, country-wide research assessment, journal impact factor and university rankings were reviewed and discussed.

Collaboration was the focus of previous studies on the correlation analysis. Despite the different measurements and types of collaboration considered, the general findings suggest that collaboration correlate with the impact of the publications (especially the lowest impact ones); and international collaborations are cited more than national ones. These results were obtained in a variety of disciplines including Psychology, Chemistry and Computer Science.

I would like to stress, at this point, that correlation does not imply causality. If one variable correlates with another it simply means there are relationships between them and they change together. Causality from one variable to another is a much stronger relationship. To confirm a causality requires very high quality data, especially the timestamps that generated the data points. As a result, there is little research trying to find causal relationships among these three variables. He *et al.*[93] touched on the causality between collaboration and productivity using the longitudinal data about the New Zealand researchers.

Little research effort has been devoted to institution-level collaboration, and the discipline coverage has been limited, which prevents us from understanding whether the effects of the collaboration on impact and productivity are uniform across disciplines.

Based on the material presented in this chapter, we have the building blocks to measure the collaborativity, productivity and impact of institutions. In the next chapter, I present how raw publication data was processed before correlational analysis was applied.

Chapter 3

Research Questions

Universities and scientists tend to assume that (1) if they publish more, they are producing higher impact research; and that (2) if they increase collaborative research, they will produce both *more* and *higher impact* research. These causal assumptions have been accepted among the research productivity, the research impact and the research collaboration.

To provide evidence to these assumptions, the following central research question is explored in this study:

> *What are the relationships, at the* ***institution*** *aggregation level, among collaboration, productivity and impact?*

This central question involves three core variables: collaborativity, productivity and impact at the aggregation level of institution. The correlation between these variables can split into three pairs: collaborativity vs impact, collaborativity vs productivity and productivity vs impact. The past studies address only one pair of the variables and rarely consider them all together in a single study[137]. This study not only correlates them in pairs, but also apply partial correlation to remove the effect of the third variable, so that the true correlation between the pairs can be revealed.

In the following section, the central question is divided into three sub-parts – the correlation between collaboration and impact, collaboration and productivity, and productivity and impact – one for each pair of the variables. A series of additional questions are asked, based on the gaps identified in the literature.

3.1 Research collaboration and impact

We have seen a wealth of articles focusing on exploring the relationship between research collaboration and impact in the literature. A positive correlation was found in a variety of fields including Nanoscience and Nanotechnology[57], Ecology[115], Economics[117] and Library and Information Science[116], but in Finance[12], Literature of Academic Librarianship[92] or Library and Information Science[116], a correlation could not be found. In these disciplinary studies, the aggregation is at article, author, journal or country level. The collaboration measurement was estimated by the number of co-authorships. A stronger collaboration between a pair of authors(or countries) is demonstrated by more papers co-authored. Gazni[80], Bordons *et. al.*[28] and Larivière *et. al.*[110] looked at the co-authorship sizes, *i.e.* the number of authors, addresses and countries listed on a paper. They have also found a positive correlation with the impact of the paper. In addition, it was further shown that not all types of collaboration have the same relationship with impact [56, 57, 72, 80, 137, 173]; the impact of the collaboration changes over time [116]; and the relationship between impact and collaboration has regional variations [115, 117].

Gazni[80] took Harvard university papers as a case study, found that at least 60% of the papers published are multi-author publications, suggesting that Harvard is a highly collaborative institution. Harvard, by its reputation, is a high impact institution, we can generalise Gazni's finding by asking the following questions:

1. *Are higher impact institutions more collaborative?*

2. *Do higher impact institutions emphasize on collaborative research?*

Collaboration and impact is rarely explored at the institutional level. The current study fills this gap by including thousands of institutions across five disciplines and conducting the analysis at the aggregation level of institution.

3.2 Research collaboration and productivity

Previous research suggested a close relationship between the research collaboration and productivity. Many found that collaboration has a positive correlation at the author level on productivity [33, 114, 205], and on author's future productivity [93]. The number of researcher's co-authors was also found to have positive correlation with the researcher's productivity [60]. Similar results were also confirmed in qualitative studies: by interviewing Nobel Laureates[205]; and by interviewing researchers in a New Zealand university[93].

At the aggregation level of entire institution, Katz[105] applied bibliometric methods to assess the collaboration status between institutions in three countries – UK, Canada and Australia. Among other questions he explored, he found a non-linear relationship between institutional collaboration and productivity, where larger institutions have proportionately fewer collaboration than smaller institutions. However, he could not generalise this finding due to the limited institutions from only three countries. Abramo *et al.*[1] studied the relationship at the institution level between Italian universities, while they found a strong correlation in information engineering, the correlations varied substantially among the remaining 7 research areas they analysed.

While the positive correlation has been found at the author aggregation level repeatedly, the studies focused on investigating at the institution level are scarce, confusing and non-generalisable. To fill these gaps, this study uses data that covers thousands of institutions to attempt to answer the following question:

> *3. Do institutions that publish more papers also collaborate more?*

We also try to generalise Katz's finding by asking:

> *4. Do institutions that publish more papers also publish proportionately more collaborative papers? (i.e. Do high productive institutions emphasize on collaborative research?)*

3.3 Research productivity and impact

Few studies consider the relationship between research productivity and impact, and those that do examined it in limited ways.

Lanjouw and Schankerman[108] investigated relationships between companies' productivity and impact of patents. They found a negative correlation between productivity and patent impact. Similarly, Bergh *et. al.*[21] found that authors having fewer articles tended to have articles that received the most citations in a journal, suggesting a negative correlation between authors' productivity and their impact. No study has yet addressed this relationship at the institutional level.

In the current study, the correlation is investigated at the institutional level and the following questions are addressed:

> *5. Do institutions that publish a large number of papers have higher impact? Are there disciplinary differences?*

> *6. Are papers published by high paper-output institutions cited more often than papers published by low paper-output institutions?*

This study uses recent advances in digital archiving and indexing services to do large-scale quantitative analyses of the relations among research productivity, impact and collaborativity, comparing effects across disciplines as well as countries. In the following chapter, the dataset and the methodology are described.

Chapter 4

Datasets and Methodology

This chapter addresses the tools and the source datasets used in this research. It splits into six sections, detailing (1) the computing resources used; (2) the datasets and the preprocessing works; (3) the descriptive statistics of the source data; (4) the metrics selection and the counting methods used for each of the three institutional research activities; (5) the correlation methods used and (6) network methods and visualisations.

4.1 Computing Resources

The entire data processing, data analysis and graph visualisation was performed on a standard issue research student PC Workstation with Quad-core processor and 12GB RAM at the University of Southampton. The majority of the data processing was completed in an Ubuntu Linux environment using the Python v2.7 programming language and the MySQL v5 database. The visualisation was performed in a Windows 7 Environment using the Network Workbench Software [182] package with GUESS[1] visualisation module.

[1] GUESS is a graph visualisation and layout software. It is written in Java and has implemented several widely used graph layout algorithms. http://graphexploration.cond.org

The analysis of the data, including data normalisation, correlation and partial correlation was performed with IBM SPSS[2] v19 statistical package. Chart drawing was completed with Microsoft Excel software package.

4.2 Datasets and Pre-processing

The Thomson-Reuters Web of Science (WoS) database collects and indexes the world's leading scientific journals in sciences, social sciences, arts and humanities. It was started by Eugene Garfield in the 1960s to index the citations between papers electronically. The citation index enables the use of the bibliometric analysis in learning about the scientific publication and practices in general. The data used in this study was obtained directly from the Web of Science under the licence that the data is solely used for research. This data covers papers published in Computer Science, Psychology, Pharmacology, Law, and Materials Science between 1973 and 2010. The metadata of the papers were provided, which includes paper's publication year, discipline, author list, institution list and the list of citing papers (papers that have cited this paper). The format of data was in database text dump, which was easily imported into a local MySQL database. The data included a total of 2,127,015 publications. WoS collects several types of publications, including articles, book reviews, letters, meeting abstracts and review articles *etc.*, but the type of each individual document was not available. The entire dataset was therefore used in analysis regardless of publication types.

The Association for Computing Machinery (ACM) is an international scientific and educational computing society, it is the world's largest computing society and it publishes leading computing journals and organises recognised conferences. The ACM Digital Library is an online service that collects the ACM's journals and conference proceedings starting in the 1950s. A dataset that contains 214,592 publications was provided by ACM directly for this research. This dataset is assumed to only contain publications on Computer Science subjects. The data is organised in journal issues and conference

[2] SPSS is a software package for statistical analysis. It has a graphical user interface and implements a wide range of algorithms.

proceedings. The publication's year, author list, institution address, citing article and reference list were included in the provided data.

Webometrics Ranking of World Universities (Webometrics ranking)[3] is a world university ranking based on data found on the web. It is calculated and published annually by the Cybermetrics Lab in the Spain. The July 2010 version of the ranking was downloaded for this study.

University Master List University names are not standardised, it is possible and often the case that the same university is spelled differently by different people publishing their papers. For example, some use abbreviations, some use different language, some even misspells. Different databases tend to use different university name conventions. For instance, the WoS dataset uses abbreviated universities' names while ACM dataset uses the universities' addresses. It poses problem in matching the universities across databases accurately.

A master university list was used to simplify the process of mapping universities' name across multiple databases. Each database maps its own university names to this master university list, which is then used to link together all of the university name variations across databases. The Webometrics ranking university list was used as the master list because: 1. it uses the common English university name spellings, which takes less effort to convert other name conversion into it; 2. it contains almost all universities in the world, regardless whether they are teaching oriented or research oriented, so it gives a wider coverage; 3. a university's country is provided when the university name is not unique. For example, University of Technology exists in many countries; 4. it provides domain and url for each of the universities on the list, which can be used as an identifier for the university. Using an intermediate university list also makes this work easily expandable to include new databases in future analyses.

[3] http://www.webometrics.info

4.2.1 ACM Digital Library Dataset Description

The metadata of articles are contained in the set of XML files. Figure 4.1 shows a sample part of an article. The ACM dataset provides the following details for all of the papers:

- article ID
- title
- publication year
- authors' names and ID
- authors' affiliation
- papers cited this paper

Due to the large size of the data (1.8 GB in total), accessing and processing the entire set was extremely slow. It took several hours just going though each article in the dataset to read its metadata. To greatly improve the processing speed, a data extraction step was performed to save only the useful data in a new format. These extracted data were stored in a file for simplicity. After this extraction process, the time takes to go thought the dataset was reduced to the orders of seconds. Figure 4.2 lists the fields extracted and the data structure and figure 4.3 shows an example of a paper metadata in the structure.

During the data extraction, I noted two major data quality issues regarding the paper metadata within the ACM dataset:

1. There was no unique ID assigned to each institution, nor were the institution names standardised. Only the institution addresses were available to us to identify the institution. These institution addresses appear to be the original text typed by the authors, and they are typically not standardised and contain (but not limited to) details like department name, post code, country *etc.*

```xml
<?xml version="1.0" encoding="ISO-8859-1"?>
<article_rec>
        <article_id>227236</article_id>
        <title><![CDATA[ISO 9000 reflects the best in standards]]></title>
        <page_from>17</page_from>
        <page_to>20</page_to>
        <doi_number>10.1145/227234.227236</doi_number>  <citation_url>http://portal.acm.org/
     citation.cfm?id=227234.227236&amp;coll=portal&amp;dl=ACM</citation_url>
        <authors>
                <au>
                        <person_id>PP39048660</person_id>
                        <seq_no>1</seq_no>
                        <first_name><![CDATA[Roy]]></first_name>
                        <middle_name><![CDATA[]]></middle_name>
                        <last_name><![CDATA[Rada]]></last_name>
                        <suffix><![CDATA[]]></suffix>
                        <affiliation><![CDATA[Boeing Distinguished Professor of Software
     Engineering at Washington State University]]></affiliation>
                        <role><![CDATA[Author]]></role>
                </au>
        </authors>
        <references>
                <ref>
                        <ref_obj_id></ref_obj_id>
                        <ref_seq_no>1</ref_seq_no>
                        <ref_text><![CDATA[Huyink, D. and Westover, C. 1SO 9000. Irwin
     Professional Publishing, New York, 1994.]]></ref_text>
                </ref>
        </references>
        <cited_by_list>
                <cited_by_number>2</cited_by_number>
                <cited_by>
                        <cited_by_object_id>232020</cited_by_object_id>
                        <cited_by_text><![CDATA[James W. Moore , Roy Rada, Organizational badge
     collecting, Communications of the ACM, v.39 n.8, p.17-21, Aug. 1996]]></cited_by_text>
                </cited_by>
                <cited_by>
                        <cited_by_object_id>245110</cited_by_object_id>
                        <cited_by_text><![CDATA[Roy Rada , James Moore, Standardizing reuse,
     Communications of the ACM, v.40 n.3, p.19-23, March 1997]]></cited_by_text>
                </cited_by>
        </cited_by_list>
</article_rec>
```

Figure 4.1: A sample article in the ACM XML dataset

2. The authors were inconsistently named and not properly identified. Most of the authors have been assigned with unique identifiers, but some have different identifiers for each of their name format. For example, Prof Les Carr in the University of Southampton was found to have several identifiers each linked to his different name format, such as Leslie Carr, Les A Carr *etc.*. Each of these identifiers has publications linked to it.

The problem 1 is specific to the ACM dataset, it is a major barrier before doing statistical analysis on the data. In the following section, I discuss my approach to identify the universities from an address and present our method to evaluate our method. The

```
Article Year,
     Author 1 firstname, lastname, ID, Address
     Author 2 firstname, lastname, ID, Address
```

Figure 4.2: Intermediate data structure used for the ACM dataset

```
2005,
     Ben, Shneiderman, PP40024227, Univ.of Maryland, College Park
     Maryann, Alavi, P193437, Univ.of Maryland, College Park
```

Figure 4.3: An example entry in intermediate data structure for the ACM dataset

problem 2 regarding the author name is more generic, the WoS dataset also has a similar problem. Our approach to the problem is presented in section4.2.3.

4.2.1.1 University Identification in ACM dataset

The ACM dataset stores the institution address as text items attached to the author (figure 4.1). These items sometimes include department, institution name, institution address, city, country and post code. They appear to be the originals the authors submitted and the ACM had not processed or identified. There were 473,634 non-empty addresses in the dataset, visual analysis on randomly selected a few hundred addresses revealed the following potential problems in matching the universities to the Webometrics list.

1. Name variance in the spelling of the university name. For example, University of California Berkeley was sometimes written as UC Berkeley, UCB and so on.

2. Language variance in the university name. For example, Universität Hannover was sometimes written as University of Hannover.

3. European accents were left as the HTML code. For example, `&auml;` found in the source data should be displayed as ä.

4. Multiple universities were described in the same text item. (These can be those authors that affiliate with multiple institutions.)

5. Mis-spelling of university names.

Targeting issue 1-3, the following algorithmic rules were implemented. The rules were executed in the order below:

1. The strings were converted into lower case, making the recognition case insensitive, thereby reducing the possible variations for addresses.

2. The HTML code in the source data was converted into the actual characters.

3. The European language variance of the word 'university', *i.e.* 'universität' was shortened into 'univ', thus reducing the range of variations the university is spelt in European countries.

4. The names of the institutions were looked up from the Webometric university list. An additional lookup table was constructed to map between the university name variations and the university name used by Webometric list.

The lookup table was constructed by a computer program that analyses the address frequency. The most frequent variations in the source data were manually identified, and the variations were added to the list of name variation for the university. This process is repeated until the increase in the address recognition rate is negligibly small for every new address variation added to the lookup table. The address recognition rate is presented in the next section.

To evaluate the scale of the problem of multiple institutions appearing in the same text item (issue 4), two variations of the algorithm were implemented. the algorithm 1 assumed that only one university was expected from any given address, while the algorithm 2 did not made such an assumption, so it finds as many universities in the address as it can. Algorithm 1 found 273,844 university instances and algorithm 2 found 275,978 university instances using the same lookup table on the same dataset. This is an increase of 0.4% more universities using algorithm 2 than using algorithm 1. This small percentage increase is not significant in the overall university recognition rate. Considering the overhead and complications in algorithm 2, I decided to use only algorithm 1 for university recognition. This means that I ignored the author's second affiliation, so fewer institutions appear to be collaborating in this dataset.

The problem of mis-spelling (issue 5) was not dealt with directly. Mis-spelling of a university name can be counted as a variation in the university name. If the mis-spelling is very common, it would be picked up by the frequency analysis process and added to the lookup table. But there was no mis-spelling common enough to enter the lookup table.

4.2.1.2 ACM Dataset University Recognition Rate

The ACM dataset includes many types of institutions ranging from hospitals, research centres, companies and so on. In order to learn the university recognition rate without treating all the unrecognised institutions as universities, we need to estimate the proportion of the university addresses. Thus we can avoid underestimating the recognition rate by basing the data on all of the addresses in the dataset.

The addresses in the ACM dataset are not standardised to be usefully aggregated. Table 4.1 shows 5 addresses all representing the same institution. The actual possible variation of the address for the same institution is far more than the listed ones. So without the knowledge of the number of unique institutions, it is only possible to estimate the *address* recognition rate, that is, the proportion of the recognised addresses out of the total recognisable addresses, with no addresses aggregated. This is to be distinguished from the *institution* recognition rate used in WoS, where the addresses have already been pre-processed and the addresses representing the same institutions are aggregated. The non-aggregated addresses include duplicated institutions. That is, an institution's address is repeated as many times as the number of papers published by the institution.

As a result, the address recognition rate is not directly comparable with institution recognition rate.

Addresses	Institutions
Massachusetts Institute of Technology	MIT
MIT	MIT
M.I.T	MIT
M.I.T, Cambridge	MIT
MIT,Cambridge	MIT
MIT, Cambridge	MIT

Table 4.1: Address variations in the dataset. The actual variations of the address for the same institution is far from the listed 5.

Address recognition rate is calculated by the formula 4.1

$$Univ.\ Address\ Recog.\ Rate = \frac{Num.\ of\ Recognised\ Address}{Total\ Addresses\ \times\ Univ.\ \%} \tag{4.1}$$

To estimate the percentage of addresses that represent universities (Univ.%), two hundred of randomly selected non-empty addresses were inspected, and their institution type were identified manually. The distribution of the institution types is shown in table 4.2.

Institution	Percentage
Company	21%
University	70%
Research Centre	9%

Table 4.2: ACM institution type distribution

The sampling reveals that the university is the major contributor in Computer Science research as seen by ACM, where 70% of the contributors are from the university. Using formula 4.1, the address recognition rate can be calculated. After several iterations of the lookup table improvement, it was able to recognise and match 273,844 address lines, which represents an address recognition rate of 83% ($\frac{473{,}634\times 70\%}{273{,}844} = 83\%$). I decided to stop here because the rate improvement with each additional address in the lookup table is negligibly small.

4.2.2 Web of Science Dataset Description

Unlike the ACM dataset, which includes only computing publications, WoS collects a broader range of disciplines. They maintain three citation indices, Science Citation Index Expanded (SCIE), which covers 150 disciplines in nature science and engineering; Social Sciences Citation Index (SSCI), which covers the social science disciplines; and the Arts & Humanities Citation Index (AHCI), which covers the arts and humanities disciplines. Its index coverage in terms of the journal is also vast. It collects 17,240 leading journals out of about 28,000 total journals in the world [161], that is more than 60% of the journals. Almost all of the current active research disciplines are covered by one of its citation indices, making WoS the largest citation index in the world.

Due to licensing and limited computing power, I can only choose a limited number of disciplines. The choice of the disciplines was decided based on the practice difference of research publications across two groups – Nature Science and Engineering (NSE) and Social Science and Humanity (SSH). In NSE, peer reviewed journal publication is the primary output of the research, while in SSH, research are more often disseminated in the form of monographs, which are not indexed in the journal-based databases such as Web of Science [86, 109, 138]. Psychology and Law were selected to represent SSH while Pharmacology and Materials Science were selected to represent NSE. Computer science was also included to compare with the ACM dataset.

4.2.2.1 WoS Data Format and Data Relations

The WoS dataset provides the following details about each paper:

- paper ID
- publication year
- subject
- list of authors
- author's order of appearance

- list of institutions
- institution's order of appearance
- institution's country
- papers cited this paper

The primary entity of this study – the institutions – were not uniquely identified by the dataset. A decision was made to use the combination of institution's name and its country as the primary key. to identify them in the university master list. Once a match was found, the URI of the university provided by the master list was used as an identifier for an institution.

The authors of the institutions were also not identified. The dataset did not provide author's information except their names, which only includes the first initial and surname. With just author's initial and surname, there is not enough information to separate two authors with the same name easily. Previous work has shown to use co-authorship and author self-citation to assist author's identification, I discuss these works in section 4.2.3.

The WoS dataset do not provide author's affiliating institutions. The authors and the institutions were separately linked to the papers, but no link exists between them. An extra field – order – was provided in both the author table and institution table, which may offer us clues to the author-institution relationship. Unfortunately I was told by the data provider that this ordering information cannot be used to reconstruct the relationship. (It appears that many of the paper's author numbers do not match the institution number, which makes reconstruction problematic.) There are 64% of the papers in the dataset that have *more* authors than number of institutions, while 3% of the papers have fewer authors than the number of institutions. The papers with more authors than institutions could be due to the source journal's convention of storing the author and institution. When authors come from the same institution, some journals may have chosen to omit the duplicated institutions on the paper. The high percentage (64%) of paper with more authors than institutions means that it is quite common for multiple authors from the same institution to collaborate on a paper. On the other hand, the 3% papers with more institutions than the authors are mainly because of an author's multiple affiliations. This phenomenon of authors specifying multiple institutions in their published paper was referred to by Katz as 'top down collaboration' [105]. He claimed that the appointment of a single researcher at multiple institutions symbolises the collaboration between institutions. However, in the UK, it is also a common practice for researchers and lecturers to visit other institutions as part of their career progression. These researchers may have reasons to put down both institutions in their published papers. This multiple affiliation resulting from visiting researchers may not necessarily be an indication of institutional collaboration. The collaborations resulted by multiple affiliation of the authors were removed when possible in this study.

4.2.2.2 WoS University Name Processing

The institution name processing for WoS dataset is very different from the ACM dataset. The ACM dataset used free text describing the institution's address while the WoS dataset has already standardised institution names. None of the free text processing problems identified for the ACM dataset in section 4.2.1.1 apply to the WoS dataset. The 12,894,008 WoS addresses were first aggregated into 539,356 institutions. The processing WoS data moves straight to match these standardised institution names to the Webometrics institution names.

There are two potential ways to match them. One is the forward method: to apply the WoS abbreviation rules to the Webometrics's institution names, then uses the obtained abbreviations to match with the WoS list; the other is the backward method: to reconstruct the WoS abbreviations into the full institution names used by the Webometrics list. Both approaches are equivalent if the rules are non-destructive, but it is not the case here.

Table 4.3 lists the four most applied rules used by WoS to abbreviate the institution names.

Since the rules used by WoS are destructive, for example, rule 2 in table 4.3, after removing 'of' from the phrase, it is not possible to reconstruct the original institution's

	Rule Description	Original	With rule applied
1	University and its European variations are shortened to UNIV	Universitatea	UNIV
2	Connecting words, ‘of’, ‘de’ are removed	of	
3	Space is converted to ‘-’	␣	-
4	European accents are converted to English alphabet	ä; à; á	a
	Example:	University of Edinburgh	UNIV-EDINBURGH
	Example:	Université de Montréal	UNIV-MONTREAL

Table 4.3: The most used WoS abbreviation rules. These rules are destructive. The original can not be obtained by applying the inverse of these rules.

name, because the location of the word was lost. This leaves us only the forward method for doing institution matching.

In addition to the abbreviation rules, WoS uses a separate list for those long established institution abbreviations. For example, California Institute of Technology is shortened to Caltech; Massachusetts Institute of Technology is shortened to MIT. These are also included in this study. Institutions with an abbreviation unrecognised or incomplete are excluded in this study. For example, about 1400 institutions labelled as ‘INCONNU’ (French word for ‘unknown’) are excluded.

4.2.2.3 WoS Dataset’s University Recognition Rate

The WoS dataset includes a range of institutions such as research centres and companies. To learn only the recognition rate for university of our matching method, I need to find out the proportion of the addresses that are actually universities.

This involves 4 steps:

- find out the total number of institutions in the dataset.
- find out the proportion of the institutions that are universities.
- work out the expected number of universities.
- work out the university recognition rate.

The calculation of the university recognition rate is shown in formula 4.2

$$University\ Recog.\ Rate = \frac{Recog.\ University\ Num.}{Total\ Inst. \times Univ\%} \tag{4.2}$$

There were in total 12,894,008 addresses, aggregation on these addresses gave a total of 539,356 institutions' addresses. This is a large number of institutions that have made scientific contributions compared to the total number of universities (12,000 establishments) in the globe. A close examination revealed two possibilities that may have elevated the number of institution: 1. these institution addresses include former institutions. For example, there were institutions with an address in former Soviet Union. 2. WoS institution abbreviation is just a shortening rule, it still relies on the source data for the correct addresses. For example, Peking University was found to have two representations: PEKING-UNIV-BEIJING and PEKING-UNIV. It may potentially have many more other representations too, increasing the institution counts.

The majority of the total 539,356 institutions contributed very few papers. There were 496,947 (92% of total) institutions only ever authored or collaborated in less than 10 papers during the past 37 years. This potentially indicates that majority of the contributors are one-off, or these addresses are just variations to the original institution.

To determine the proportion of *universities* in the dataset, a random sample of 200 institutions were selected, and 11% of them were found to be universities. Scaling it to the entire dataset, the expected number of universities is about 59,000. This is almost 5 times more compared to Webometrics' world wide university listing of 12000. This is revealing that the WoS has not been standardising institution names very well (issue 2 above), making alternative spellings of the university name been categorised as a different university.

Executing the matching algorithm on all of the institutions gave 7,183 matched Webometrics universities. Further analysis reveal that out of the 7183 matched universities, only 1972 have been actively publishing (at least 3 authorship per year). The decision was made to use only active universities because (1) very low contribution university will not change the big picture; (2) the methodology adopted in this study (correlation analysis) is unsuitable for analysing a large quantity of less quality data. Applying formula 4.2, I obtained the university recognition rate of 12%.This is equivalent of nearly 60% of the current world university found to have contributed to the domain analysed.

WoS Institution Types The institution type distribution – whether they are a company, research centre or university – for each discipline was also examined. The WoS data was categorised into five disciplines and institutions with less than 100 papers were removed. A random sample of 200 institutions was selected in each discipline. Table 4.4 shows the institution type for Pharmacology (Phar.), Law, Materials Science (M.S.), Psychology and Computer Sciences (C.S.). From the data, university takes the biggest proportion in the institution types in all five of the disciplines. Since there are only about 12000 university establishments in the world, the majority of the institutions filtered out were not universities. Universities were also the highest contributing institutions in terms of the number of papers published. In SSH disciplines, companies make no contributions. This is of no surprise since research in these areas is less linked to revenues. On the other hand, Pharmacology research receives the highest industry support. Pharmaceutical companies are actively involved in research as well as publishing their research.

WoS	Phar.	M.S.	Law	Psych.	C.S.
Company	12%	2.5%	0%	0%	8%
Research Centre	28%	10%	15%	10%	5%
University	60%	87.5%	85%	90%	88%
Total	100%	100%	100%	100%	100%

Table 4.4: Institution types of the WoS data. University is the biggest contributor in all of the disciplines; Psychology and Law have no company contributions while Pharmacology has the biggest company contribution.

4.2.3 Author Disambiguation

Although the overall aim of this study deals with research activities at the institutional aggregation level, it is necessary to obtain the author estimations in order to calculate certain descriptive statistics. The author estimations require disambiguation of the author names provided by the dataset.

Author name disambiguation is the problem of the non-uniqueness of people's names. Within a small group(*e.g.* with in a research group), the full names can probably identify people without ambiguity because the likelihood of two people named the same is very low. But when the full name is used to identify researchers globally, it is far from appropriate. For instance, authors with the same name cannot be distinguished using their name; people may change names over the course of their career(*e.g.* changing surname after getting married); many journals use first initial in their publications, making the author names even more ambiguous. Inaccurate author identification brings issues to publication based the author evaluation metrics (*e.g.* *h*-index), making them less authoritative and credible.

4.2.3.1 ORCID and ResearcherID

In recent years, this problem is becoming more urgent with accelerated research output and output based performance evaluations (*e.g.* the UK and Australian's research assessment exercise). Efforts are currently being made to resolve this problem from the root. A community based project, Open Researcher and Contributor ID (ORCID)[4], is maintaining a database of identifiers for authors. This identifier, unlike the author's name, is globally unique and does not change over time. Linking it with the Document Object Identifier (DOI)[5] of the author's publications, it creates this unambiguous relationship between authors and publications. A similar ID system, ResearcherID[6] is currently used by Thomson Reuters to allow authors to link to their publication in the WoS database. It is only early stages for these systems, and the data I have obtained from WoS do not contain any ResearcherIDs.

Until the ORCID and ResearcherID systems resolve this problem from ground up, we can only use heuristic approaches to this problem.

[4] http://www.orcid.org

[5] DOI is a currently well adopted document identification system. Each DOI uniquely identifies one document.

[6] http://www.researcherid.com/

4.2.3.2 Author Identification and Institutional Repository

With the open access initiative to make the research publications openly and freely available online, hundreds of institutions around the globe[7] have mandated their scholars to deposit the publications into their institutional repository. (Some take further steps to use these publications as promotion evidence.) The scholars within each individual institution are well identified, and they tend to have an institutional wide ID. Journals and conference proceeding publishers also make efforts to identify the authors within their own database. However, when these publication data was aggregated by the secondary publisher, *e.g.* WoS, authors cannot be matched up across different journals even if they have IDs in each of the journals, so the identified authors are often lost.

4.2.3.3 Fuzzy approach to Author Identification

Given a set of publications with the author's name attached, there are two practical problems in identifying these names. First is the multi-name format problem. That is, one individual's name may have multiple forms of spellings recorded in publications, including the name spelling variations, mis-spellings and OCR errors. The impact of this problem in bibliometric analysis is that it splits a single author's publications into several distinguished identities, splitting this author output. When conducting co-authorship network analysis, this split could also potentially disconnect a large, connected network into several small, isolated ones, thus strongly affecting the network structure and the subsequent observations. At the individual level, researcher metrics and evaluation methods, such as h-index could give false results due to incomplete data. On the other hand, conducting the same analysis at the institution level is not affected by this problem, because the institution level only count the papers and discards the number of authors. So there is no difference between 10 authors each publishing 1 paper, or 5 authors each publishing 2 papers. Both situations count as one institution publishes 10 papers.

The second is the duplicated-name problem. It is frequently seen in the Chinese name's English representation, where common surnames, for example Zhang with a short first

[7] For a complete and more up to date list, please refer to http://roarmap.eprints.org/

name (or just first name initial) clashes with different researchers with the same spelling. This problem has the opposite effect on the bibliometric and network analysis at the author level, where these common names stand out because they collect all of these researchers' publications into a single 'individual'. However this problem does not affect analysis at the institution level. At this level, author's name is not used as an identifier to merge the publications into, the institution's name is used instead. For example, Zhang's paper from University of Southampton is counted towards Southampton and Zhang's paper from University of Oxford is counted towards Oxford, instead of counting towards the common identifier Zhang.

Co-authorship and social network analysis techniques have been experimented in name disambiguation [99, 147]. It works based on the similarities of the same author's co-authors. For example, if several 'names' have the same or a similar set of co-authors, then these names are likely referring to the same individual and can be safely treated as the same person. The problem with this technique is that it is computationally expensive and unsuitable for large scale analysis.

4.2.3.4 ACM and WoS Author Identification

The ACM dataset contains an ID field to each authors. However, these 'IDs' do not uniquely identify individual authors as its name suggested. There were instances found that the same individual has been assigned with multiple of these IDs, splitting his/her work into multiple authors. In order to evaluate the extent of the ID's re-assignment, I have to use an alternative method to estimate the number of authors. (Figure 4.4)

The name conversion and collapse was an attempt to address the multi-name format problem while not introducing unacceptable duplicated-name problem. The institution-wide collapse assumes that first initial is a good enough identifier for authors in discipline within an institution.

The number of authors identified using this method is 323,419, which approximates the number of ACM IDs. For simplicity, ACM ID was used as the identifier for authors in the ACM dataset.

1. An author's institution is identified and matched to the corresponding Webometrics entry.
2. The authors' full names are converted to surname with first initial, which attempts to remove the different spellings of names on various publications.
3. The surname and first initials are merged *institution-wide* to limit the chances of mis-merging due to the shortened names.

Figure 4.4: Alternative method to estimate the number of authors in ACM dataset.

The WoS data has the authors represented in the form of surname with first initial without any identifiers attached. As a result, the duplicated-name problem is expected for common names, while the multi-name problem would be less likely to exist. However, the author information provided by WoS does not include affiliation, so the same technique of merging names within each institution was not possible. Author level analysis is not the focus of this study, some author ambiguity will not affect the outcome of this study. It was decided to use author names as provided.

4.3 Descriptive Statistics

The overview of the two datasets is presented in figure 4.5 organised by discipline. There are two datasets for Computer Science – ACM and Web of Science, each collects a different sets of journals. For Pharmacology, Materials Science, Law and Psychology, the Web of Science datasets were used.

	ACM C.S.	WoS C.S.	Phar.	M.S.	Law	Psych.
Period/Years	1957-2009/52	1973-2010/37				
Total papers	214,592	479,913	728,721	583,640	126,675	208,066
Mean papers per year	4,049	12,970	19,695	15,774	3,423	5,623
Institutional collaborative papers	20,043	164,553	277,939	214,491	11,543	68,141
Papers per author	1.56	3.85	4.16	4.38	1.99	2.77
Total citation counts	1,225,561	2,711,196	12,488,473	4,895,448	592,857	3,514,787
Papers received one or more citations	113,836	267,666	602,611	403,364	66,887	156,992
Mean citations per paper	5.71	5.65	17.14	8.39	4.68	16.89
Publishing universities	2,523	3,742	3,081	3,413	1,425	2,896
Total authors	334,150	310,683	729,427	452,191	78,373	175,731
Authorships	522,369	1,195,081	3,033,987	1,979,209	155,937	486,750
Authors per paper	2.43	2.49	4.16	3.39	1.23	2.34

Table 4.5: Dataset overview by discipline

Figure 4.6 lists the statistics based only on the collaborative papers in each discipline.

	ACM C.S.	WoS C.S.	Phar.	M.S.	Law	Psych.
Collaborative papers	20,043	164,553	277,939	214,491	11,543	68,141
Mean papers per year	542	4,447	7,512	5,797	312	1,842
Papers per author	1.41	2.80	2.77	3.18	1.40	2.07
Citations	138,138	1,066,609	4,802,108	1,938,636	80,790	1,348,530
Papers received one or more citation	12,534	103,356	236,208	161,273	7,361	55,880
Mean citations per paper	6.89	6.48	17.28	9.04	7.00	19.79
Mean institutions per paper	2.27	2.37	2.59	2.41	2.33	2.47
Authors	53,393	194,246	541,695	308,046	17,326	108,613
Authorships	75,548	544,655	1,498,862	978,275	24,219	225,142
Authors per paper	3.77	3.31	5.39	4.56	2.10	3.30

Table 4.6: Collaborative paper overview by discipline

4.3.1 Paper Productivity

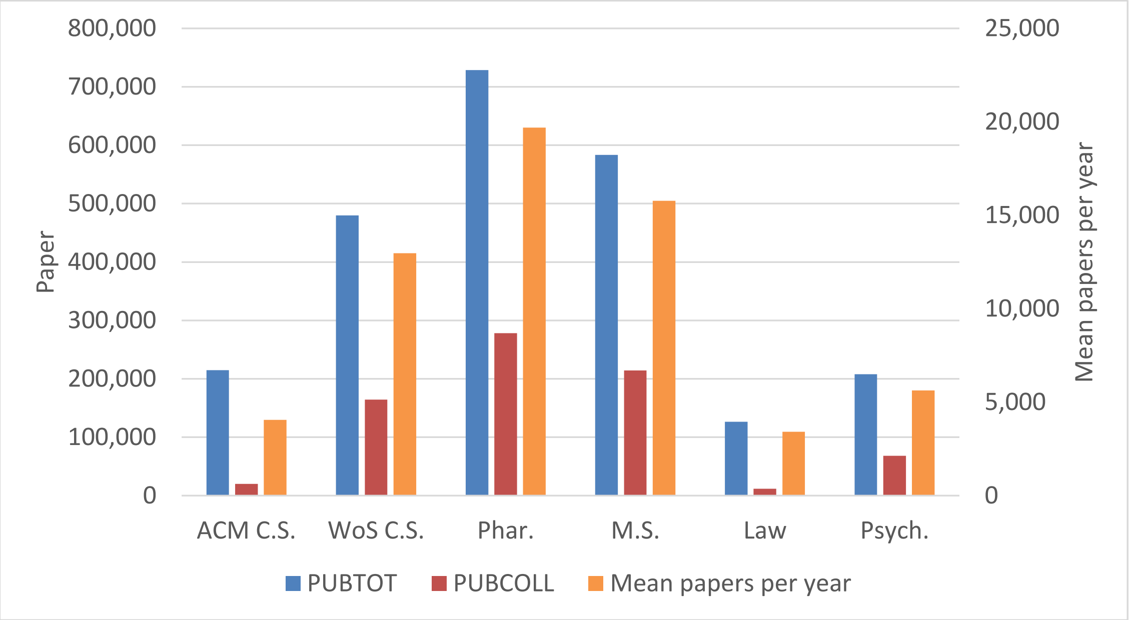


Figure 4.5: Paper productivity overview. Paper productivity varies across disciplines, demonstrating the disciplinary differences toward publishing. Productivity must be taken into account when comparing publication metrics across disciplines.

Figure 4.5 presented the number of published papers, collaborative papers and mean papers per year for each discipline. From the WoS dataset, Pharmacology has the most number of published papers in total, per year and in collaborative papers. The other two NSE disciplines – Computer Science and Materials Science also have above 400,000 paper productivity. On the other hand, Law and Psychology papers are much fewer. This could be because social science disciplines published in other formats which are not included in our dataset[109].

Fewer Computer Science papers were recorded in the ACM database than in WoS database, despite the ACM dataset covering a longer period of time (52 years vs 37 years).

We have seen large differences in paper productivity across disciplines, the number of unique universities involved in the publication stays about the same (Table 4.5 publishing universities). Most universities are involved in publication and play a big role in research output.

4.3.2 Citation

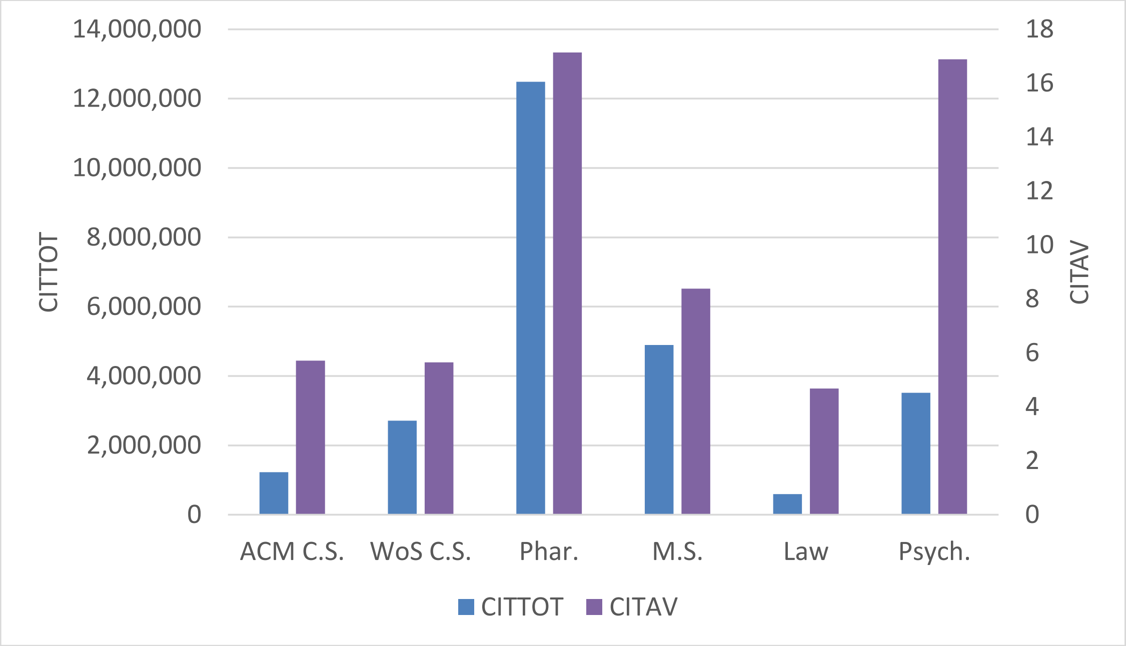


Figure 4.6: Citations received by each discipline in the period studied. Total citation shows a similar shape as the per paper citation in all disciplines, the more citations received in total, the more citations were received by each paper, despite the increase of the papers. Psychology papers receive much higher number of citations per paper, compared to two similarly cited datasets: Materials Science and WoS Computer Science.

Figure 4.6, Pharmacology received the highest number of citations to its papers, both in its total citations received by all of its papers and the mean citations per paper. Psychology only received a third of citations compared to Pharmacology, but its mean citations per paper is as high as Pharmacology. The remaining disciplines receive a fraction of the total citations of Pharmacology. Pharmacology has a much higher rate of publication and citation practice compared to the other disciplines, while Psychology is much higher cited.

The *mean citations per paper* has grown for each discipline once the non-collaborative papers are taken out of the statistics (Table 4.5 compared to Table 4.6). Collaborative papers in Law received an increase as high as 50% (from 4.68 to 7.00 citations per paper).

4.3.3 Collaborative Papers

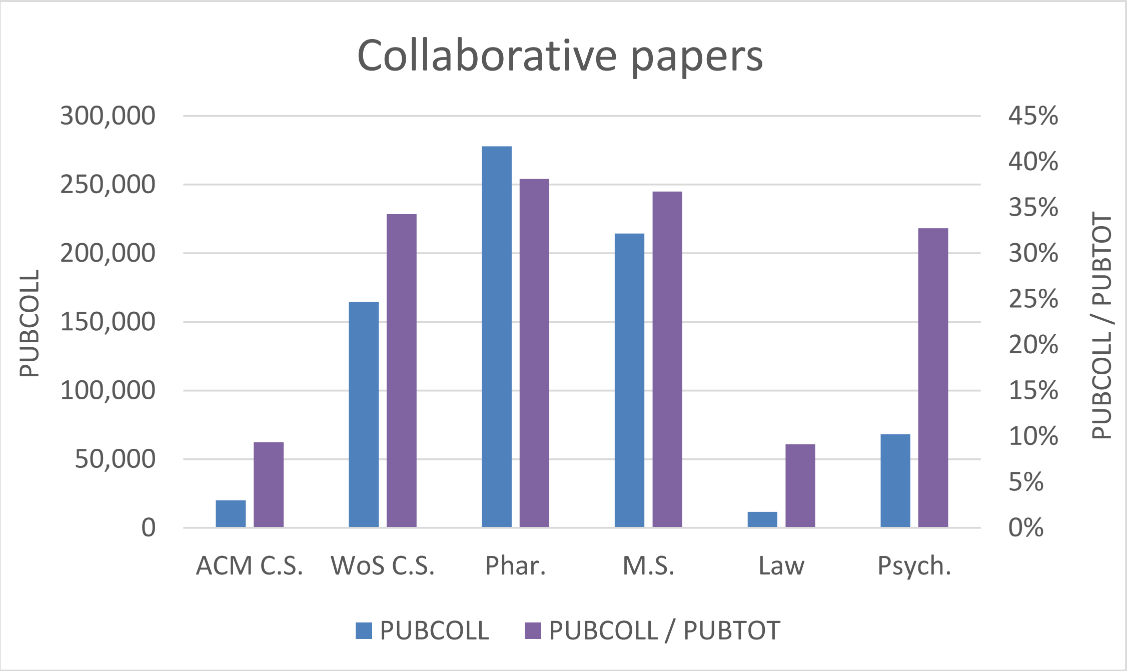


Figure 4.7: Collaboration Papers. The number of collaborative papers published by each discipline varies strongly. The proportion of the collaborative paper over the total paper is also discipline dependent, where Law has a very low proportion of papers collaborated.

In figure 4.7, the ACM Computer Science dataset contains fewer collaborative papers than the WoS Computer Science dataset, the ratio between collaborative papers and the total papers is also much lower. Pharmacology has the highest number of collaborative papers as well as the highest proportion of the collaborative papers / total papers. Psychology, while it has a smaller number of collaborative papers, its proportion of collaborative paper is much higher than ACM CS and Law. Law has the lowest proportion of collaborative papers of all disciplines.

4.3.4 Collaboration Size

In figure 4.8, the mean number of institutions involved in each paper is relatively stable across disciplines at about 2. At institution level, collaborations do not tend to be very large in these five disciplines. At the author level, the collaboration size varies slightly more, with Pharmacology the largest of more than 5, while Law has the smallest of 2. The size of the collaboration may be closely related to the mode of research in that the experimental based disciplines (*e.g.* Pharmacology and Materials Science) have a larger number of co-authors than those that are not [140].

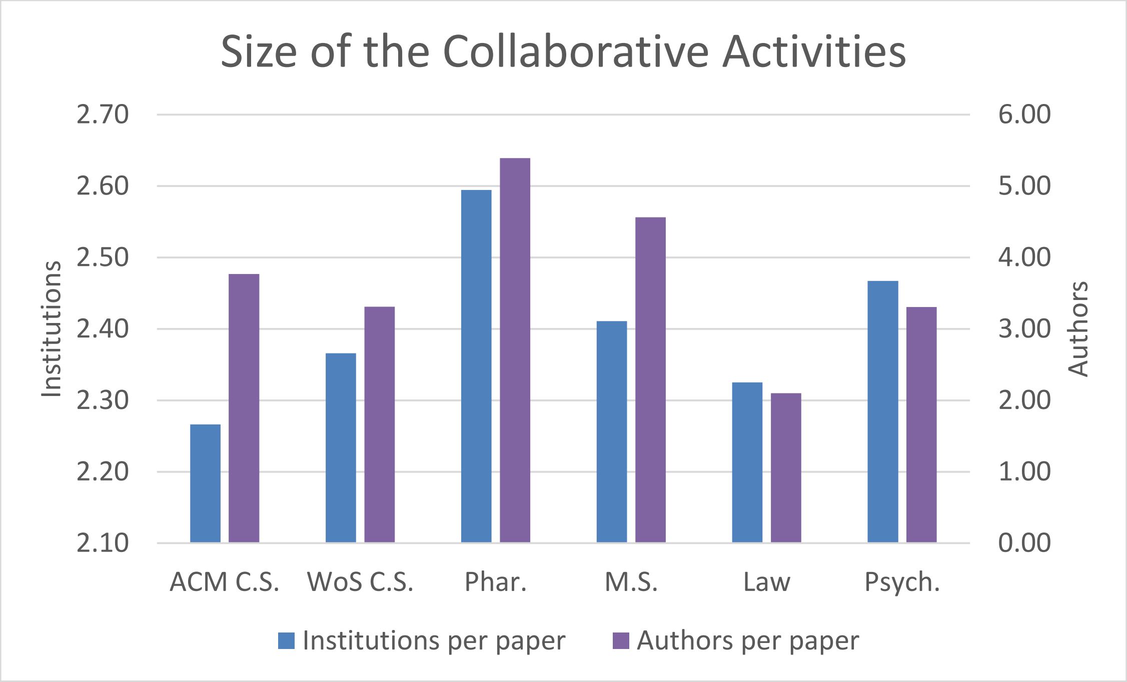


Figure 4.8: Collaboration Size. The mean number of institutions involved in each discipline is similar while the number of authors can vary across disciplines.

4.4 Metrics Selection and Counting Methods

The aim of the metrics selection is to include as broad a range of measurement in our analysis as possible, while striking a balance of not including too many redundant metrics. The limitation of the metrics selection in this case lay in the data availability.

In this section, I discuss how the metrics were selected to estimate the institution's productivity, impact and collaboration, and how these metrics were counted and aggregated.

4.4.1 Institutional Productivity Measurements

Research productivity was discussed and it was concluded in the literature review that it can be approximated by publication productivity. Publication productivity is the number of publications an entity (a researcher, an institution or a country) publishes within a period of time. The highly productive ones are those which publish more within the same period. Using publication productivity to measure research output is frequently used in previous studies [33, 69, 101].

It is important to recognise that since the input of publication was not factored into this measurement, productivity is not normalised, which means that larger institutions or larger countries which have more researchers may have a higher productivity due to their size.

To measure a university's publication productivity, papers published with the university name contained in the address is counted. The counting process used in this study is as follows: the university's productivity (***PUBTOT***) is incremented once if a paper has listed the university's name at least once. Multiple appearances of the same university on a single paper are counted *once* towards the university.

4.4.2 Institutional Impact Indicators

In section 2.3, we discussed that the overall research quality of an institution can be approximated from individual research's quality. The four ways to measure the quality of a piece of research are: methodological quality, report quality, peer review based quality and bibliometric. The first two quality measurements are very specific to disciplines and these measurements are difficult to use in a quantitative analysis. Peer-review based measurement, although it offers the authoritative expert opinions about the research, the associated cost is often much higher than the alternative. The UK and the Australian governments have conducted and published their nation-wide expert-review based institution research quality measurements. But these national data limited to only two countries are not useful in this study concerning the global context. The bibliometric method offers a practical and reproducible impact indicator, in particular, the citations offer impact at a quantitative level. Based on the availability of the data, three citation based metrics were chosen to measure the impact of an institution's research activities: citation per institution; PageRank weighted citation per institution, and citations per paper. In addition to bibliometrics, I have also included a university ranking – Webometrics Ranking of World Universities – in our study. It offers a completely different perspective to the citation based metrics and potentially gives us new insights.

4.4.2.1 Citations per institution

Citation count has been widely used in the literature as a research impact indicator [18, 93, 137]. It is easy to count and the data is readily available in databases. I include it to offer a baseline for the analysis.

In this study, the citation count is aggregated at the institution level and it is referred to as citations per institution (***CITTOT***). All of the citations received by a paper are counted towards the listed institution's citation. For collaborative papers that have multiple institutions, each of these institutions gets a copy of all the citations received by the paper. This is also referred to as the "full counting" in the literature.

However, the raw citation count has a few disadvantages: 1. The citation count varies across fields of study due to different scholarly practices, the size and the nature of the audience and the size of the community. As a result, the citation number is not comparable across disciplines. 2. The older the publication, the more time it has to receive citations, making a pattern that older publications generally have bigger counts, thus giving older publications an unfair advantage in measuring impact using citation counts. 3. The citations that come from low impact articles are weighted with the same value as the citations that come from high impact articles. In other words, citation count does weight the value of citations according to source impact. This is addressed by using a weighted citation count that is based on PageRank.

4.4.2.2 Citation PageRank

Citation PageRank (***CITTOTw***) is a qualitative improvement on the raw citation counting, in that it gives weighting to those cites that come from highly cited institutions. Compared to the citation count, not only the citation number is considered, but also how well cited the citing institution is. This metrics requires more strict source data since both the citing and cited institution need to be available to calculate the weighting.

The calculation of the citation PageRank was completed using the Network Workbench software package. A citation network of institutions generated from the dataset was the input to the program, and the PageRank value of the institution was the output of the program.

4.4.2.3 Citations per paper

Citations per paper (***CITAV***) normalises the citation measurement by dividing an institution's total citation counts by the total number of published papers. I include it in this study because it provides a productivity normalised impact for the institution. This normalised view can highlight institutions that have a low number of published papers, but a high number of average citations (*i.e.* high impact but low volume research).

Formula 4.3 calculates the CITAV for institution i.

$$CITAV = \frac{C_i}{P_i} \tag{4.3}$$

where C_i is the total citations received by all of the publications associated with institution i and P_i is the total number papers published by the institution. The citations per paper measures the average number of citations papers receive at an institution and estimates the average impact of individual papers.

4.4.2.4 University League Table

Four university league tables were studied in the literature review, but only the Webometrics ranking (***WRANK***) offers the complete world university ranking, so it is the only one included in this study. The July 2010 version of the ranking was used.

This version of the ranking is based on four aspects:

- University website's popularity. How many external websites are linking to it.
- University website's size. The number of webpages the website has.
- University's output and accessibility. The number of open format documents which can be found on the university's website.
- University's research output. The number of publications which can be found online.

The quality represented by the ***WRANK*** is not a single aspect of the university's quality, but a weighted mixture of four sub-metrics. (In fact these sub-metrics are evolving over the years, with latest ranking measures a slightly different set of these sub-metrics [8].) This "non-pure" quality indicator may give high correlations with our other measurements, without necessarily being related to the university's quality. For instance, Webometrics assigns 20% of the score for the number of papers published by the university, which is expected to be correlated with the ***PUBTOT***. It is important to recognise the features of this indicator before making interpretations.

League tables generally put high ranking universities in front, with smaller rank values. (*e.g.*, rank 1 is better than rank 20.) To make this metrics consistent to the other metrics and make the results easier to interpret, the ***WRANK*** is inverted, so that a higher ranking value indicates the better universities.

[8] Please see http://www.webometrics.info/ for the latest sub-metrics used in ranking universities

4.4.3 Institutional Collaboration Measurements

Using the institutional collaboration model discussed in section 2.1.3, these are the two important factors affecting the strength of the institutional collaboration: 1. the number of times institutions have co-authored papers; 2. the size of a collaboration. The collaboration measurements used in the literature mostly considers the frequency of institutions' collaboration, without paying attention to the individual collaboration's size. Both of these factors are measured in this study. In addition, I also include a measurement of the institution's proportion of collaborative papers. This measurement helps us to address the questions related to whether shifting to a collaborative mode of research alone would improve the productivity and the impact.

4.4.3.1 Number of times collaborated

This indicator is frequently used, it counts the number of collaborative papers the institution has published. This directly reflects the institution's involvement in collaboration. I denote it as $\boldsymbol{PUBCOLL}$. Similar to the publication counting and citation counting, "full counting" was used, which means one collaborative paper is counted as many times as the number of distinct institutions presented in the paper. For example, a paper by Southampton and Oxford University is counted once for each university, even if there are multiple authors from either university. This measurement provides a baseline for the collaboration measurement.

4.4.3.2 Size-weighted collaboration

Building on the collaborative paper count, size-weighted collaboration ($\boldsymbol{PUBCOLLw}$) aimed to add collaboration size into the metrics. The more authors a paper has, the more self-citation as well as the their colleagues' citation would potentially be guaranteed, driving up the impact.

In a collaborated paper, there are two author-number related parameters:

- the total number of authors participating in the collaboration. For instance, a paper with 10 co-authors forms a larger collaboration than a paper with 2 co-authors.
- the number of authors the institution has brought to the collaboration. For instance, an institution with 5 authors in a collaboration is contributing more compared to an institution in the same collaboration with 1 author.

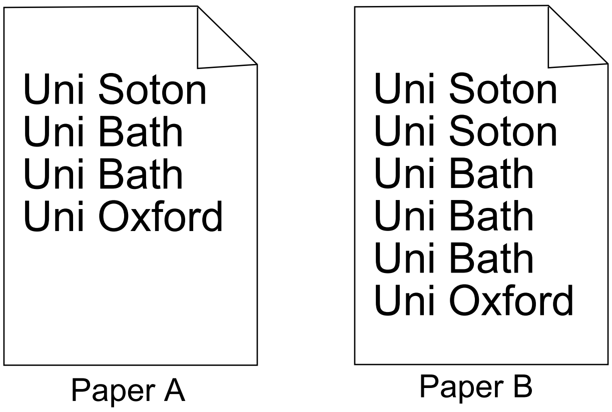
Both of these collaboration variables are proportional to the institution's measured size-weighted collaboration, and they are summarised over all of the papers an institution produced. When calculating an institution's contribution to the paper, to avoid double counting the institution's participating researchers, the authors from the institution are subtracted. Formula 4.4 calculates size-weighted collaboration for institution i.

$$CS_i = \sum_{pi} A_{pi} \times (TA_p - A_{pi}) \tag{4.4}$$

where A_{pi} is the number of authors from institution i on paper p, and TA_p is the total number of authors for paper p.

For example, figure 4.9 shows two collaborated papers, both have the same number of institutions, but have different *total authors* and different number of *authors from each institution*. All three institutions have the same $\boldsymbol{PUBCOLL}$ count, because each paper give them 1 collaborative paper, so the $\boldsymbol{PUBCOLL}$ count is 2 for Southampton, Bath and Oxford with no differentiations. $\boldsymbol{PUBCOLL}$ count ignores the fact that paper B was a bigger collaboration and Bath had contributed the most in both of the papers. $\boldsymbol{PUBCOLLw}$ will be able to differentiate based on their collaboration sizes.

Using formula 4.4, we calculate the size-weighted collaboration for the three institutions (Figure 4.9). In Paper A, $\boldsymbol{PUBCOLLw}_{Soton}$ is the same as $\boldsymbol{PUBCOLLw}_{Oxford}$, while less than $\boldsymbol{PUBCOLLw}_{Bath}$, because Bath contributed 2 authors in the paper A while Southampton and Oxford each contributed 1. Oxford has contributed the same number of authors in both Paper A and Paper B (1 author for paper A and 1 author for paper B), it gets a higher $\boldsymbol{PUBCOLLw}$ value in Paper B because Paper B is a



$$\begin{array}{ll} \boldsymbol{PUBCOLLw}_{Soton} = 1 \times (4-1) = 3 & \boldsymbol{PUBCOLLw}_{Soton} = 2 \times (6-2) = 8 \\ \boldsymbol{PUBCOLLw}_{Bath} = 2 \times (4-2) = 4 & \boldsymbol{PUBCOLLw}_{Bath} = 3 \times (6-3) = 9 \\ \boldsymbol{PUBCOLLw}_{Oxford} = 1 \times (4-1) = 3 & \boldsymbol{PUBCOLLw}_{Oxford} = 1 \times (6-1) = 5 \end{array}$$

Figure 4.9: Size-weighted collaboration calculation

larger collaboration with more authors involved. Treating these two papers as a dataset, Bath is the top in ***PUBCOLLw*** as it has contributed the most in both of the papers; Southampton is the second while Oxford is the third.

4.4.3.3 Percent collaboration

The percent collaboration is calculated using the ratio between total collaborative papers and total papers (formula 4.5). An institution with high collaborative paper to total paper ratio pays more attention to collaboration. This variable allows us to learn whether institutions that are involved in more proportion of collaborative research are linked to impact or productivity.

$$PUBCOLL\%_i = \frac{PUBCOLL_i}{PUBTOT_i} \tag{4.5}$$

where $PUBCOLL_i$ is the collaborative papers for institution i and $PUBTOT_i$ is the total number of papers for institution i.

4.5 Correlation Methods

Given two sets of variables or data, correlation analysis finds the statistical relationships between them. It is a widely used method in bibliometric studies [18, 51, 102, 137, 154–156, 159]. The correlation analysis is implemented by a few algorithms, but the underlining assumptions of each algorithm is different. It is important to learn about the assumptions made by the algorithms and it can help us avoid wrong conclusions about the relationships between the data. For example, an exponential relationship between two variables can be mistakenly tested significant using a linear correlation analysis. So before choosing a correlation method, it is often useful to plot the distribution of the variables and visually identify that the distribution is not far from the assumption used by the correlation method. The correlation measurement is often represented in a triple of three values (Figure 4.10).

$n\ r\ p$

n – the size of sample.
r – the correlation coefficient, varies from -1 to 1, where -1 means total negative correlation and 1 means total positive correlation. 0 means no correlation between the pair.
p – the probability of such correlation occurs by random chance. When $p > 0.05$, it is often considered that such correlation is not significant, therefore, no correlation exists between them.

Figure 4.10: The correlation measurement triple

Given the distribution of the variables, we discuss two correlation analyses: linear correlation and non-parametric (non-linear) correlation.

We must emphasize that correlation does not mean causation. A pair of variables correlating with each other provides no information on whether one causes (happens before) the other. It is important to not interpret correlation as causation. Correlation can be understood as a prediction. Given a reading of variable A that correlates with variable B, we can predict, within certain range, the value of variable B. The same can be done from B to A. The two variables are correlated, but it is not enough to determine whether one causes the other.

4.5.1 Linear Correlation

Linear correlation measures the linear dependence between two variables. Pearson's product-moment correlation coefficient, or Pearson's r is the most used algorithm in literature to measure linear correlation. Mathematically, Pearson's correlation coefficient is defined as the covariance of the two variables divided by the product of their standard deviations.

$$r_{x,y} = \frac{covariance(x, y)}{\sigma x \sigma y} \tag{4.6}$$

The assumption of this correlation is that the two variables (x,y) are distributed normally, that is a distribution where 95% of the values lie within two standard deviations of the mean and the plot appears as a bell on a frequency distribution diagram. However, this is often not the case for many of the variables used in this study. When the variables to be correlated do not naturally distribute normally, one common approach is to apply a transformation to the variables, so that the eventual distribution resembles a normal distribution.

An invertible function is used to transform this function to obtain a normal distribution. It is a mathematical function such that the original values can be calculated by applying the inverse of the function. In other words, the transformation can be undone by another function. The commonly used functions include square, quadrupedal and square root. To check if the resulting distribution is normal, both visual techniques (*e.g.* by plotting the transformed variables in a histogram) and numerical tests (*e.g.* Kolmogorov-Smirnov test) are commonly used.

4.5.1.1 Box-Cox technique for selecting transformation function

The process of trying each invertible function in the hope that it will convert variable distribution into normal distribution is slow. Especially as there are nearly 10 variables which need to be transformed in this study.

Box-Cox technique integrates the process of selecting a transformation function, applying the function and evaluating the quality of the normal distribution after the transformation.

The functions often used to transform the variables are a class of functions based on the equation 4.7

$$y = x^{\lambda} \tag{4.7}$$

The square root and square functions are special cases when $\lambda = 1/2$ and $\lambda = 2$. Using equation 4.7, it is possible to construct power functions at finer steps in order to find a function that produces the best transformation given the data distribution.

I have implemented this procedure in SPSS with the starting λ set to -2.1 and the difference between each λ set to 0.1 (these values were found to be the most appropriate, although both of them were configurable). The most appropriate transformation was selected by visual observation according to the histogram plot as well as the Kolmogorov-Smirnov test.

4.5.2 Non-parametric Correlation

Having normally distributed data is not always possible even after a transformation. Non-parametric correlations were developed to address these cases. Spearman's non-parametric correlation describes how well the two sets of data can be mapped with a monotonic function. It assumes that the pair of variables to be correlated are measured at least on a rank order scale. If a pair of variables have positive Spearman's correlations, it means that the *rank order* of the variable can be predicted from the *rank order* of the other variable. The Spearman's correlation coefficient also varies from -1 to 1 and when two datasets have no correlation the coefficient is 0. The Spearman's correlation is a weaker correlation compared to Pearson's correlation, because only the rank order is predicted from one variable to another, while the actual value of the variable is not predicted.

4.5.3 Null Hypothesis and Significance Testing

Researches in quantitative methods have developed very rigorous practices in examining correlations. When determining the correlation between factors, a null hypothesis is first assumed. that is, there is no correlation between the pair of variables under examination unless shown otherwise. Only if the result is showing that the null hypothesis is false, the alternative hypothesis (i.e. the hypothesis that the correlation exists) is proven true. I will adopt this null hypothesis technique throughout this work.

In practice, correlation analysis of two independent variables may never be exactly 0. When the sample size (n) is small, the chance of having a sizable correlation coefficient by random occurrence is quite large. A non-zero correlation is not always an indication of a relationship between variables. Significance testing (p) of a correlation result is the process of determining the likelihood that a coefficient occurred by chance. This significance testing is dependent on the size of the sample, the larger the sample size, the smaller the likelihood that the correlation occurred by chance. $p < 0.01$ was used in many previous bibliometric analyses [102, 156] as a cut-off point for correlation to be significant. It is used as the deciding probability for most of the correlations, but some have used larger probabilities, which will be indicated next to the coefficient.

4.5.4 Normalisation

Normalising a variable is also referred to as “correcting” the variable against some variations. It is a process of removing or partialling the effect of unwanted variables from the variable. In making real world measurements, the measured variables are often intercorrelated with other variables that we would like to isolate the effect from. For example, the length of a person’s hair has been found to have significant negative correlation with a person’s height. However, if we control the gender of the person (by splitting the data into male and female), the correlation disappears. The significant correlation between the length of hair and the height really lies between the genders. In order to give insight into the real factors that relate to the correlation, normalisation is necessary.

If the unwanted effect is categorical, (*e.g.*, male and female), we could split the data into the categories. If the data is continuous, division is one of the common methods. The denominator is the effect to be removed and the numerator is the variable to remove it from. Two of our variables in study are in fact division of two existing variables: citation per paper – division of citations over total papers; and percent collaboration – division of collaborative papers over total papers. The division of the existing variables creates new interpretable variables that describe the institution and offer new insight. In addition to division, linear regression offers an alternative.

4.5.4.1 Linear regression

In statistics, linear regression is the process of finding the linear composition of the dependent variable (y) from a list of independent variables ($x_1...x_n$). Equation 4.8 describes this relationship. Linear regression assumes the relationship between the dependent variable and the independent variable(s) is linear.

$$y = \beta_1 x_1 + ... + \beta_n x_n + \epsilon \tag{4.8}$$

Where y is the dependent variable and $x_1...x_n$ are the independent variables, ϵ is the residual. The sum of all the residuals is zero in a linear regression to minimise unpredictability.

The residual ϵ is the unpredicted part of the dependent variable y from $x_1...x_n$, which is the effect of the variable y with variables $x_1...x_n$ removed. The technique is applied before the partial correlations are calculated.

4.6 Network Methods and Visualisation

Network visualisation techniques have been widely used in the field of bibliometric, data analysis and network analysis [32, 165, 201]. They have been used to visualise the relationships and to help spot interesting patterns that are otherwise buried in massive data.

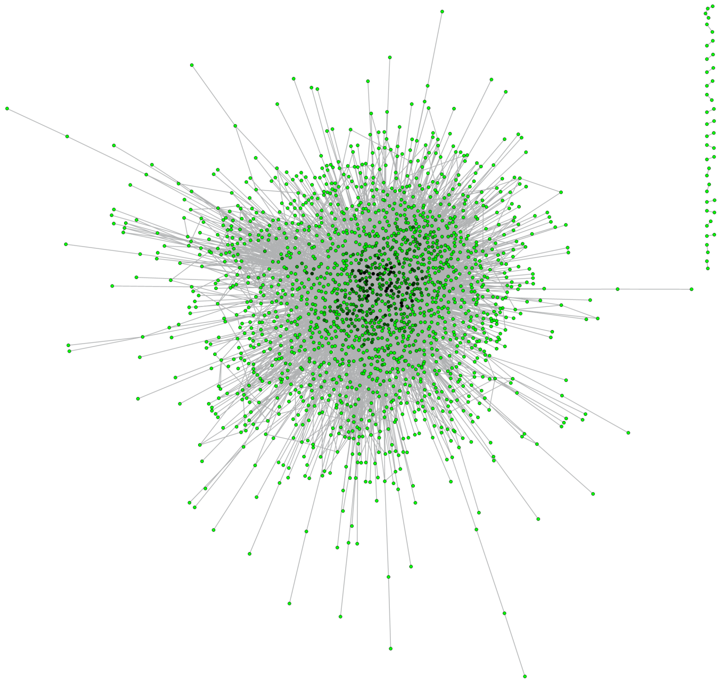


Figure 4.11: The institutional collaboration network recorded by ACM between 1951-2010. The nodes represent institutions and the edges represent collaborations between the institutions. Total nodes: 1843, total edges:12615, diameter: 9, average shortest path: 3.32, there are 21 connected components with 0 isolates. The largest connected component consists of 1801 nodes, which is 97.7% of the total. Nodes listed on the right are disconnected islands. The colour of the node represents its degree (how many other nodes this one connects to) the darker, the higher the degree.

A graph consists of nodes and edges, where edges are connected by nodes. In the context of this work, the nodes represent institutions and the edges represent the collaborations between them.

We attempt to visualise the relationships using the ACM Computer Science data. The institutional collaboration network is constructed, the institutional impact as well as the institution's country is visualised using node sizing and node colouring techniques.

4.6.1 Collaboration Network Analysis on ACM Computer Science

The institutional collaboration graph was constructed based on the set of inter-institutional collaborative papers in the ACM dataset. Figure 4.11 shows the collaboration network as recorded by ACM between 1957 and 2010. The colour of the node represents the number of institutions it has collaborated with, the darker the colour, the more institutions it has collaborated with.

There are in total 1843 nodes (institutions) on the graph, 1801 nodes are connected and form the largest component. The diameter of the largest connected component is 9 and the APL is 3.32. The number of dark coloured nodes is low compared to lightly coloured ones, meaning that institutions with a large number of collaborators are a small portion in the network.

The global institutional collaboration network for Computer Science is well connected. The collaborating institutions form a large connected component consisting of 97.7% institutions. On average there are only just over 3 steps for any institutions to connect to any other. This is a short distance considering that there are nearly 2000 of them and the condition for the connection are strong (compared to, for example, acquaintanceship.). This means information not only can reach most of the institutions in the network, but it can also reach them quickly and efficiently.

On the other hand, although information has passed on to a researcher in the target institution through some form of *inter*-institution collaboration, this information needs to travel further through the *intra*-institution communication network to reach the target researcher. The efficiency of the *intra*-institution network depends on the organisational structure of the institution and it is out of scope for this study.

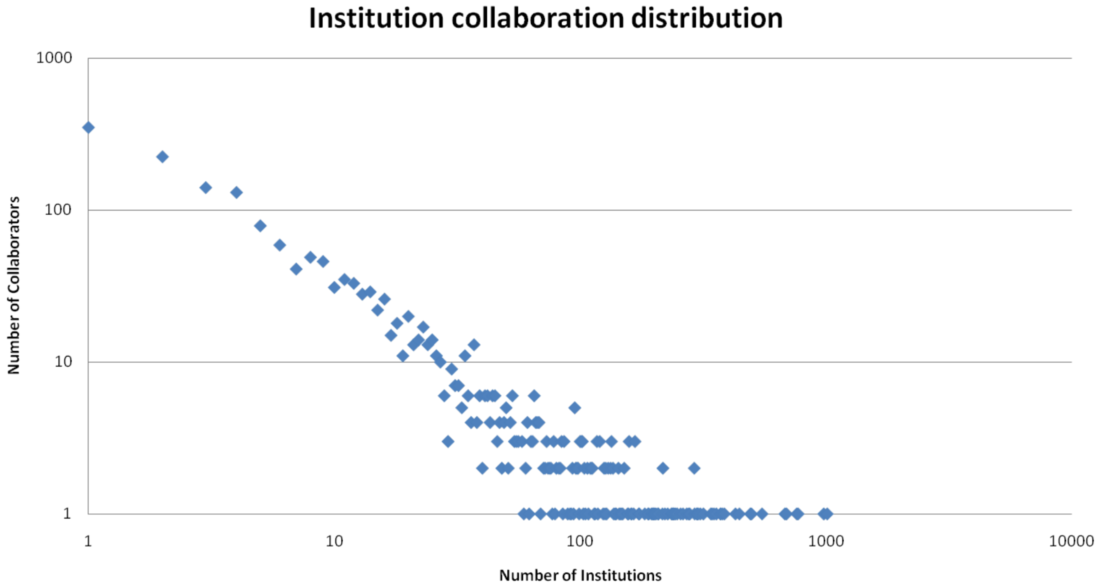


Figure 4.12: ACM Institution collaboration distribution. Each dot is a plot of the number of institutions having the number of collaborators. It is plotted on an exponential axis. A fat-tail plot is shown, indicating a power law distribution of the institution's collaborators.

The majority of the institutions do not have many collaborators, while a few institutions have a lot of collaborators. This is reflected in figure 4.12. The figure showed a fat tail plot, where many institutions only have a few collaborators, while only a few institution in the ACM dataset had more than 100 collaborators. This type of network degree[9] distribution exhibits power law degree distribution. Power law degree distribution is often found in networks where the rich get richer, and the structure and the connectivity of the network is robust to random node removal [65, 100]. "Richer gets richer" means that institutions that already have many collaborators will attract more collaborators, leaving those institutions with fewer collaborators even fewer opportunities to collaborate.

4.6.2 Improving the Visualisation

Excessive numbers of edges in a graph visualisation can obscure the discovery of interesting patterns. e.g., figure 4.11 is not very informative due to the large number of edges presented in the graph. Removing excessive edges is a common network analysis technique to improve the visualisation. We present here the three methods commonly used.

4.6.2.1 Threshold on Edge Weight

This is the naive method to remove excessive edges on a network where edges are weighted [204]. In an edge weighted network, values are attached to edges between the connected nodes. Depending on the meaning of the value, higher or lower valued edges are favoured over the other. In our institutional collaboration network, edge weight represents the number of publications the two institutions have collaborated on. The more publications the institutions have worked on together, the stronger the relationship between the two; on the other hand, institutions that have only worked on few publications indicate a weak collaboration relationship. A threshold value is determined based on the number of edges needing to be removed in order to show a clearer graph.

[9] Degree of a node is the number of other nodes it connects to. In this network, degree value is the number of collaborator the institution has

In practice, there are two main problems with this method. First, it does not take in to account the original network structure. A collaboration network is a type of social network, where the important information lies not only in the number of times two institutions have worked together, but also in the position where the institution is situated in the network, and also how many other institutions it has connections with. These structural features of the network are as important as the edge weight connecting the nodes together. The second problem is that it creates a large number of isolated institutions, rendering the remaining graph less informative. Due to these major flaws, the ACM institution collaboration graph with edge weight thresholding did not led to useful results and therefore is not shown.

4.6.2.2 Minimum Spanning Tree

The concept of the spanning tree comes from graph theory [123]. A spanning tree of a graph $G = \{V, E\}$ is a subgraph of G containing all nodes V and a set of edges $F \subset E$ such that these edges connect all nodes, but do not form a cycle. A minimum spanning tree is a spanning tree such that the sum of its edge weights is minimum. In our situation, edge weight is a positive indicator, the higher the weight, the stronger the collaborations. So we want to calculate the maximum spanning tree where the sum of the edge weights is maximum. Minimum spanning tree is often not unique for a given graph, there can be many minimum spanning trees due to alternative routes having the same value. This potentially causes problems and is therefore not applied.

4.6.2.3 Path Constraints

Another method to reduce the edges and to improve the clarity of a graph is to impose constraints on the path between the nodes. A path is defined as a series of edges that connects two nodes in a graph. Paths that do not satisfy defined constraints are excluded in the resulting network. Pathfinder network-scaling is a widely used path constraints algorithm. The underlying concept of pathfinder networks is pairwise similarity. Given a network with edge connect nodes to represent proximity, pathfinder extracts edges that best describe the core similarities in the network.

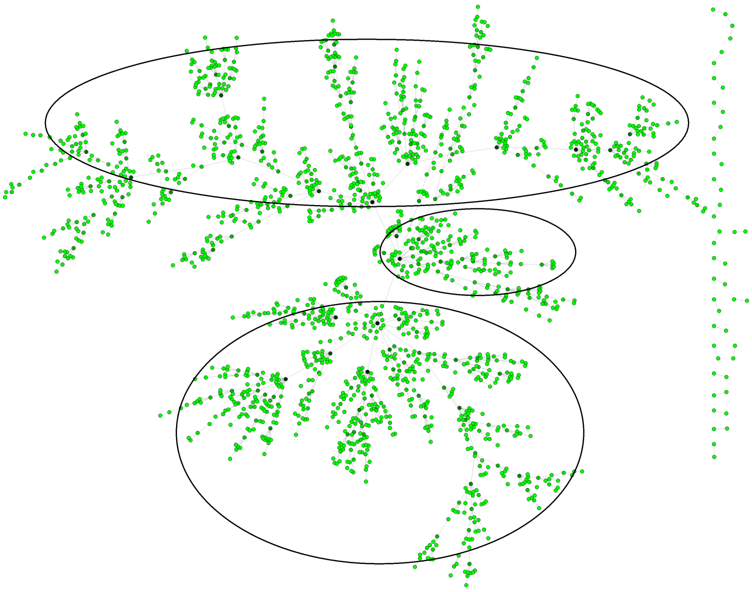


Figure 4.13: Institutional collaboration Graph Applied PathFinder. The three most collaborative institutions in this dataset – MIT, UC Berkeley, and Carnegie Mellon University – form the root of the three sub-trees (circled). The top tree is rooted at UC Berkeley, the centre tree is rooted at MIT and the bottom tree is rooted at Carnegie Mellon University

Pathfinder relies on the triangle inequality to eliminate redundant edges. Given two edges or paths in a network that connect two nodes, the edge or path that has a greater weight by Minkowski metric [136] is preserved. In our case, the stronger collaboration. Two parameters affect the output of a pathfinder network. the R-parameter influences the weight of a path based on the Minkowski metric. In other words, R is part of the formula used to calculate the path weight. The Q-parameter defines the number of edges in alternative paths up to which the triangle inequality must be maintained [168, 169].

Pathfinder network-scaling is used by many knowledge analysis works. Chen and Carr [41–43] applied the technique on author co-citation analysis and demonstrated the effectiveness of reducing edges in the network. Börner *et al.*[29, 44, 45] used the technique to visualise the knowledge domains and semantic spaces.

4.6.3 ACM Institutional Collaboration Graph

Figure 4.13 is figure 4.11 applied pathfinder with $R = \infty$ and $Q = N - 1$. After removal of the excessive edges by applying the path finder algorithm, a clearer representation of the structure of the collaboration was unveiled. The number of edges was reduced from

12615 to 1822. The three most collaborative institutions in this dataset – MIT, UC Berkeley, and Carnegie Mellon University – form the root of the three sub-trees circled out in Figure 4.13.

In the following sections, the node size and node colour are altered to reflect the collaboration, country of origin, and quality to visualise the relevant effects.

4.6.3.1 Visual Analysis of Institution's Country

Figure 4.14 colours the node by the country of the institution's origin. A homophily effect on country is clearly observed: on the left top corner, there is a "Japanese tree"; on the lower bottom there are two "Brazilian trees"; on the right centre is a "Korean tree". US universities dominate the entire graph (dark red nodes) – almost every corner has a few US universities, demonstrating the breath of its research in Computer Science and its connection to almost every country. They also have teams of universities attacking particular areas of research such as the right top "US tree" rooted at Villanova University and the right bottom tree rooted at Purdue University.

In contrast, European universities do not form observable large single-country trees similar to Asian countries do in the graph. They are inter-connected with each other. For instance, the UK universities circled out on the right side of the figure 4.14 are immersed in universities from Spain, France, Portugal, Belgium, Sweden, Denmark, Netherlands, Turkey, Liechtenstein and Russia. This effect may well be a result of the European Union's funding strategies, which promotes collaborative projects within the member states. On the other hand, it appears that *intra*-national collaboration within Japan, Brazil and Korean in Computer Science is very strong, which could be related to the funding policies of these countries.

4.6.3.2 Visual Analysis of Institutional Impact and Collaboration

Figure 4.15 shows the same graph as in the figure 4.13, with the node size representing institution's total collaborative papers.

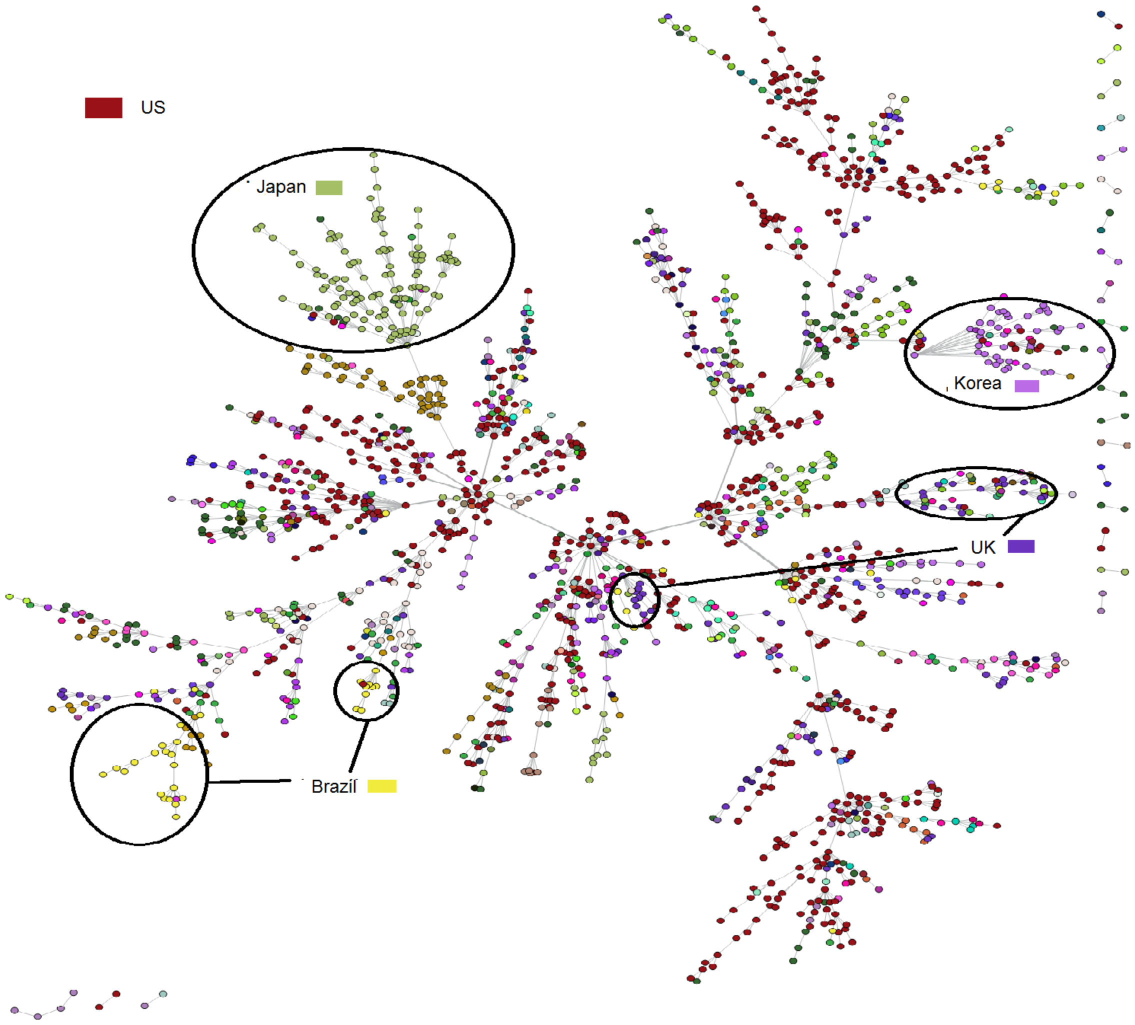


Figure 4.14: Institution collaboration graph colour by country

This graph shows interesting patterns: the main branches of the graph have big, green nodes, while leaves of the graph are small and darker nodes. This suggests that the core network of the institutional collaboration is held together by high quality and highly collaborative institutions. These institutions have strong channels with each other, like the main pipes connecting knowledge transfer across the world; those ranked lower and less collaborative institutions depend on the more specific areas of research, attached to one of the trunk institutions, and the even less collaborative ones attached to them.

The correlation of higher ranked universities which collaborate more and are lower ranked which collaborate less is visually presented on the graph. In fact, this correlation is so strong that we can use it to spot errors in our university recognition algorithm. For example, the large dark node located at the right top of graph appears to be too big to

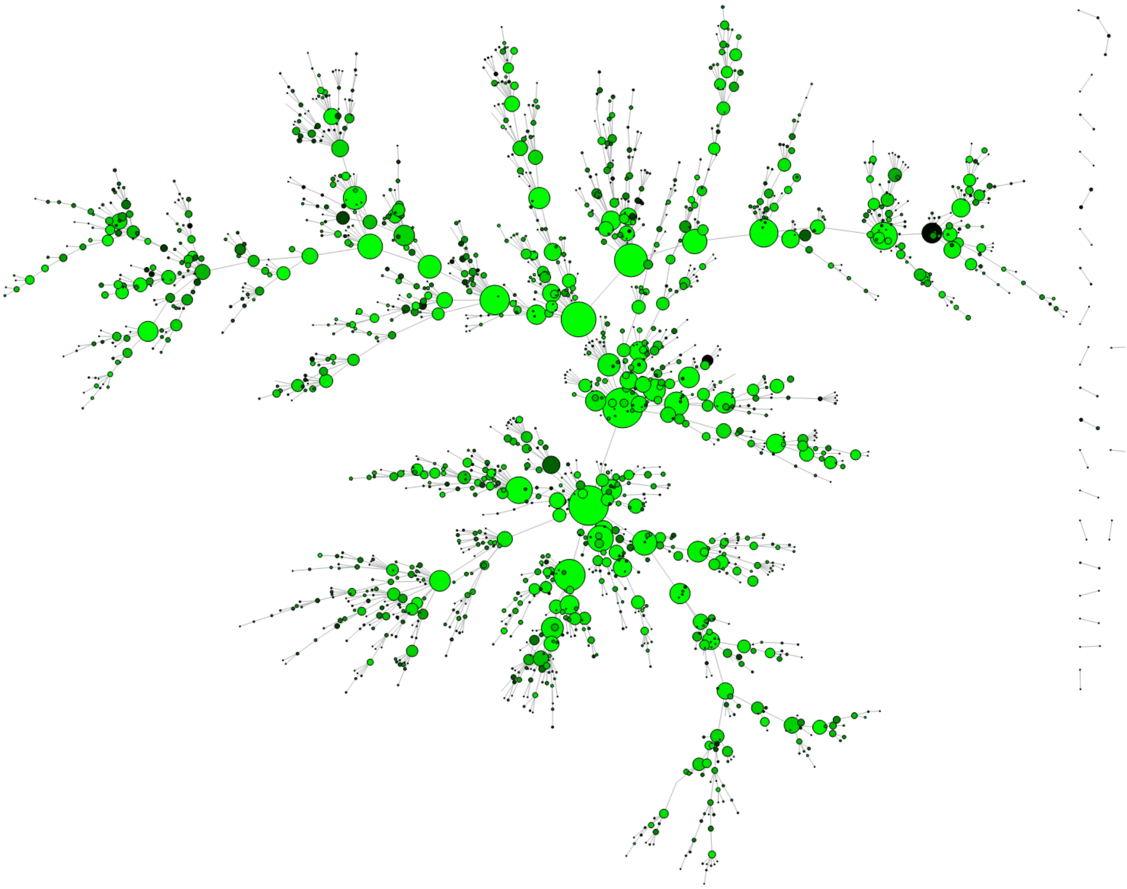


Figure 4.15: Institution collaboration graph with colour representing ranking and size representing collaboration

be dark. We looked into it, the node represents Indiana University, which is a multi-campus university system in the US. Overall it is ranked 25th in the world, which would be coloured light green in our graph. However, instead of using 25th as its ranking, our algorithm chose to use rank 7373, which is the rank for one of its small campuses, resulting in the dark colour; similarly, on the top left corner, the dark node is "National University", which is a name used by many universities in various countries and it has confused our matching algorithm.

4.7 Chapter Summary

The preceding discussion details the dataset, the preprocessing procedures, and the statistical methodologies, which will be used in the coming chapters. The two datasets – WoS and ACM – required different approaches in unifying the university addresses due to the different levels of data cleanness. A master university list(based on Webometrics Ranking) was used to align both datasets to, so that the problem of university matching across datasets becomes matching from a dataset to the master list. The matching percentage of the addresses to the master list from each dataset was calculated and evaluated. Improvements were made to the process until the matching rate was satisfactory. The university name variation lookup table resulted from this process could potentially be useful in other studies trying to identify universities from their addresses.

The author disambiguation problem and previous attempts solutions were discussed. But author level analysis is not the focus of this study, so no advanced author disambiguation algorithm was implemented.

A descriptive statistic table for both of the dataset across five disciplines were presented. Two sets of data were computed, one for all papers and one for institutional collaborative papers. Some features of the disciplines were discussed.

Three institutional collaboration variables, four institutional impact variables and one productivity variable were selected based on the literature and data available in the datasets. The aim was to include as many relevant metrics as possible for each of the institution's research activity. The counting method for each of the metrics was also presented.

The statistical analysis methods which are frequently used in the literature were presented. Correlation, normalisation, null hypothesis and significance testing were introduced. These methods will be applied in the coming chapter to find whether correlations exist between research activities metrics.

Finally, we applied social network analysis techniques on the institutional collaboration graph and we used the ACM Computer Science dataset. We demonstrated that these techniques can be effectively applied to networks based on publication data. Various graph drawing techniques were also applied to resize the nodes and to change the nodes' colour based on institution's country. A strong homophily effect was observed in the collaboration network, where institutions from the same country collaborated more frequently. The correlation of the institutions' productivity and collaboration was also intuitively visualised.

The social network analysis techniques were able to reduce the mass numerical data into effective visualisations and extract the key information. However, visualisation techniques often prone to deliberate manipulation to show specific patterns, it is not a strong enough evidence on its own to prove relationships. In the following chapters, we use linear correlation and partial correlation to confirm the existence of the relationships.

Chapter 5

Relationships Among Research Productivity, Research Impact and Research Collaboration

In this chapter we examine the relationship between the institutions' research productivity, research impact and research collaboration by determining their correlations. We use the variables identified in previous chapters to measure the three main factors: collaboration is measured by number of collaborative papers (***PUBCOLL***), size-weighted collaboration (***PUBCOLLw***) and percent collaboration (***PUBCOLL%***); productivity is measured by total institutional paper output (***PUBTOT***); and impact is measured by citations per institution (***CITTOT***), PageRanked citations per institution (***CITTOTw***) and citations per paper (***CITAV***). [1]

Each pair of the variables are analysed separately, and their pairwise non-parametric correlation is calculated. Pairwise correlation reveals the general relationships between the pair of variables. A positive relationship between the variables is shown as a significant positive coefficient; while a negative relationship are shown as a significant negative coefficient. If the two variables do not have any linear correlation, then a non-significant

[1] For clarity, the bold typeface of the abbreviations represent the unfiltered raw variable, as opposed to the non-bold typeface that used to represent the partialled variables used in the partial correlation analysis in the next chapter.

coefficient would be found. Many previous studies in this area used pairwise correlation [155, 159], so the results presented in this chapter can be compared and contrasted with them. This is the initial attempt to understand the relationship between these phenomena. The pairwise correlation result also provides a basis for comparing with the partial correlation result in the next chapter.

Three factors yields three top level pairs of correlations:

- collaboration versus productivity
- productivity versus impact
- collaboration versus impact

Since each factor is represented by multiple variables, the actual pairs of correlations further splits into the variable pairs (Figure 5.1). These correlations were then repeated for all 6 disciplines included in this study.

	CITTOT	CITTOTw	CITAV	PUBTOT
PUBCOLL	✓	✓	✓	✓
PUBCOLLw	✓	✓	✓	✓
PUBCOLL%	✓	✓	✓	✓
PUBTOT	✓	✓	✓	

Table 5.1: Variable pairs for pairwise correlation computation. All permutation of the variables describing different factors are computed for correlation.

The remaining of the chapter is organised as three sections presenting the correlation results by each pair of factors: collaborativity vs productivity, productivity vs impact and collaborativity vs impact . The results are discussed and a summary is presented at the end of this chapter.

5.1 Collaborativity versus Productivity

In 1966, Price and Beaver [159] found positive relationships between co-authorship and the number of publications at the individual author level. They found that the authors who published more papers are also those that published more co-authored papers. Similar observation was also reported by Defazio *et al.*[55], Lee and Bozeman[114]. Defazio attempted to summarise the reasons why collaborative activity would lead to higher publication productivity. Firstly, collaboration provides opportunity for knowledge combination, and knowledge combination tends to create new knowledge, which is reported in papers and gives higher research output. Secondly, collaboration provides learning opportunities for scientists. The skills acquired during collaboration can help scientists to potentially increase their future productivity. Finally, collaboration also provides social network and connections for scientists, which broaden the information channels for latest developments, opportunities and funding, leading to more research carried out and potentially higher output. These claims were mostly made against individual researchers. We will investigate whether these relationships can still be observed at the institution level in our data, and whether they vary across disciplines.

In the coming sections, we listed the top institutions by each of the collaborativity and productivity variables. These numbers are summarised over the entire period for which we have the data of. By analysing the institution's position movement across these lists, we were able to perform the similar analysis Price and Beaver had done so that we can compare with their results. The collaborative paper and collaborator analysis were performed next, to find out if institutions, like country collaborations, form small circles of frequent collaborators[46]. The correlation coefficients between each permutation of the collaborativity variables and productivity variables are then presented for each discipline.

Total Papers		Collaborative Papers	
Massachusetts Institute of Technology	4686	Massachusetts Institute of Technology	771
Carnegie Mellon University	4213	Carnegie Mellon University	760
Stanford University	3209	University of California Berkeley	523
University of California Berkeley	3008	Stanford University	586
University of Maryland	2106	Georgia Institute of Technology	385
Georgia Institute of Technology	2091	University of Washington	478
University of Washington	1984	University of Maryland	419
University of Michigan	1804	Purdue University	327
University of Texas Austin	1733	University of Texas Austin	346
University of Toronto	1713	University of Michigan	295
Size-Weighted Collaboration		**Percent Collaboration**	
Carnegie Mellon University	1974	University of Hong Kong	53%
Massachusetts Institute of Technology	1819	Villanova University	51%
Georgia Institute of Technology	1369	University of Connecticut	50%
University of California Berkeley	1356	Pace University	47%
Stanford University	1268	Fudan University	46%
University of Washington	1031	City University of Hong Kong	45%
University of Texas Austin	983	University of Strathclyde	45%
Purdue University	869	Bentley College	44%
University of Illinois Urbana Champaign	824	University of Electro-Communications	40%
University of Maryland	788	Grinnell College	40%

Table 5.2: Top Institutions in Computer Science (ACM) by paper count and collaboration variables. Institutions with lower than 100 papers in total is excluded in the percent collaboration table.

5.1.1 Computer Science

5.1.1.1 Top Institutions

Table 5.2 and Table 5.3 lists the top institutions for each variable measured based on ACM dataset and WoS dataset respectively. The data is gathered across the entire dataset available to us. So a smaller number of years are included for the WoS data than the ACM data (WoS covered 37 years while as ACM dataset covers 52 years, although ACM shows a smaller number in all straight count variables)

From the ACM dataset, the most prolific institution – MIT – is also the most collaborative; the next 8 out of 9 most prolific institutions have the next most collaborative papers, higher than what Price and Beaver observed at individual level. Comparing size-weighted collaboration ranking with collaborative paper ranking, MIT and Carnegie Mellon swapped places and Georgia Tech moved up two places. Institutions moving up in these ranks may mean that they participated in larger collaborations and (or) contributed more researchers in collaborations.

Total Papers		Collaborative Papers	
Massachusetts Institute of Technology	4620	Massachusetts Institute of Technology	2674
University of Illinois Urbana Champaign	4595	University of Illinois Urbana Champaign	2570
Carnegie Mellon University	4441	Stanford University	2309
Stanford University	4014	Carnegie Mellon University	2269
University of Maryland	3607	University of California Berkeley	1973
University of Texas Austin	3479	University of Maryland	1929
University of California Berkeley	3316	University of Texas Austin	1887
Technion Israel Institute of Technology	2961	Purdue University	1698
Purdue University	2942	Technion Israel Institute of Technology	1670
University of Southern California	2693	University of Waterloo	1481
Size-Weighted Collaboration		**Percent Collaboration**	
Massachusetts Institute of Technology	5574	Singapore Management University	94%
University of Illinois Urbana Champaign	5303	Providence University	87%
Carnegie Mellon University	4619	Wonkwang University	87%
Stanford University	4617	University of Crete	87%
University of Maryland	4114	Kookmin University	81%
University of California Berkeley	3850	East China Normal University	81%
University of Texas Austin	3290	University of Lausanne	79%
Purdue University	3202	University of Lyon	79%
Technion Israel Institute of Technology	2830	Universitat Pompeu Fabra	79%
University of Southern California	2754	Xiamen University	79%

Table 5.3: Top Institutions in WoS Computer Science by paper count and collaboration variables

The percent collaboration ranking has a completely different set of institutions. None of the institutions appeared in the first three rankings remain in this list. This is an indication that high paper output or high collaborative institutions do not have a high percent collaboration in Computer Science as seen by ACM.

The WoS dataset tells a very similar story, MIT is the most prolific and also the most collaborative, same as what was found in ACM dataset; the next 8 out of 9 prolific institutions are also the most collaborative institutions. The percent collaboration table also gave a different set of universities.

5.1.1.2 Comparison of ACM and WoS

Comparing the WoS with the ACM, six out of ten top institutions in the top total paper ranking are the same. The number of published papers is in the similar range across two datasets, especially for those top ranked institutions, *e.g.*, MIT, Carnegie Mellon and Stanford had almost the same number of papers presented in both, despite the journal coverages being quite different between the two databases. However, the gap in the paper number expands as the rank moves down. This could be due to the fact that those institutions focused on publishing in journals collected by a single database only, while more prolific institutions published in a wider range of journals collected by multiple databases.

Even though the institution's number of collaborative papers in the ACM dataset is only about a quarter of what found in WoS dataset, six of the top ten institutions were repeated in both datasets. The smaller collaborative paper counting may due to ACM's exclusion of the university and non-university collaboration papers. Both types of collaboration participants (Uni with Non-uni and Uni with Uni) were included in WoS dataset.

Size-weighted collaboration has the highest number of repeated institutions across the two datasets, eight out of ten are the same, despite that the raw value in the ACM dataset is only about a third of WoS.

The percent collaboration tables' institutions do not overlap at all across the two datasets. The percentage value, is quite different too. ACM is in the range of 40% -50 %, which is only half of the WoS's 80% - 90%.

The institution's ranking based on the paper count, collaborative paper count and size-weighted collaboration is quite stable, the two datasets have produced a very similar set of institutions for the top values of these metrics, even though the actual raw values varies largely across the two datasets. Percent collaboration produced different institution lists, where the two datasets do not have any overlap in this ranking at all. These institutions also did not appear in the previous three lists.

5.1.1.3 Institutional Collaborative Papers

Figure 5.1 shows the Computer Science collaborative paper count for each institution (left ACM dataset, right WoS dataset), ordered from high to low, and the number of collaborative paper axis (vertical axis) is in log scale.

The growing rate of the number of collaborative papers from low collaborative institutions to high ones shows a long tail-like distribution, where the highly collaborative

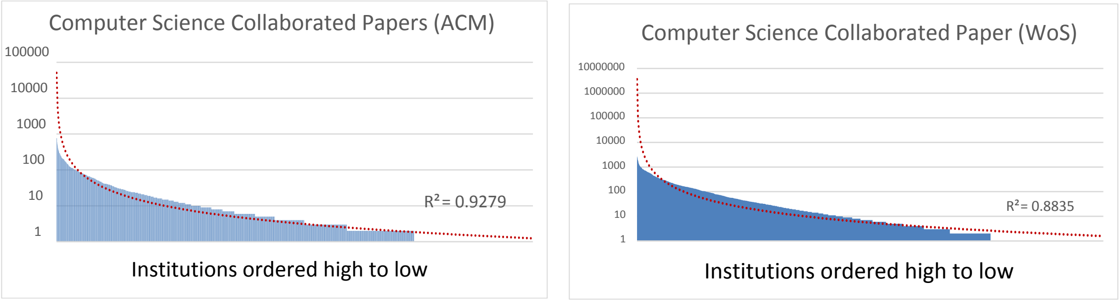


Figure 5.1: Institution's collaborative paper count in descending order in Computer Science. Most institution's name are omitted due to space limitation. The vertical axis indicates the institution's collaborative papers and is in log scale. The dotted line is a trend line based on power, which fits the institution's collaborative paper quite well, indicating a power decrease of institution's collaborative papers from high to low.

institutions are relatively few, but they have thousands of collaborations; while the institutions having about 10 collaborations account for almost half of all institutions. A power trend line (the dotted line) approximates a power reduction fits both of the diagram, with $r^2 = 0.93$ in ACM diagram and $r^2 = 0.88$ in WoS diagram.

On the other hand, deviation can be recognised in both ACM and WoS towards the top collaborative institutions (left side of the diagram), where the power distribution fitting out grows the actual number of collaborative papers published by the top collaborative institutions.

This power growth on the number of collaborative papers across institutions suggests that there are orders of magnitude as more papers are collaborated by the top collaborative end institutions than by lowly collaborative institutions. However, it is important to keep in mind that unlike publishing singly authored papers, the requirement of an institutional collaboration is to have more than one institutions working together. It is not possible for a single institution to have a lot of "collaborations" while all the rest have none. Collaborations must happen among institutions. So were all these papers the result of repeated collaborations between the top institutions themselves? This leads to one of the questions we would like to investigate: Do institutions form a core-periphery structure similar to the country level collaboration? We investigate this question in section 5.1.6.

5.1.1.4 Correlation of Collaborativity and Productivity in Computer Science

		PUBTOT PUBCOLL%	PUBTOT PUBCOLLw	PUBTOT PUBCOLL
CS (ACM)	rho	-.635	.858	.909
	sig.	.000	.000	.000
	n	1609	1609	1609
CS (WoS)	rho	-.326	.931	.962
	sig.	.000	.000	.000
	n	3742	3742	3742
Pharmacology	rho	-.496	.947	.976
	sig.	.000	.000	.000
	n	3251	3251	3251
Materials Science	rho	-.469	.958	.978
	sig.	.000	.000	.000
	n	3305	3305	3305
Psychology	rho	-.302	.872	.931
	sig.	.000	.000	.000
	n	2827	2827	2827
Law	rho	-.065	.820	.858
	sig.	.019	.000	.000
	n	1300	1300	1300

Table 5.4: Spearman's correlation between Productivity and Collaboration. Across disciplines, strong correlation between institutional total paper and collaborative paper were observed, demonstrating that both variables have direct impact on each other. The size-weighted collaboration, which taken into account of the collaboration size, shows a somewhat smaller correlation but still high and significant. The correlation between institutions' total paper and the percent collaboration have negative effect on each other, meaning that the more an institution published papers, the less proportion the collaborative papers they have produced. Law has the smallest, but significant negative correlation of all.

Table 5.4 and figure 5.2 top two figures show the correlation between the three collaboration variables and the productivity variable using the ACM and the WoS data. Significant positive correlation was found between the number of collaborative papers and the total number of papers. The effect of correlation is stronger in WoS dataset than in ACM, with 0.962 in WoS and 0.909 in ACM. When collaboration size – measured by size-weighted collaboration – is taken into account, both datasets showed a reduced correlation, but still high and significant.

The percent collaboration showed a significant negative correlation in both datasets. The ACM data have a higher negative correlation of -0.635, larger than the WoS's -0.326. A negative correlation implies that the more the institution published papers, the fewer the proportion of total papers directly the result of a collaboration. In other

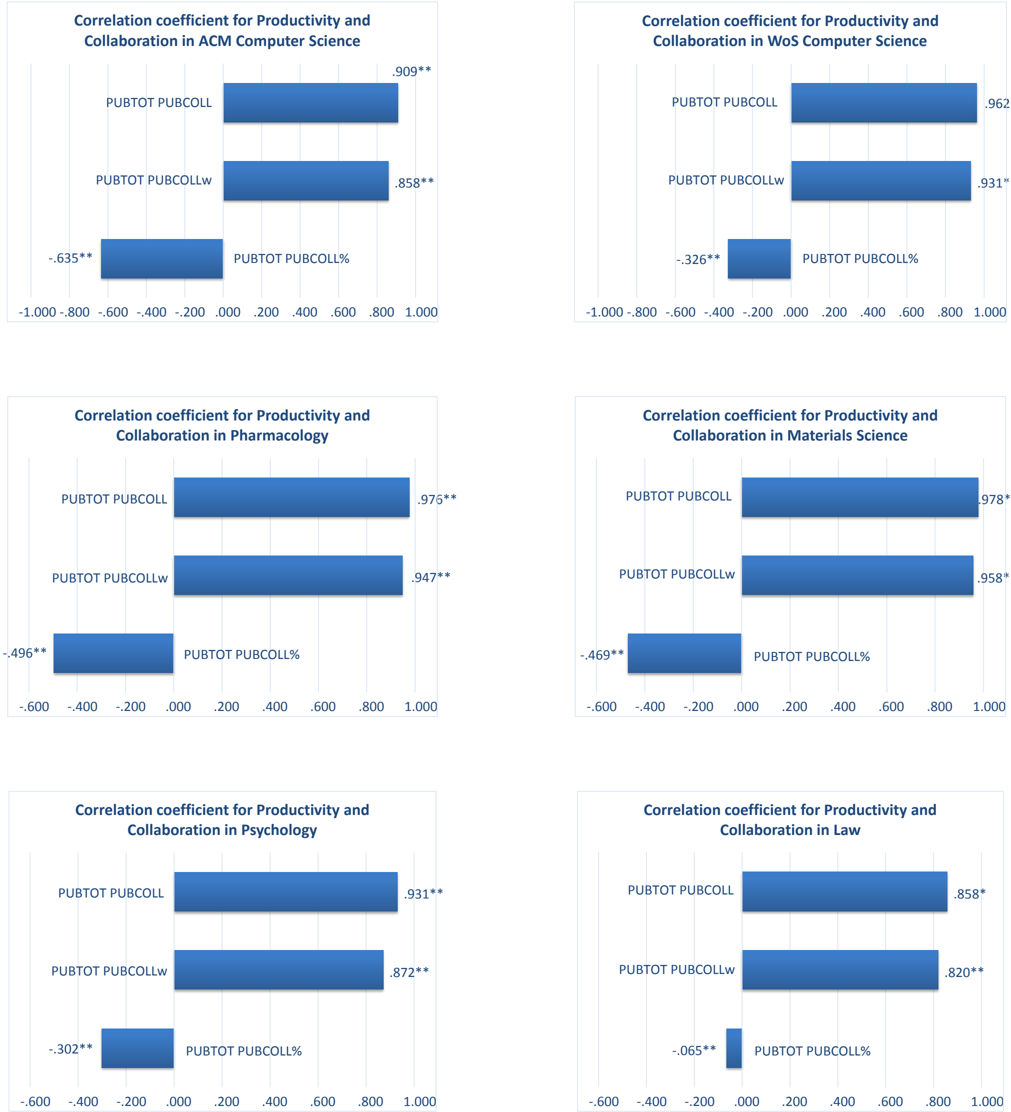


Figure 5.2: Correlation results of institutional productivity and collaboration (Visualisation of table 5.4)

words, in institutions that have published more papers, the higher percentage of the papers are singly authored without a collaborating institution. Despite the percent collaboration which is lower for high paper count institutions, their raw number of collaborative papers was in fact higher than those institutions with higher proportion of collaborative paper. The higher paper count is more important than the proportion when counting the collaborative papers.

Total Papers		Collaborative Papers	
University of Texas Austin	6320	Harvard University	3621
Harvard University	4824	University of Texas Austin	3038
University of North Carolina	4165	University of North Carolina	2584
University of California San Francisco	3993	University of Toronto	2502
University of Minnesota	3844	Karolinska Institute	2308
University of Michigan	3835	University of California San Francisco	2182
University of Toronto	3695	University of Milan	2101
University of Illinois Urbana Champaign	3612	University of Tokyo	2011
University of Milan	3431	University of Minnesota	1997
Karolinska Institute	3412	University of Michigan	1823

Size-Weighted Collaboration		Percent Collaboration	
Harvard University	11654	Dong-A University	96%
University of Texas Austin	8364	Hong Kong Polytechnic University	95%
University of Toronto	8200	National Yang Ming University	94%
University of North Carolina	7936	National Sun Yat-Sen University	94%
University of California San Francisco	6808	Catholic University of Daegu	92%
Karolinska Institute	6068	Campbell University	91%
Seoul National University	5898	Dongguk University	91%
University of Washington	5897	Yanbian University	91%
University of Minnesota	5806	Konkuk University	90%
University of Pittsburgh	5710	University of Toulouse	90%

Table 5.5: Top Institutions in Pharmacology by paper count and collaboration metrics

5.1.2 Pharmacology

5.1.2.1 Top institutions

Table 5.5 lists the top institutions in Pharmacology according to the ordering criteria specified. The top productive institution – Texas Austin which published almost 1500 more papers than the next most productive institution, has lose out on the collaborative papers by more than 600 to the second most collaborated institution. The next 8 out of 9 institutions remain on the next most collaborated institutions. Taking the collaboration size into account, University of Toronto and Karolinska Institute moved one rank higher; Seoul National University, University of Pittsburgh and University of Washington jumped into the top 10 list, while University of Milan, University of Tokyo and University of Michigan dropped out.

Same as all other disciplines investigated, the percent collaboration list is vastly different from the other three lists. There are 6 out of 10 institutions which published only just over 100 publications for the duration of 40 years. The paper output for these high percent collaboration institutions is low.

5.1.2.2 Correlation of Collaborativity and Productivity in Pharmacology

Figure 5.2 middle left shows the correlation coefficient between the productivity of an institution and its collaboration metrics in Pharmacology. The institutions' total paper and collaborative paper show significant and strong positive correlation. This suggests that the more papers published by an institution, the more papers it has co-authored with other institutions. Size-weighted collaboration slightly reduces the correlation coefficient from .976 to .947, but it still represents a very high correlation between the two variables. Significant and negative correlation is also found between papers and the percent collaboration in Pharmacology, thus agreeing with the results in the Computer Science domain.

5.1.3 Materials Science

5.1.3.1 Top institutions

Total Papers		Collaborative Papers	
Indian Institute of Technology Bombay	6171	Tohoku University	2553
Tohoku University	4256	Indian Institute of Technology Bombay	2301
Kyoto University	3767	University of Tokyo	2290
University of Tokyo	3668	Kyoto University	2102
Pennsylvania State University	3557	Tokyo Institute of Technology	1674
Tokyo Institute of Technology	3160	Seoul National University	1663
Nanyang Technological University	2875	Osaka University	1628
Harbin Institute of Technology	2816	Korea AIST	1588
Tsinghua University	2798	Pennsylvania State University	1513
Osaka University	2771	University of Cambridge	1511
Size-Weighted Collaboration		**Percent Collaboration**	
Tohoku University	4945	Dongguk University	93%
University of Tokyo	4765	Paris Diderot University	92%
Kyoto University	4111	University of Havana	92%
University of California Berkeley	3640	University of Potsdam	91%
Massachusetts Institute of Technology	3579	Andong National University	90%
Seoul National University	3510	Goteborg University	89%
Indian Institute of Technology Bombay	3379	Autonomous University of Barcelona	89%
University of Cambridge	3060	Kwangwoon University	88%
Tokyo Institute of Technology	3032	Changwon National University	88%
Osaka University	3031	Kongju National University	88%

Table 5.6: Top institutions in Materials Science based on total papers and collaboration metrics.

Figure 5.6 lists the top institutions in Materials Science according to the four metrics. This time, the top productive institution – Indian Institute of Technology Bombay is no longer on the top of the list for collaborative papers, even though it has published nearly 2000 papers more than the next prolific institution, It falls short by 252 papers to the most collaborative institution.

7 top productive institutions remain on the list for the top 10 collaborative institutions, this is the lowest number of all the disciplines investigated. Comparing the institution's position in the size-weighted collaboration with collaborative paper lists, Indian Institute of Technology Bombay dropped from 2nd to 7th in collaborative papers; MIT and UCB, both of which were not represented in the first two lists, appeared in the size-weighted collaboration list. Cambridge moved from 10th in the collaborative paper list up 2 places into the 8th in the size-weighted list. The same observation was found for the percent collaboration list as the other disciplines – no institution overlap with the other three lists.

5.1.3.2 Correlation of Collaborativity and Productivity in Materials Science

Figure 5.2 middle right, a similar correlation pattern was found in Materials Science with the other disciplines – two positive and one negative. The institutions' total paper and collaborative paper showed a significant and strong positive correlation of $r = 0.978$. The size-weighted collaboration was $r = 0.958$, slightly reduced. Significant and negative correlation was found between the papers and the percent collaborations in agreeing with the two previous disciplines.

5.1.4 Psychology

5.1.4.1 Top institutions

9 out of 10 of the top prolific institutions remain in the top ten on the collaboration list; Stanford University moved up three places in collaboration list from 9th to 6th; University of Missouri Columbia dropped out while University Minnesota entered the list. There were no big position changes for the rest of the institutions. In size-weighted collaboration list, Yale moved above Stanford from 7th to 5th, possibly indicating that

Total Papers		Collaborative Papers	
University of Illinois Urbana Champaign	2652	Harvard University	1648
Harvard University	2421	University of Illinois Urbana Champaign	1413
University of California Los Angeles	2397	University of California Los Angeles	1378
University of Texas Austin	2353	University of Michigan	1246
University of Michigan	2122	University of Texas Austin	1220
University of North Carolina	2047	Stanford University	1185
University of Wisconsin Madison	1959	University of North Carolina	1114
University of Missouri Columbia	1946	Yale University	1049
Stanford University	1887	University of Wisconsin Madison	1005
Yale University	1791	University of Minnesota	993
Size-Weighted Collaboration		Percent Collaboration	
Harvard University	4056	Karolinska Institute	81%
University of California Los Angeles	3123	Skidmore College	80%
University of Illinois Urbana Champaign	2686	Medical College of Georgia	80%
University of Michigan	2478	University of Saint Andrews	78%
Yale University	2383	Brown University	78%
University of Texas Austin	2372	Baylor College of Medicine	78%
Stanford University	2356	Nanyang Technological University	77%
University of North Carolina	2279	University College London	75%
Boston University	2209	University of Konstanz	75%
University of Pittsburgh	2180	Colgate University	75%

Table 5.7: Top Institutions in Psychology by paper count and collaboration metrics.

Yale involved in larger collaborations than Stanford. The same observation was found with percent collaboration rank as the other disciplines.

5.1.4.2 Correlation of Collaborativity and Productivity in Psychology

A similar correlation pattern was also found in Psychology. The institutions' total paper and collaborative paper showed a significant and strong positive correlation at $r = 0.931$. The size-weighted collaboration was slightly smaller of $r = 0.872$. Both values are significant at 99%.

The correlation in percent collaboration was $r = -0.302$, it is the second smallest coefficient, just above Law.

Total Papers		Collaborative Papers	
Harvard University	2574	Harvard University	700
University of Chicago	2093	University of Chicago	448
University of California Berkeley	1767	New York University	428
Yale University	1629	Yale University	419
Georgetown University	1583	Georgetown University	352
New York University	1563	University of California Berkeley	323
University of Pennsylvania	1366	Northwestern University	308
Columbia University New York	1220	University of Pennsylvania	291
University of Virginia	1168	Stanford University	259
Northwestern University	1116	Columbia University New York	240
Size-Weighted Collaboration		**Percent Collaboration**	
Harvard University	1217	Johns Hopkins University	63%
University of Chicago	624	University of Oxford	43%
New York University	658	University of Missouri Columbia	30%
Yale University	683	University of Maryland	28%
Georgetown University	678	Northwestern University	28%
University of California Berkeley	577	New York University	27%
Northwestern University	458	Harvard University	27%
University of Pennsylvania	482	University of North Carolina	26%
Stanford University	472	Yale University	26%
Columbia University New York	417	Stanford University	24%

Table 5.8: Top Institutions in Law by paper count and collaboration metrics.

5.1.5 Law

5.1.5.1 Top institutions

Harvard University was found to lead in the Law discipline. It published nearly 500 more papers than the second most prolific institution; also 250 more co-authored papers than the second most collaborative institution; it almost doubled the size-weighted collaboration of the second highest on the list.

9 out of 10 of the most prolific institutions remain in the top 10 most collaborative list. As for the size-weighted collaboration list, there are no ordering changes at all when compared to the collaboration list.

What differentiates Law from the other disciplines is the percent collaboration list. 5 out of 10 of the top percent collaboration institutions re-appeared in the previous three ranks, which none of the other disciplines did; the top productive institution – Harvard –

ranked 6th in the percent collaboration. This observation suggests an inverse relationship between percent collaboration and productivity measurement.

In addition, Law is not a discipline that has a high percent collaboration. Even within the top percent collaboration list, the ratio rapidly dropped from 63% to 25% towards the 10th rank, compared with other discipline's 90% to 75% rate.

5.1.5.2 Correlation of Collaborativity and Productivity in Law

Figure 5.2 bottom right. Law demonstrated the lowest correlation in all of the examined subjects, with $r = 0.858$ between paper count and collaborative papers and $r = 0.820$ between paper count and size-weighted collaboration. Even though a negative correlation was found between paper count and percent collaboration as all other disciplines, its coefficient size was the smallest at $r = -0.065$ (only significant at 5%).

5.1.6 Institution's Collaborator and Collaborative Papers

In section 5.1.1, we found that the top collaborative institutions publish a lot more papers than the next most collaborative institution, and the next most collaborative institution collaborate yet a lot more than the next again, and so on. The reduction speed of the published collaborative papers is power law like from the top institutions to the lowest ones. Since collaboration is a mutual process and at least two institutions must be involved in order to collaborate, such distribution of the collaborative papers may only be a result of two processes: 1. the highly collaborative institutions formed a 'collaboration club', where institutions collaborate frequently within the club, pushing up their collaborative papers while leaving out the non-club institutions for collaboration; or 2. the higher collaborative institutions worked with a wider range of institutions, thus the more collaborative the institution is, the more the institutions it has collaborated with.

In fact, the first process was confirmed by Choi [46] at the country level. He found that international collaboration forms a core-periphery structure, where advanced nations form a circle of frequent collaborations, while the rest of the countries are like the periphery structure attached outside of this core, with little collaboration compared to the core.

To investigate which process happened at the institution level, we calculate the ratio between collaborative papers and unique collaborators for each institution within a discipline. This ratio tells us how many papers on average institutions collaborate with another institution. If this ratio is about the same across institutions, then it means institutions work with more collaborators when they publish more collaborative paper and they do not form a collaboration club.

Figure 5.3 plots the institutions' collaborative papers against its unique collaborators for all disciplines. Each dot represents an institution.

As the figure suggest, institutions that have more collaborative papers also have more collaborators, and it is true in all of the disciplines. The relationship between the two variables can be fitted with a linear fitting. The ratio between the collaborator and collaborative papers is in the range of 1.4 to 1.8 across disciplines. Some outliers do exist towards the high collaborative end, for example, in ACM Computer Science(Figure 5.3 top right), the institutions represented by the two right most dots are MIT (right most) and Canegie Mellon University. Their papers per collaborator is approaching 2.8, slightly higher than the average 1.8, which deviated from the dotted line. Even with these outliers taken into account, this ratio does not vary largely. If the top institutions formed the 'collaboration club', we would observe top collaborative institutions having much fewer collaborators, (so their vertical position on the diagram would be much lower). However, these diagrams do not rule out the possibility that the top institutions collaborating with each other the most, while only collaborating once or twice with the other institutions, resulting in a high number of collaborative papers as well as collaborators.

5.1.7 Discussion and Summary of Collaboration vs Productivity

In all of the five disciplines, a close connection between paper count and collaborativity among the top institutions was clearly observed: a high number of prolific institutions

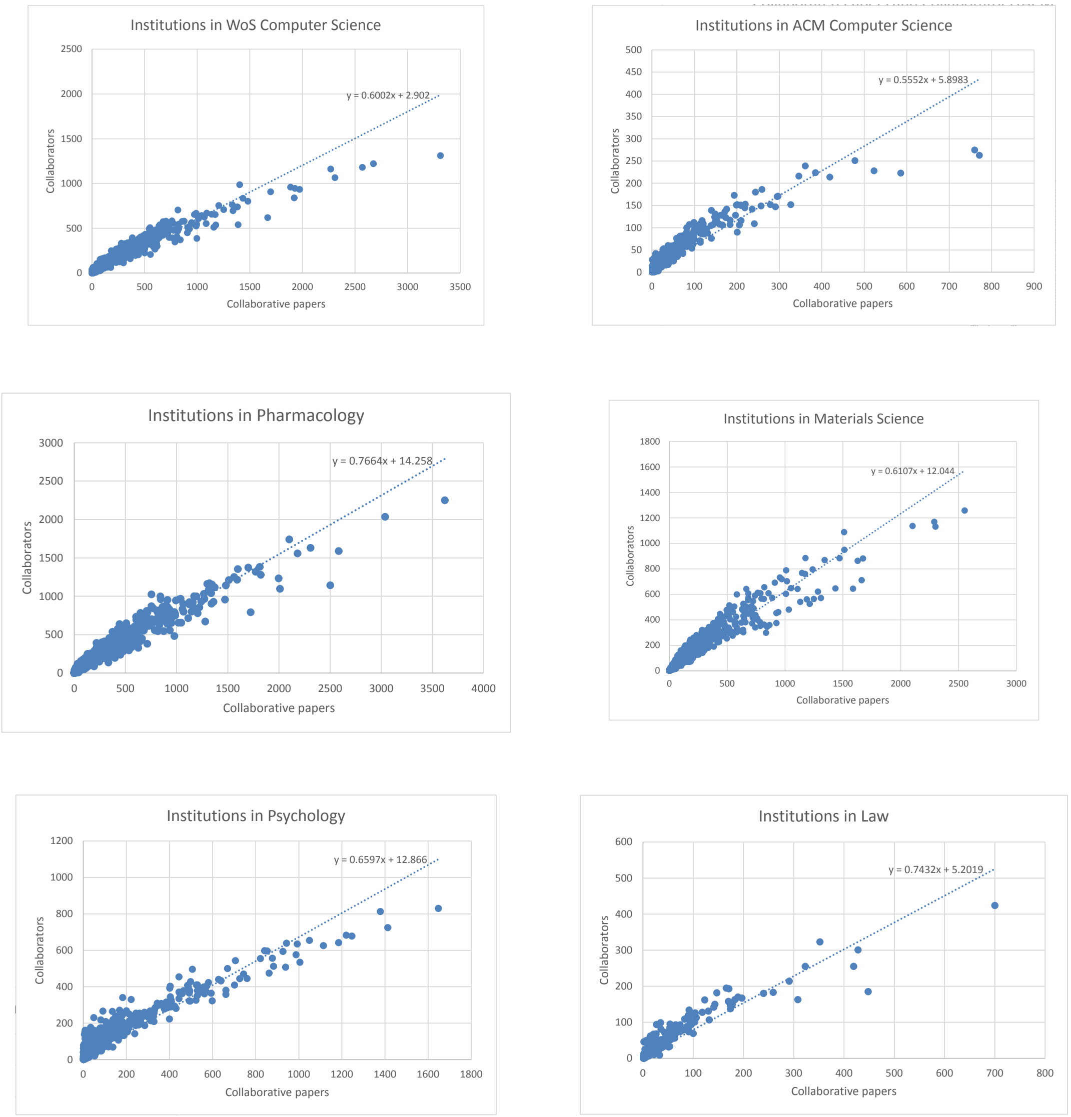


Figure 5.3: Institution's collaborative paper and collaborators. The points on the figure indicate institutions. Horizontal axis is the total number of collaborative papers the institution has published, vertical axis is the unique collaborators the institution has worked with for producing these collaborative papers. The relationship between the number of collaborators and collaborative papers is approximately linear. Some institutions tail off at the top end, where there was a smaller number of collaborators for those high collaborative institutions, such as the right top two institutions in ACM Computer Science diagram. Despite these outliers, the ratio between collaborative paper and collaborator is very much stable around 1.4 to 1.8 papers per collaborator. This means that a highly collaborative institution also have a larger number of unique institutions it has collaborated with. The large number of collaborative papers observed in the top institutions are not a result of these institutions collaborated frequently themselves, instead, it is more likely that they have collaborated with a wider range of collaborators

were found in the top collaborative paper and top size-weighted collaboration tables.

The results from the two Computer Science datasets – ACM and WoS – agrees with each other in the top institution analysis: 6 out of 10 institutions are the same in the top paper rank; 7 out of 10 are the same in the collaborative paper rank, and 8 out of 10 are the same in the size-weighted rank respectively. There are also differences between the two datasets. The number of collaborative papers in ACM is four times fewer than the WoS dataset, while the total number of papers were the same. The institution's percent collaboration (collaborative paper / total paper) is lower in ACM dataset than in WoS, which was due to the lower number of collaborative papers in ACM dataset.

The increase of collaborative papers from low to high institutions follows a near power growth rate for all the disciplines examined. The Law and Psychology has the highest rate of increase, while Computer Science in WoS, Pharmacology and Materials Science have a slightly slower rate. A power law growth means that either the high collaborativity institutions collaborate frequently with each other, or they have a large number of collaborators. Subsequent study showed that the ratio between the collaborative paper and collaborator remains a small number – with the largest ratio at around 6 papers per collaborators. This shows that institutions with higher number of collaborative papers have proportionally higher numbers of collaborators. That is, highly collaborative institutions do not form a circle of collaborators, so the lower collaborative institutions have opportunities to collaborate with them.

The correlation analysis between the number of total papers published and the number of collaborative papers gives positive results for all of the subjects. Computer science, Pharmacology and Materials Science, which belong to the NSE, have consistently higher correlation coefficients compared to Psychology and Law. The correlation coefficient dropped to and below 0.9 in Psychology and Law, while all the rest are in the high 0.9 range. This suggested a stronger relationship in NSE disciplines between the number of published papers and the collaborated papers than the SSH discipline. These differences could due to the collaboration patterns and publication practices between these disciplines.

A high positive correlation was found between the papers published and size-weighted collaborations. This correlation coefficient was reduced in size compared to the unweighted collaborative paper count (***PUBCOLL***). Size of the collaboration does not appear to give positive impact on the productivity.

The correlation between the productivity and percent collaboration is consistently negative in all subjects examined. A negative correlation indicates that the more papers an institution publishes, the fewer the collaboration. Despite that, the highly productive institutions have lower percent collaboration, and they publish more collaborative papers than the lower productive institutions. This is because they have a proportionately larger publication output.

Of all the subjects studied, Law has the weakest negative correlation between the productivity and percent collaboration. This suggests that Law as a discipline may have a completely different publication practice and collaboration pattern.

In the disciplines examined, there are significant portions of the institutions which publish almost only collaborative papers, with a very small number of singly authored papers. This is suggesting that some institutions have strong barriers to publish papers alone. A future study can probably verify this from the relationship between rejection rate and the productivity of the institution: A higher rejection rate of the lower productive or lower impact institutions for singly authored papers should be observed.

5.2 Productivity versus Impact

Lanjouw and Schankerman [108] studied the relationships between productivity and impact on company patents. They used the ratio between number of patents generated and resources spent on them as the measurement of productivity; and the revenues generated from these patents as the measurement of impact. They reported a negative correlation between their productivity and impact measurement of the companies. They also showed that the amount of resource spent on the patent is a good predictor of the patent impact. We explore whether the same is true in the process of institutions publishing papers. In addition to the impact of the output items, we are also interested to learn whether the impact of entities, *i.e.* institutions, is correlated with their productivity.

The productivity variable is the total paper count and the impact variables are as follows:

- CITTOT – Total citation count of institutions' research in the given field. This measures the impact of the institution's research in the domain. This variable is not normalised.
- CITTOTw – Citation PageRank – fine tuned citation measurement, taking account of the citation impact of the citing institution. A qualitative improvement of the citation count metrics.
- CITAV – Citation Per Paper – the number of citations each individual paper of the institutions receive on average. This measures the mean impact of a given institution's research output.
- WRANK – Web based view of the importance of an institution. This gives an overall position of the institution and represents the institution's overall quality.

Applying Spearman's non-parametric correlation algorithm, we obtained the correlation coefficient between the institutions' paper and the four impact metrics.

5.2.1 Institution Citation

In Figure 5.4, the pairwise correlation coefficient with the unfiltered citations variables (CITTOT) were highly positive (around $r = 0.9$) in all investigated disciplines across two datasets in both NSE and SSH. The weighted citation variable (CITTOTw) showed a slightly smaller correlation compared to the un-weighted citations. Despite this, it demonstrated that the two variations of the impact measurements are robust when applied at institution level.

The lowest correlation coefficients were found in the SSE disciplines, with Law having the lowest of all. SSE disciplines are showing some difference compared to NSE disciplines in the institutional impact measurements.

		PUBTOT CITTOT	PUBTOT CITTOTw	PUBTOT CITAV	PUBTOT WRANK
CS (ACM)	rho	.899	.863	.550	.728
	sig.	.000	.000	.000	.000
	n	1609	1609	1609	1609
CS (WoS)	rho	.905	.885	.501	.490
	sig.	.000	.000	.000	.000
	n	3742	3588	3742	3742
Pharmacology	rho	.916	.905	.431	.393
	sig.	.000	.000	.000	.000
	n	3251	3251	3251	3251
Materials Science	rho	.923	.905	.411	.384
	sig.	.000	.000	.000	.000
	n	3305	3305	3305	3305
Psychology	rho	.879	.856	.537	.608
	sig.	.000	.000	.000	.000
	n	2827	2827	2827	2827
Law	rho	.852	.815	.441	.571
	sig.	.000	.000	.000	.000
	n	1300	1300	1300	1300

Table 5.9: Spearman's correlation results (rho) of the productivity vs impact. Five disciplines from two datasets showed very similar overall correlations: correlations with the CITTOT(weighted and un-weighted) have similar values, while higher than the CITAV and WRANK. All four pairs of productivity and impact variables showed positive and significant correlations. These correlation results means that the more the institutions published papers, the more the citations they receive, and the higher they have been ranked in the world ranking, as well as the higher impact their publications are. sig. is the significance for a 2 tailed test; n is the number of institutions.

The high correlation in both weighted and un-weighted citation metrics means that the more the institutions have published, the higher their citation impact. However, we must view this result with caution as these are unfiltered variables, the high correlation maybe contributed from a common variable (such as collaboration) that presented in both variables.

5.2.2 Paper Impact

The average citation (CITAV, Figure 5.4 third bar from top in all sub-figures) showed significant and positive correlations with the number of papers institutions have published. This correlation was found in all six figures of the studied disciplines, both NSE and SSH. Within the WoS dataset (*i.e.* all figures except first one), Psychology had the highest correlation of $r = 0.537$, Materials Science had the lowest of $r = 0.411$. It appears that SSH discipline has a higher correlation coefficient than NSE disciplines.

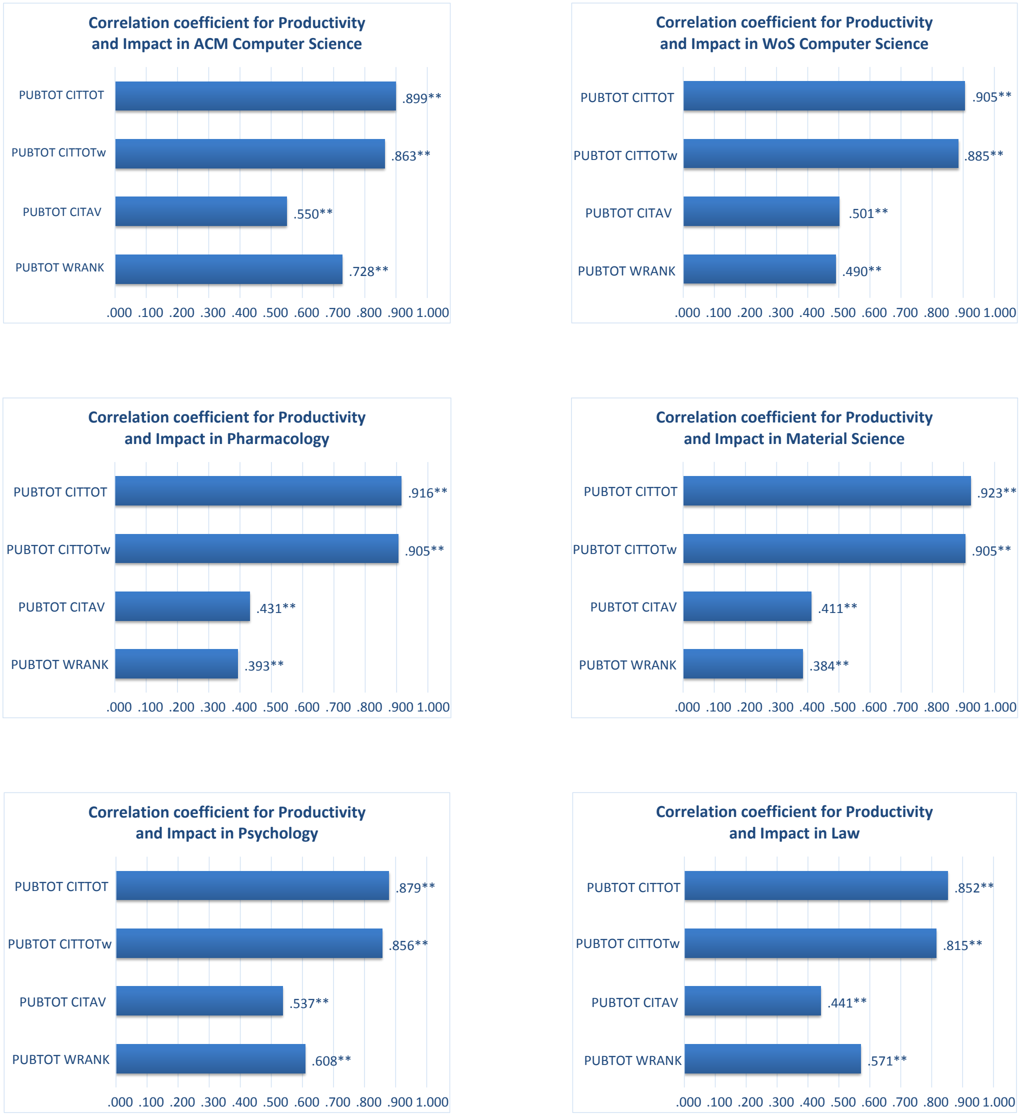


Figure 5.4: Correlation results of institutional productivity and impact (Visualisation of table 5.9)

A positive correlation indicates an opposite result as found by Lanjouw and Schankerman[108]. This gives evidence that highly productive institutions have intrinsic factors that attract citations more than the low productive institutions. Factors such as world leading researchers the top institutions employ and the reputations they have can affect the visibilities of these papers, which can then have a knock on effect to the received citations.

To understand how the institution level productivity and average paper impact unfold onto the individual researchers; what the relationship between the number of papers researchers publish and the impact of these published paper is, we analyse the same data and conduct correlation at the individual level.

5.2.2.1 Researcher Productivity and Paper Impact

In all of our studied subjects, we found that institution's total paper output is positively correlated with the average citations these papers received. In other words, the more the papers an institution had published, the more the citations each of those papers received. This gives the opposite result as reported by Lanjouw and Schankerman [108] in their company's patent productivity study. They found negative correlations between the productivity and impact of the patent at the company level. That is, in companies which have higher numbers of patent filed, their patent impact tend to be lower. According to their research, it is less useful for companies to file more patents because the impact of the patents would just get lower, which would result in less revenues than fewer but higher impact patents.

A further correlation analysis at the researcher level is conducted. We would like to investigate for researchers whether there is a correlation between the number of papers they publish and the average citations they receive on their papers.

The authors are first aggregated into groups according to their paper count. Then the average citations received by the authors who published the same number of papers is calculated. Finally, the average citations per paper is calculated by dividing the number of papers. Formula 5.1 describes this calculation.

$$\bar{c}_p = \frac{\frac{\sum_a c}{a}}{p} \tag{5.1}$$

$\bar{c}_p$ is the average citations per paper received by authors publishing p papers; $\sum_a c$ sums over the citations for all authors who published p papers, a is the number of authors.

For example, if there are in total 5 authors who published 20 papers in a Computer Science dataset, each received 40, 45, 30, 69 and 150 citations respectively. To calculate the average citation per paper for authors published 20 papers is:

$$\bar{c}_{20} = \frac{\frac{40+45+30+69+150}{5}}{20} \tag{5.2}$$

$$\bar{c}_{20} = 3.34\ (\textit{citations per paper}) \tag{5.3}$$

In this example, the average citations received by each paper is 3.34 for authors who published 20 papers.

This calculation is repeated for each author group. The resultant data looks like table 5.10 using the Materials Science data.

Papers published	Avg. citations per paper
1	6.78
2	7.55
3	8.19
4	8.76
...	...
970	8.41
1129	8.36
1154	9.96
1162	8.24

Table 5.10: Sample data for productivity vs impact at the individual level

The correlation coefficient is then calculated between these two variables. This process is then repeated for each discipline to obtain the disciplinary correlation.

Figure 5.5 shows the correlation results for each discipline at the individual researcher level. The correlation result is mixed:

WoS Computer Science, Materials Science and Pharmacology show significant and negative correlation; Psychology and ACM Computer Science show significant and positive correlation; Law is found to have no correlation between paper productivity and paper impact at researcher level.

It was surprising to find that ACM and WoS dataset demonstrates such big difference – one having the highest positive correlation while the other having the biggest negative. However, due to the factors such as author identification, which is an unsolved problem of its own, this difference cannot be interpreted any further. Putting the ACM dataset aside, from the WoS dataset, the NSE gave a negative correlation and Law gave no correlation at the individual level. Because the institution level correlation is formed by

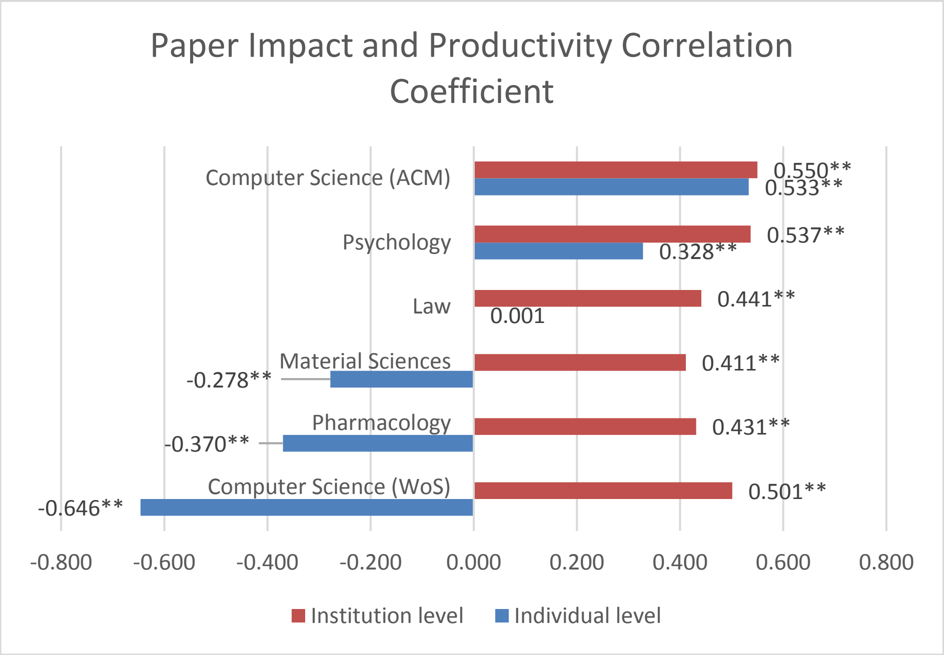


Figure 5.5: Paper impact and productivity non-parametric correlation coefficient for institution level and individual level. Institution level correlates the institutions' total paper (PUBTOT) with the mean citation per paper(CITAV), while individual level aggregates the same productivity authors and uses the mean citation per paper for the authors publishing same number of papers. The ACM and WoS dataset seems to have dramatic differences as shown in Computer Science result. In WoS dataset, NSE shows significant negative correlation; Psychology shows significant positive while Law does not show any correlation for individual researchers.

these individual authors, the individual level correlation must match up with the institution level. This suggests that certain types of authors must show negative correlation while some show positive correlations between their published papers and the citations.

To explore this further, the individual researchers were put into three equal groups according to their paper count rank. We would like to see whether productivity is related to the each author's citation. Due to the number of authors sharing the same low number paper counts, (*e.g.* more than half of the authors published 1 paper in Pharmacology dataset), splitting all authors into three groups would give unhelpful result (because the lowest group will only contain authors published 1 paper). A threshold of minimum 100 papers was applied to remove the vast majority of the very low productive authors. The correlation is calculated for researchers in the high and low groups. Figure 5.6 presents the results.

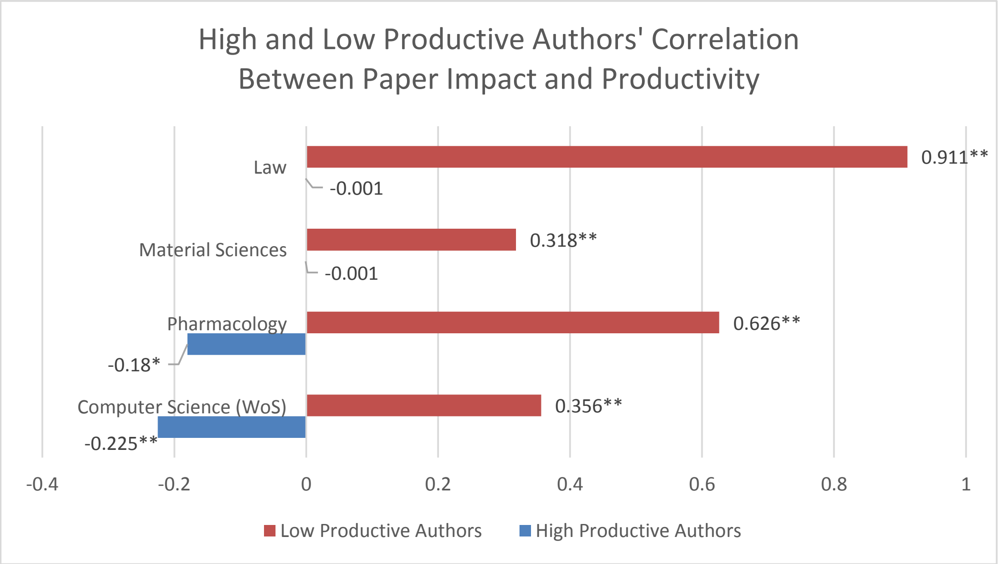


Figure 5.6: All the disciplines show significant and positive correlation among the low productive authors, meaning that among the low output researcher, a few more paper brings up citations received per paper. At the high productivity end, the picture inverses in Pharmacology and Computer Science: the more paper researchers publishes, the less citation per paper they are expected to receive.

Firstly, all of them have shown significant positive correlation among low productive authors, which suggests that for low productive authors, the more the papers they published, the more the citations they receive for each of their papers. In Law, the correlation coefficient even went past $r = 0.9$ in the low productive end, the highest of all disciplines. However, no significant correlation was found among the highly productive authors. In Pharmacology and Computer Science (WoS), negative correlation was found among the high productive authors, which means highly productive authors would receive fewer citations per paper when they publish extra papers.

5.2.2.2 Discussion

A positive correlation between the paper number and citation per paper means that a few more published paper can give the authors an accelerated rate of receiving citations, making the ratio between citation and papers larger, hence giving a positive correlation.

According to this correlation, for example, say an author published 1 paper per year received 2 citations to his paper; another author made more effort and published 2 papers in the same period received 3 citations to each of his papers (the increase is perhaps due to the publicity of this extra paper). This would result in totally 6 citations to his work.

As we can see, publishing the extra paper accelerated this person's citation accumulation. The increase of the citation is *faster* than the number of papers, hence the observed positive correlation.

However, this accelerated citation rate slows down as the author becomes more productive. We even found negative correlation in the top-tier productive author group. When the number of papers published per year reach a certain threshold (in our case, when researcher's productivity reaches the top-tier as we defined it), the increase in citation slowed down so much that the rate of the papers published over takes it. Each additional paper published above this threshold receives lower-than-average citations, hence it bring down the researcher's average citation, which is the negative correlation we have observed.

Using this result, we can also infer the institutions composition of high-low productivity researchers.(Institution in this context is just a particular combination of researchers that satisfy the overall correlation we have observed). We have found positive correlations between institutional productivity and paper impact, that is, the more papers institutions publish, the higher the average citations the papers receive. We also found that only the low productive authors show such positive correlation, while the high productive authors show a negative correlation. In order for the positive correlation to form at the institution level, the most of authors should be low productive, with some medium productive ones, but it can only have a few high productive authors, so that the overall correlation would remain positive.

Despite our large sample size, we still need to pay attention to the limitation of this analysis. The following factors can potentially affect the institution's productivity, hence altering the assumptions made regarding the institution's productivity.

- An institution's researcher number. An institution can have a lower publication output simply due to lower numbers of researchers.
- Productivity varying between individual researchers. For example, publication rate, publication impact, working hours, the ability of the researcher, the resources available to the researcher and so on.
- An institution's research resource. i.e. funding, equipment etc.
- The effort spend on individual publications. For example, more time spent working on one paper leading to a lower overall paper publication.

5.2.3 Institution Rank

The WRANK used in this analysis is a mixture of four factors, one of which is the institution's paper count. This paper count measures the same underlying factor as our PUBTOT variable. The paper count measure in WRANK contributes to 15% of the overall ranking score, which gives about $r = 0.4$ correlation alone, assuming the paper count in the WRANK highly correlates with PUBTOT used in the current study.

The correlation between the institutions' WRANK and their total papers gives a significant coefficient. In Pharmacology and Materials Science, the correlation fall below $r = 0.4$, which is the minimum amount given the overlapping paper count variations in both variables. The actual meaningful variation of the WRANK as predicted by the Pharmacology and Materials Science papers is therefore very limited. In the same WoS dataset, other disciplines did not give exceptionally high correlations. In the ACM Computer Science the correlation is the highest of $r = 0.728$.

The correlation between the WRANK and disciplinary productivity is not particularly high. The WRANK contains the institutional impact data in all disciplines offered by the institutions, while the data the WRANK is correlating with is in single discipline. Using one of a few dozen disciplines an institution offers to estimate the entire institutional output is not appropriate.

5.3 Collaboration versus Impact

In the literature, the exact aspect of collaboration or impact varies from one another. These variations change the interpretation of the result. Presser [156] used co-authorship as proxy to collaboration, and put collaboration into two categories: non-collaborative as defined by single author papers and collaborative paper as defined by 2+ authors. The quality of the papers is estimated by the editorial decision of a journal's review panel. Narin *et al.*[137] compared the number of citations received across *three types* of collaboration based on the paper's address line – European country with European country's collaboration, European country with non-European country's collaboration and single country's paper. He *et al.*[93] studied collaboration's effect on authors, so his collaboration is measured by the number of collaborators and the affiliation of their collaborator. i.e. within-university collaborator, within-country collaborator or international collaborator. The paper impact was measured by the paper's n-year average impact factor.

Even though the exact aspect of collaboration and impact is different, the outcomes indicate that collaboration and impact are positively related. Presser [156] claimed that collaborative papers are "less bad" than the single authored papers. Narin *et al.*[137] found that multiple-country collaborative papers are twice as heavily cited as single country papers, while He *et al.* [93] showed a positive correlation between an article's impact and its within-university collaborator or international collaborator.

This section examines the relationship between the institutional collaboration and impact for the five disciplines, and presents the result of the analysis.

Our collaboration metrics are:

- Institution Collaborative Paper – Strength of the collaboration
- size-weighted collaboration – Strength and size of the collaboration
- Percent Collaboration – Institution's percentage of collaborative paper. A measure of institution's focus on collaboration.

Our impact metrics are:

- Webometrics Rank – Ranking of world institutions, according to various metrics.
- Institution Citation Count – Institution's overall citation impact.
- Citation PageRank – Institution's citation impact based on the network work effect.
- Citation Per Paper – Institution's average paper impact.

Each pair of the collaboration and the impact metrics for the five disciplines from two datasets are correlated. Figure 5.7 shows the correlations between the impact and the collaboration metrics. The result of PUBCOLLw is very close to PUBCOLL, therefore PUBCOLLw is omitted from the figure. Similarly, CITTOTw and CITTOT are also very close, therefore CITTOTw is also omitted to improve clarity of the figure.

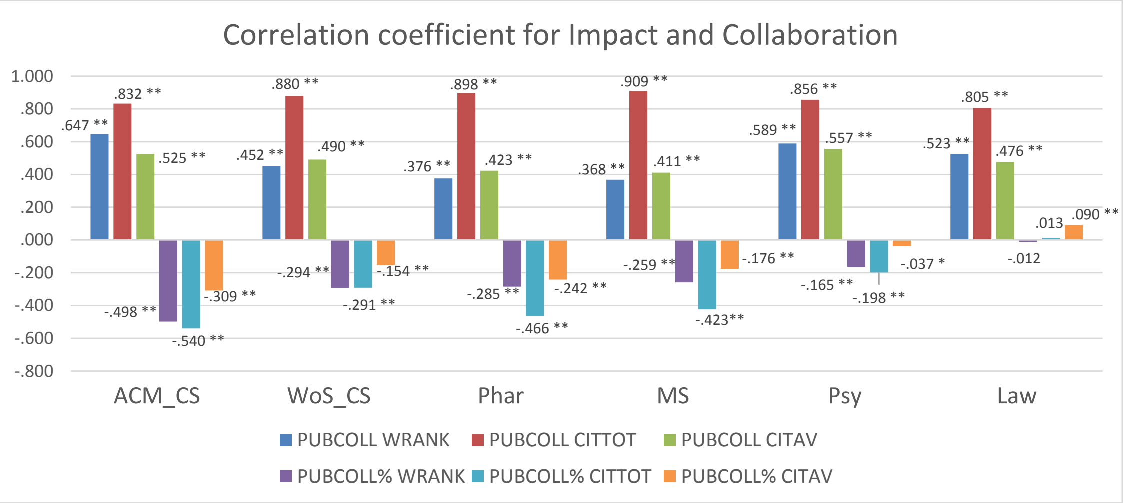


Figure 5.7: Correlation results of institutional collaboration and impact (Visualisation of table 5.11). PUBCOLLw and CITTOTw are omitted from the figure to improve clarity.

5.3.1 Ranking and number of collaborated papers

The ranking of institutions showed generally significant high to medium correlations with the number of institutional collaborative papers. This means that the higher the institution's rank, the more the collaborative papers the institutions have published.

Institutions that have been ranked high in the Webometrics ranking are also the ones that published more collaborative papers in our datasets. The highest correlation was observed in ACM Computer Science with $r = 0.647$, while the lowest is in Materials Science with $r = 0.368$.

The Webometrics ranking is a multiple factor variable, with productivity, size, visibility and rich file all contributing to the final score of the ranking. The positive correlation found here could be the result of one or several sub-variables this ranking algorithm considers.

5.3.2 Ranking and percent collaboration

The ranking of institutions show significant negative correlation with the collaboration percentage (purple bar) for all disciplines (except Law). The negative correlation dropped below $r = -0.4$ in the Computer Science (ACM) dataset, which has the largest negative correlation of all. The NSE disciplines, Computer Science (WoS), Pharmacology and Materials Science have a very close coefficient of about $r = -0.3$. As we move into SSH disciplines, the negative correlation diminishes. Psychology has $r = -0.165$ and no correlation was found in Law.

An interpretation of the negative correlation between these two variables is: the better ranked institutions are publishing a smaller proportion of collaborative papers than lower ranked institutions. From a different perspective, the lower the institution's rank, the greater proportion of its paper are collaborated, and the less proportion of its papers are singly authored.

Although the proportion of the collaborative paper reduces as institutions move higher in ranking, due to the large total output of the higher ranking institutions, the actual number of collaborative paper still increases as the ranking increases.

5.3.3 Citations and collaborated papers

The relationship between Citations and Collaborative Paper (and the omitted size-weighted collaboration) has the highest positive correlation coefficient in each of the disciplines. Materials Science has the highest correlations of reaching $r = 0.909$, while Law has the lowest significant correlation of $r = 0.805$. Although the high correlation indicates strong relationships between the measured variables, it is also an indication of the amount of information overlap carried with each of the variables. Since it has already been shown that productivity and collaborative paper have high $r = 0.9$ range of correlation, and productivity also has about $r = 0.9$ correlation with citation counts, it is therefore expected to find a high correlation between citation and collaborative papers, simply due to the overlapping information between these variables. In order to have a deeper understanding of the real relationships between these variables, chapter 6 employs linear regression methodology to remove the effect of one variable from another, before calculating the correlation coefficient.

5.3.4 Citations and percent collaboration

CITTOT showed the strongest negative correlation with the percent collaboration, approaching $r = -0.5$ in Computer Science, Pharmacology and Psychology. CITTOTw showed very similar results to CITTOT, so it was omitted from the figure for clarity. All impact metrics – CITTOT, CITTOTw, CITAV and WRANK are showing negative correlations with PUBCOLL%.

Law is, again, an exception where no significant correlation was found, but the other SSH discipline – Psychology, however, is demonstrating a significant correlation. This showed again that Psychology is closer to NSE in its research collaboration and citation patterns.

5.3.5 Citations per paper and collaborated papers

The institutions' average individual paper impact shows more than 0.4 significant correlation with the institution's collaborative papers in all of the discipline sets. That is, the more papers collaborated, the better the individual paper's impact. However, this is not to be interpreted as the collaboration caused the higher paper impact, or vice versa.

5.3.6 Citations per paper and percent collaboration

Average single paper impact of an institution shows significant negative correlation with the percent the institution's collaborative paper in NSE disciplines. A negative correlation means that the higher percentage the institutions' collaborative papers are, the lower the individual paper's impact.

In SSH, the picture is different, Psychology does not show any correlation between the two variables; while Law shows a significant but small positive correlation, which is in contrast with what was found in NSE. This is an evidence that Law and Psychology have different collaboration strategies, and are different from the NSE disciplines.

5.4 Chapter Summary

In this chapter, we investigated the pairwise relationships between the three factors relating to institutional research activities: research productivity, research collaboration and research impact. Five disciplines across two datasets were analysed and compared. This chapter used the methodology commonly applied in previous studies to establish relationships between variables, the results were compared to the previous works, and our results were mostly consistent with previous works.

However, we found that institutional impact and productivity positively correlate with each other, which contradicts previous findings between company's patent impact and patent productivity. We further investigated the question on the individual level, to see whether at individual level an author publishes more paper receives more citations. The result showed large disciplinary differences where Computer Science (ACM) and Psychology demonstrated significant positive correlations; Materials Science, Pharmacology and WoS Computer Science showed negative correlations, while Law showed no correlation. This means in Law, publishing more papers makes no difference in the number of citations one would receive on average for each paper. In Computer Science (ACM) and Psychology, the more an author publishes, the more citations he receives on average for each paper, making him receiving an increased total citations (total citations = paper x citations per paper). This implies that researchers in these disciplines should publish more paper when they can because the extra papers are beneficial to their total citation impact. In Materials Science, Pharmacology and WoS Computer Science, the more an author publishes the fewer citations they receive per paper. So researchers in these disciplines should not aim to publish more papers, instead, should perhaps publish more higher impact papers.

In order to explain why Law, Materials Science, Pharmacology and WoS Computer Science have inconsistent institution and individual level correlation, we further split the authors into three productivity tiers – high, medium and low – for these disciplines. We found consistent strong positive correlations for low productive authors while no correlation or negative correlation existed among high productivity author. This indicates that low productive authors receive more-than-average citations for each extra paper they publish. High productive author, on the other hand, may see their citation per paper reduced when publishing extra papers. (i.e. the extra papers published receive below-average citations, so lowers the new average citation with this extra paper counted in.)

This analysis reveals the institution's composition in terms of high and low productive authors. At the institution level, a positive correlation was found, same as the low productive authors group. The high productivity authors give no/negative correlations, which is opposite of what found at the institution level. So in order for the institution level's correlation to remain positive, the majority of the authors need to come from the low productive group, while the number of high productive authors need to be fewer.

Previous study indicated that the research collaboration at the country level cluster together, where highly developed countries mostly collaborate with each other, leaving out the less developed counties [46, 81]. We have not found the same pattern at the institution level. At the institution level, top institutions in terms of productivity and collaboration have an equally large number of institution collaborators. Unlike countries, top institutions do not solely collaborate with a small group of institutions while leaving out the rest.

The collaboration percentage (the ratio between collaborative papers over total papers) for institutions showed negative correlation with impact and productivity. This means that institutions with higher proportion of the papers that are collaborated are generally less impact and productivity. This could be due to the lower research capabilities of these institutions that collaboration is one of their only ways to be involved in research publication.

The weighted variables – Size-weighted collaboration and PageRank weighted citation – don't particularly alter the correlation coefficient compared to their original counterpart. The correlation coefficient of these weighted variables gave consistently smaller values across subjects.

The external variable – Webometrics university ranking – showed positive correlation to productivity and collaboration in all of the disciplines, which to a certain extent, confirms that to publish more or to collaborate more would increase the institutions' rank in this league table. However, since the metrics used to compose the league table includes sub-variable such as 'number of papers found on web', there is no surprise that institutions with more papers published would be ranked higher.

Finally, the un-weighted variables showed strong, positive correlations across all disciplines consistently. These three variables also circularly correlate with each other, which prevent us from understanding the true relationship between any of the pairs. In the next chapter, we use partial correlation to isolate the circular correlations and attempt to reveal the true relationships between these variables.

		PUBCOLL	PUBCOLL	PUBCOLL	PUBCOLL	PUBCOLLw	PUBCOLLw	PUBCOLLw	PUBCOLLw	PUBCOLL%	PUBCOLL%	PUBCOLL%	PUBCOLL%
		CITTOT	CITTOTw	CITAV	WRANK	CITTOT	CITTOTw	CITAV	WRANK	CITTOT	CITTOTw	CITAV	WRANK
CS (ACM)	rho	.832	.822	.525	.647	.788	.785	.502	.624	-.540	-.481	-.309	-.498
	sig.	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	n	1609	1609	1609	1609	1609	1609	1609	1609	1609	1609	1609	1609
CS (WoS)	rho	.880	.865	.490	.452	.861	.851	.488	.453	-.291	-.319	-.154	-.294
	sig.	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	n	3742	3742	3742	3742	3742	3742	3742	3742	3742	3742	3742	3742
Pharmacology	rho	.898	.891	.423	.376	.874	.873	.410	.377	-.466	-.451	-.242	-.285
	sig.	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	n	3251	3251	3251	3251	3251	3251	3251	3251	3251	3251	3251	3251
Material Sciences	rho	.909	.898	.411	.368	.899	.895	.420	.384	-.423	-.395	-.176	-.259
	sig.	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	n	3305	3305	3305	3305	3305	3305	3305	3305	3305	3305	3305	3305
Psychology	rho	.856	.845	.557	.589	.811	.804	.534	.560	-.198	-.173	-.037	-.165
	sig.	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.047	.000
	n	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827
Law	rho	.805	.785	.476	.523	.780	.766	.472	.500	.013	.031	.090	-.012
	sig.	.000	.000	.000	.000	.000	.000	.000	.000	.644	.269	.001	.670
	n	1300	1300	1300	1300	1300	1300	1300	1300	1300	1300	1300	1300

Table 5.11: Correlation coefficient between collaboration and impact. PUBCOLL showed significant positive correlations with all three citation metrics in all disciplines investigated. On the other hand, PUBCOLL% showed significant *negative* correlation with all of the citation metrics in Computer Science, Pharmacology and Materials Science. Interestingly, the negative correlation diminishes in Psychology and Law, which happen to be NSE disciplines. This could suggest a division in collaboration attitude between NSE and SSH disciplines.

Chapter 6

Partial Correlations Among Research Productivity, Research Impact and Research Collaboration

In the previous chapter, statistical correlation was applied between each pairs of the metrics that measure institutions' productivity, impact and collaborativity. The results demonstrated that strong relationships exist among them. These are in agreement with past research [137, 159, 205] and gave positive evidence on the assumptions made by funding agents and institution policy makers.

However, the strong inter-correlations exist between all pairs also suggested that there can be common factors presented in these variables. The correlations found in any pair of the variables may merely reflect the common variable presented, masking the real correlations between them. This chapter tries to use partial correlation to isolate the third variable in an attempt to find the true correlations.

Partial correlation calculation among the institutional research collaborativity, productivity and impact has not been attempted in the scale such as the current study.

This chapter splits into three sections. Section 6.1 interprets the effect partialling, describes what it means to remove a particular variable from another and gives an interpretation of the resulting variables after effects have been partialled; section 6.2 presents the correlation outcome for each permutation of the variable pairs for each discipline; section 6.3 discusses and interprets the correlation results.

6.1 Interpretation of the Factor Partialling

The partialling of a factor can be understood as *removing* the effect of the factor from a variable (controlling, removing and partialling out will be used interchangeably in this thesis). The new variable with the effect removed will not respond to the changes in the removed factor. That is, the new variable is *independent* of the removed variable. This way, the measurements of this new variable can be compared fairly between institutions, knowing that it is not affected by the variable removed. e.g, institution A published 10 papers and received 100 citations; institution B published 40 papers and received 200 citations. Without removing the productivity effect, institution B received more citations and hence a better result. However, when the paper count is controlled (*e.g.* divide the citations by the number of papers), institution A receives 10 citations per paper while B receives 5 citations per paper. With productivity removed, A comes out better than B.

It is often difficult to interpret the remaining variable after the effects have been controlled. The naive way is to understand and read the variable as the removing variable "per" removed variable. For example, after the collaborative paper effects have been removed from the total papers at the institution level, the new variable can be interpreted as the *papers per collaborative paper*, that is, the number of papers an institution publishes for every collaborative paper they publish. This offers us a start point for understanding the new variable after the partialling. In the following sections, we explain why one factor may be presented in a variable and give an interpretation of the variables after the effects have been partialled.

6.1.1 Factor Partialling Between Productivity and Impact

Institutions with high impact work published are more visible in the field, this is because the high impact work is published in reputable journals and generally is cited more, so more researchers have read it. Collaborations and projects opportunities, as a result of the higher visibility, will find the authors more easily. In addition to the visibility advantage, researchers with high impact output also make collaborators more willing to work with them. More collaboration opportunities and more willing collaborators make the institutions publish more compared to the less visible institutions.

Partialling out the institutions' impact from their productivity variable adjusts their productivity metric (paper count) in a way that those high impact institution's advantages are removed. So that the remaining variable can be compare and correlated fairly.

This way, institutions' past performance will not affect the adjusted productivity of the institutions. The less productive institutions or smaller institutions are not disadvantaged for their high impact research output with this new variable. The smaller institutions with fewer researchers may publish fewer papers, but their work could be field leading. Simply compare raw citation number would disadvantage them, resulting an unfair comparison of their impact.

One productivity variable (***PUBTOT***) and four impact variables (***CITTOT, CITTOTw, WRANK*** and ***CITAV***) are considered in this study. The front three variables measure the institutions' *overall* impact while the last one measures the institutions' average paper impact. When partialling out impact variables from productivity, all of the four variables were partialled out one by one from ***PUBTOT***. This process was completed by an algorithm provided by SPSS. The productivity with the impact effect removed are represented as $PUBTOT$ – the same symbol without bold face. The four impact variables with ***PUBTOT*** removed are represented as: $CITTOT$, $CITTOTw$, $CITAV$ and $WRANK$.

6.1.2 Factor Partialling Between Collaboration and Impact

From the results shown in chapter 4, as well as in previous research [137, 156], institutional collaboration has been found to have a relationship with the impact of the institutions. Institutions that made higher research impact and ranked high in the league tables are in a better position to collaborate. They have more chance to have the leading experts and necessary equipments that collaborators desired to have in a collaboration (hence the high impact papers as well as high impact of the institution's output), giving them the advantage in producing more collaborative research.

Partialling out the impact variables adjusts the collaborations so that the extra collaborations gained due to these high impact effects are removed. The four impact variables were removed one by one from each of the collaboration variable, resulting: $PUBCOLL$, $PUBCOLLw$ and $PUBCOLL\%$.

On the flip side, institutional collaboration also affects institution's impact. The more collaborations institutions participate in, the more papers are published by this institution, so more citations are likely to be received. Collaboration also provides opportunities for a researcher's publicity, social network and learning opportunities which all, in one form or another, leading to an institution's higher research impact. Partialling out the collaboration variables removes these potential biases in measuring institution's research impact.

6.2 Partial Correlation Results

We apply the partial correlation to control the overlapping variables that may have increased the correlation coefficients. All three possible permutation of the partial correlations are computed:

- Collaboration vs. Productivity with Impact controlled
- Collaboration vs. Impact with Productivity controlled
- Productivity vs. Impact with Collaboration controlled

We remove the controlled variable from *both* variables prior to the computation of the correlation. For instance, to compute the correlation between collaboration and productivity with impact controlled, impact is partialled out from both collaboration variables and productivity variable, before the collaboration variables and productivity variables are computed for correlation.

6.2.1 Collaboration versus Productivity Controlled for Impact

		PUBTOT PUBCOLL% -Q	PUBTOT PUBCOLLw -Q	PUBTOT PUBCOLL -Q
CS (ACM)	rho	-.357	.468	.592
	Sig.	.000	.000	.000
	n	1609	1609	1609
CS (WoS)	rho	-.134	.587	.706
	Sig.	.000	.000	.000
	n	3588	3588	3588
Pharmacology	rho	-.153	.421	.610
	Sig.	.000	.000	.000
	n	3251	3251	3251
Materials Science	rho	-.126	.369	.721
	Sig.	.000	.000	.000
	n	3305	3305	3305
Psychology	rho	-.446	.474	.586
	Sig.	.000	.000	.000
	n	2827	2827	2827
Law	rho	.013	.071	.023
	Sig.	.706	.020	.450
	n	1083	1083	1083

Table 6.1: Correlation coefficient between Productivity and Collaboration, with Impact factors controlled for. Disciplines except Law held a significant and high correlation between PUBTOT and PUBCOLL. The non-significant correlation in Law suggests that impact factors are the primary reason for the high correlations between productivity and collaborativity. While in the remaining disciplines, impact factors did not play as deep role as in Law.

6.2.1.1 Computer Science

The correlation in Computer Science (Figure 6.1) showed an uniform reduction in all pairs compared to the unpartialed results. Although the size of the correlation was reduced, all correlation remain significant at $p < 0.01$.

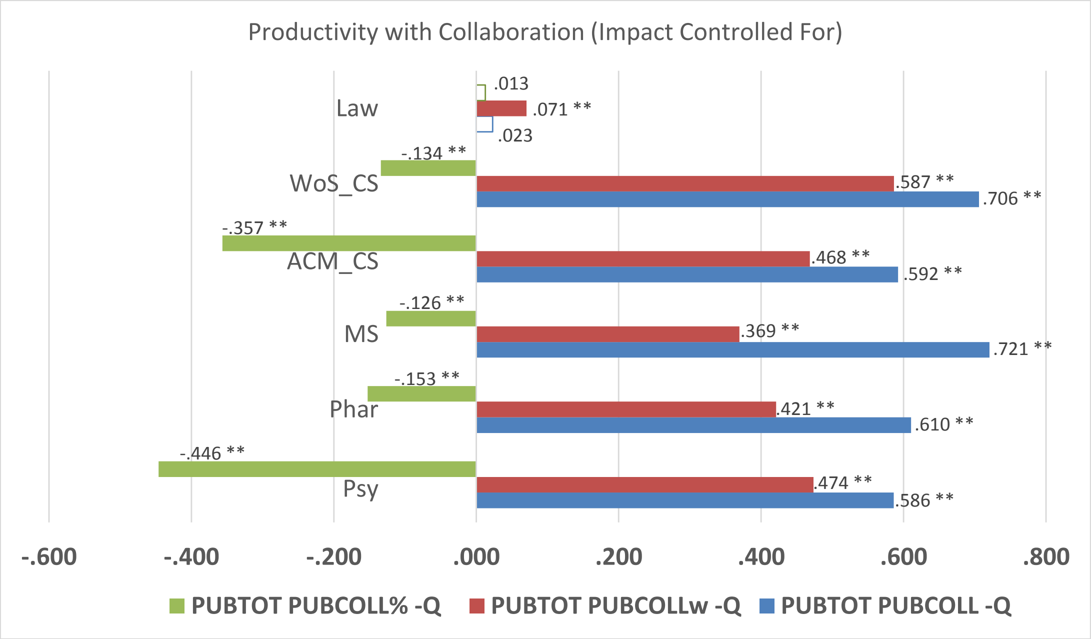


Figure 6.1: Correlation results of institutional productivity and collaboration, with impact factors controlled. (Visualisation of table 6.1) ** indicates $p < 0.01$

In ACM dataset, the *PUBTOT PUBCOLL* correlation was $r = 0.592$, a reduction from $r = 0.909$; the *PUBTOT PUBCOLLw* was $r = 0.468$, dropped from $r = 0.858$; the direction of the correlation coefficient remained negative for *PUBTOT PUBCOLL%*, the size was dropped from $r = 0.635$ to $r = 0.357$.

In WoS dataset, the *PUBTOT PUBCOLL* correlation reduced to $r = 0.706$ from $r = 0.962$; the *PUBTOT PUBCOLLw* is $r = 0.587$, a reduction from $r = 0.931$ without partialling impact; the direction of the correlation remained negative, same as the uncontrolled correlation before, only the size decreased from $r = 0.326$ to $r = 0.134$.

With impact factors controlled, the correlation between the productivity and collaboration proportionately reduced in both datasets in Computer Science. This true correlation between productivity and collaboration is not as high as it was found in pairwise analysis. Impact as measured by citations, average citations and ranking, is a partial indicator for productivity and collaboration, this effect of institutional impact is presented in both productivity and collaboration, and positively affected them.

Both before and after partialling, the two measurements of the collaborative papers (unweighted and weighted) demonstrated significant positive correlations. If an institution were found with high number of Computer Science collaborative papers, then their total Computer Science paper output should be high too. However, the casual direction is not determined.

Comparing the ACM result to the WoS result, ACM dataset showed a smaller correlation size in the positive pairs ($PUBTOT\ PUBCOLLw\ \ PUBTOT\ PUBCOLL$), while larger correlation size in the negative pair ($PUBTOT\ PUBCOLL\%$). This could be due to that fewer university samples were available in the ACM dataset for the relatively smaller number of collaborative papers, (20,043 collaborative papers in ACM vs 164,553 collaborative papers in WoS) which lead to a less predictability between paper count and collaboration metrics, giving a smaller correlation coefficient.

6.2.1.2 Psychology

In figure 6.1, the $PUBTOT\ PUBCOLL$ has reduced from $r = 0.931$ to $r = 0.586$ and $PUBTOT\ PUBCOLLw$ has reduced from $r = 0.872$ to $r = 0.474$, but $PUBTOT\ PUBCOLL\%$ has gained correlation from $r = -0.302$ to $r = -0.446$. All correlation were significant with $p < 0.01$.

Impact did not found to be a strong correlating factor to either productivity or collaboration in Psychology (as opposed to Law, which impact had strong impact for both productivity and collaboration). The two positive correlations – $PUBTOT\ PUBCOLLw$ and $PUBTOT\ PUBCOLL$ had demonstrated the similar level of correlation as the NSE discipline. In this respect, Psychology has aspects of the publication practices that is closer to NSE than SSH discipline.

The increase of the $PUBTOT\ PUBCOLL\%$ suggests that institutional qualities (institutional citations, ranking and paper impact) have negative impact on the predictability between paper count and percent collaboration. Fewer cited institutions have larger chance in Psychology to publish less proportion of collaborated papers than in other disciplines. However, this result must be viewed with caution as the small correlation size increase may came from errors.

6.2.1.3 Pharmacology

For Pharmacology in figure 6.1, each of the correlation coefficient had an uniform reduction from the previous uncontrolled correlation, maintaining the shape of the bars. The *PUBTOT PUBCOLL* reduced from $r = 0.976$ to $r = 0.610$; the *PUBTOT PUBCOLLw* had a higher reduction than *PUBTOT PUBCOLL*, from $r = 0.947$ to $r = 0.421$. The size of the negative correlation *PUBTOT PUBCOLL%* has dropped from $r = 0.496$ to $r = 0.153$. All three correlations were still significant at $p < 0.01$.

Similar to WoS Computer Science, institutional impact has shown a moderate effect on the productivity and collaboration in Pharmacology. Impact of the institution plays a role in Pharmacology, but it is not the primary factor. Institutional productivity is a stronger predictor for institutional collaboration and vice versa.

6.2.1.4 Materials Science

Materials Science (Figure 6.1) has the highest *PUBTOT PUBCOLL* correlation in all disciplines, reaching $r = 0.721$, however, it has the second lowest *PUBTOT PUBCOLLw* correlation of $r = 0.369$, a drop from one of the highest uncontrolled ***PUBTOT PUBCOLLw*** of $r = 0.958$. The *PUBTOT PUBCOLL%* correlation although holding the negative correlation, it became the smallest of all disciplines with $r = -0.126$. All correlations were significant at $p < 0.01$.

The larger reduction in all three pairs in Materials Science could indicates the stronger institutional impact's influence in the institutional productivity and collaboration. In particular, the largest reduction in size-weighted collaboration reveals that the impact of the institutions (as measured by citation, rank and paper citation) closely linked to the size of Materials Science research collaborations. The size of a collaboration (the number of collaborators in each of the collaboration) and the number of researchers to contribute to the collaboration are both important factors the institutions have to consider in Materials Science.

6.2.1.5 Law

Law (Figure 6.1) is the only discipline investigated that had a drastic correlation reduction after partialling the impact. The *PUBTOT PUBCOLL* dropped from $r = 0.858$ to insignificant; the *PUBTOT PUBCOLLw* reduced from $r = 0.820$ to $r = 0.071$(sig. at $p < 0.05$); the correlation $r = -0.065$ between ***PUBTOT PUBCOLL%***(sig. at $p < 0.05$) had disappeared once the impact has been controlled.

These large reductions suggest that the impact is one of the most important factor that relates to institutions' productivity and collaboration in Law.

However, the results regarding the collaboration must be viewed with caution. Collaboration is not a popular activity in Law when publishing scientific results. Only less than one tenth of the papers were collaborated in our dataset, in contrast to nearly half of the papers were collaborated in NSE. The collaboration patterns demonstrated may not be a good indicator of the overall institutional publication practice in Law.

6.2.1.6 Productivity and proportion of collaborated papers

Previous studies have demonstrated that high productivity tend to linked with lower proportion of the collaboration. Katz [105] found that universities with higher number of paper output, have a lower ratio of collaborative papers over total papers, than those published fewer. Davidson and Carpenter [52] reported the same finding at the country level: the more the papers a country publish, the less the percentage of co-authored papers. The same finding is confirmed by Luukkonen *et al.* [122]. Schubert and Braun [167] observed that foreign co-authorship can be approximated by national publication productivity through a power law in which the exponent is *less than one.* Big countries have thus, in general, lower shares of international co-publications than medium-sized or small countries have.

In this study, Psychology, Pharmacology, Materials Sciences, ACM Computer Science and WoS Computer Science have all been shown a negative correlation of number paper and percent collaboration, which means that institutions with higher number of papers, their proportion of collaborated papers is lower. This finding is in agreement with the previous studies, with exception in Law, in which no significant correlation was found.

6.2.1.7 Discussion of Collaboration vs Productivity with Impact Controlled

After the impact variables were partialled out, the correlation coefficient between $PUBTOT$ and $PUBCOLL$ were reduced. A higher reduction was observed with the weighted collaboration variable $PUBCOLLw$, but they were still significantly positive. A reduction indicates that the impact was contributing to the correlations between $\boldsymbol{PUBTOT}$ and $\boldsymbol{PUBCOLL}$. While impact contributed in Pharmacology, Psychology, Materials Science and Computer Science moderately, however, it was the most important factor in Law because the correlation became non-significant after impact was partialled out.

The negative correlations found between $\boldsymbol{PUBTOT}$ and $\boldsymbol{PUBCOLL\%}$ before the partialling were confirmed again after impact variables were partialled out. Even though the size of the correlation was reduced in many disciplines, they remain significant. This means that *regardless of the institution's impact, the more the institutions published papers, the less percentage of their papers were collaborated.* Despite the collaboration has been heavily favoured in the recent years, institutions, especially those highly productive ones, do not seem to collaborate heavily compared to their total output. They still focus on singly authored papers. On the other hand, the lowly productive institutions seem to have taken collaboration quite seriously and we observed many low productive institutions with almost all of their output collaborated.

However, Law is an exception in this regard. Correlations in Law disappeared as the impact variables have been partialled out, only the $PUBTOT$ and $PUBCOLLw$ correlation remained significant, but very small. This gives an strong evidence that the removed factor – institutional research impact – was affecting directly with productivity and collaboration, while the real relationships between institution's productivity and collaboration is non-existent.

		PUBTOT	PUBTOT	PUBTOT	PUBTOT
		CITTOT	CITTOTw	CITAV	WRANK
		-C	-C	-C	-C
	rho	.091	-.263	.107	.002
Law	sig.	.006	.000	.001	.960
	n	888	992	888	1047
	rho	.438	.276	-.006	.124
CS (WoS)	sig.	.000	.000	.705	.000
	n	3742	3588	3742	3742
	rho	.082	-.074	.361	.202
CS (ACM)	sig.	.000	.003	.000	.000
	n	1841	1609	1841	1841
	rho	.011	-.347	.108	-.046
Materials Science	sig.	.528	.000	.000	.009
	n	3196	3121	3196	3196
	rho	.180	-.184	.095	.145
Pharmacology	sig.	.000	.000	.000	.000
	n	3104	3086	3104	3104
	rho	.364	.056	.122	.088
Psychology	sig.	.000	.005	.000	.000
	n	2557	2557	2557	2557

Table 6.2: Correlation coefficient between productivity and impact, with collaboration factors controlled.

6.2.2 Productivity versus Impact Controlled for Collaboration

6.2.2.1 Computer Science

Figure 6.2, the correlation of the Computer Science datasets showed strong reduction to the size of the coefficient in general, some even turned negative. In the ACM dataset, the correlation of $PUBTOT$ with $CITTOT$, $CITAV$ and $WRANK$ reduced to $r = 0.082$, $r = 0.361$ and $r = 0.202$ respectively, but they were all significant. The correlation between $PUBTOTandCITTOTw$ has turned negative to $r = -0.074$. In the WoS dataset, the correlation of $PUBTOT$ with $CITTOTw$, $CITTOT$ and $WRANK$ remained significant and positive, with values $r = 0.276$, $r = 0.438$ and $r = 0.124$ respectively. The correlation of $PUBTOTandCITAV$ has become statistically insignificant.

Although there are large disagreements between the two datasets once the collaborativity variables have been removed, the correlation with $CITTOT$ and $WRANK$ remained positive in both dataset, confirming the positive relationship between productivity and impact in two of the four variables after partialling.

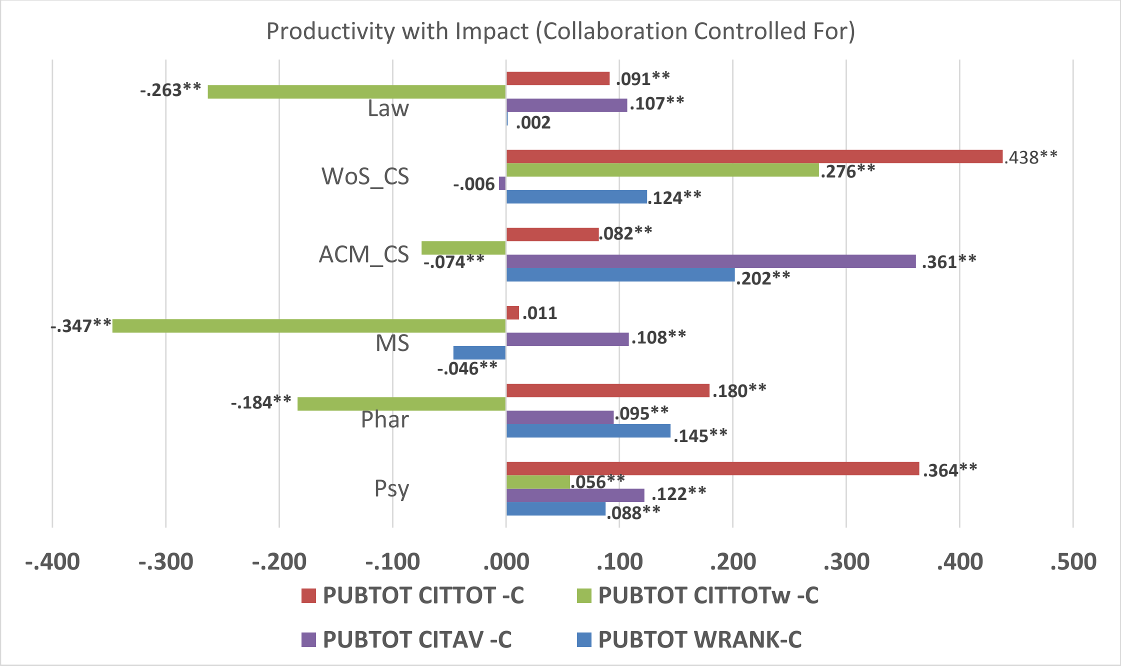


Figure 6.2: Correlation results of institution productivity and impact, with collaboration factors controlled (Visualisation of table 6.2) ** indicates $p < 0.01$

6.2.2.2 Psychology

All four pairs of correlations were still significant and positive but reduced in size compared to the un-partialled results. The size reduction of the correlation coefficient was large. The correlation of $PUBTOT$ and $CITAV$ has reduced from $r = 0.537$ to only $r = 0.122$; the correlation of $PUBTOT$ and $CITTOTw$ has reduced from $r = 0.856$ to $r = 0.056$; while the $PUBTOT$ $WRANK$ reduced from $r = 0.608$ to $r = 0.088$. The correlation of $PUBTOT$ and $CITTOT$ was the largest, with $r = 0.364$.

The reduction of the correlations suggested that the effect of removed variables was presented in either/both of the institutional productivity and impact.

6.2.2.3 Pharmacology

All pairs of correlations were significant in Pharmacology. The correlations of $PUBTOT$ with $CITAV$, $CITTOT$ and $WRANK$ were significant and positive, but small, with $r = 0.095, r = 0.180$ and $r = 0.145$ respectively. The $PUBTOT$ and $CITTOTw$, which were high and positive before the partialling, has turned negative with $r = -0.184$.

6.2.2.4 Materials Science

The correlation of $PUBTOT$ and $CITAV$ was significant and positive, with $r = 0.108$; the correlations between $PUBTOT$ and $CITTOT$ were no longer significant, while $CITTOTw$ and $WRANK$ has significant, but negative correlation with $PUBTOT$. The correlation between $PUBTOT$ and $CITTOTw$ has one of the largest negative correlation of $r = -0.347$; the correlation between $PUBTOT$ and $WRANK$ was $r = -0.046$.

Materials Science is the only discipline that had a negative correlation between PUBTOT and WRANK. Even though it is not big, it gives indication that after removing Collaboration, the higher the institutions were ranked, the lower their published papers in Materials Science would be.

6.2.2.5 Law

The correlations of $PUBTOT$ with $CITTOT$ and $CITAV$ are significant and positive with $r = 0.091$ and $r = 0.107$ respectively; $PUBTOT$ and $CITTOTw$ had a negative correlation of -0.263; while no correlation was found between $PUBTOT$ and $WRANK$.

6.2.2.6 Discussion of Productivity vs Impact with Collaboration Controlled

In this section, we shifted our focus to the relationship between institutional productivity and impact. Previously, productivity was found to have strong, positive correlations with impact before collaborativity was controlled. This result was in agreement with works presented by Presser and He[93, 156].

However, with collaboration variables partialled out, the resulting correlation between productivity and impact varies from discipline to discipline. Some disciplines' correlation turned negative while others remained strong and positive. These complex correlation changes implies the collaboration's deep and non-trivial role in the institution's productivity and impact. In addition, there is no identifiable SSH and NSE pattern split in the institutional productivity and impact correlations.

6.2.3 Collaboration versus Impact Controlled for Productivity

Figure 6.3 shows the correlation result between impact and collaboration with the productivity factor controlled. After the partialling of the productivity factor, the correlation coefficient has reduced in all disciplines.

6.2.3.1 Computer Science

Partialling for productivity removed many of the significant and high correlations between impact and collaboration variables (Figure 6.3 A and B). The three correlation pairs: *CITTOT PUBCOLL%*, *CITTOT PUBCOLLw* and *CITTOTw PUBCOLL* in WoS has become non-significant. The other pair of the correlations has reduced strongly, though still significant.

An interesting change in the correlation was between *CITTOTw* and *PUBCOLL%* in the ACM dataset. The coefficient has turned positive from a negative value, $r = -0.481$ to $r = 0.165$.

In fact, the correlation of Webometrics ranking and collaborativity has turned sign too for both of the datasets (Figure 6.3 D). The correlation with *PUBCOLL*, *PUBCOLLw* had turned negative in both of the Computer Science datasets. They were medium and positive before partialling the productivity.

The correlation between *PUBCOLL%* and *WRANK* maintained the negative correlation, with the size slightly reduced to $r = -0.151$ in ACM dataset and $r = -0.147$ in WoS dataset.

6.2.3.2 Psychology

The correlation between untransformed citation metrics (represented by *CITTOT* and *CITTOTw* in the graph) and untransformed collaboration metrics (*PUBCOLL* and *PUBCOLLw*) has reduced from in range of $r = 0.8$ down to $r = 0.1$, with the pair *CITTOTw* and *PUBCOLL%* no longer significant.

		CITTOT PUBCOLL -P	CITTOT PUBCOLLw -P	CITTOT PUBCOLL% -P	CITTOTw PUBCOLL -P	CITTOTw PUBCOLLw -P	CITTOTw PUBCOLL% -P	CITAV PUBCOLL -P	CITAV PUBCOLLw -P	CITAV PUBCOLL% -P	WRANK PUBCOLL -P	WRANK PUBCOLLw -P	WRANK PUBCOLL% -P
Psy	rho	.138	.140	.101	.205	.169	.028	.251	.100	.011	.067	.042	.003
	sig.	.000	.000	.000	.000	.000	.164	.000	.000	.593	.000	.023	.887
	n	2896	2896	2896	2827	2827	2534	2896	2896	2557	2896	2896	2557
Phar	rho	.012	-.001	-.033	.113	.088	-.001	.010	-.006	-.043	-.168	-.082	-.101
	sig.	.481	.966	.066	.000	.000	.965	.567	.718	.016	.000	.000	.000
	n	3287	3287	3104	3251	3251	3086	3286	3286	3104	3287	3287	3104
MS	rho	.052	.179	.000	.166	.241	.031	.044	.124	.006	-.036	-.026	-.058
	sig.	.002	.000	.987	.000	.000	.084	.011	.000	.723	.037	.127	.001
	n	3413	3413	3196	3305	3305	3121	3413	3413	3196	3413	3413	3196
CS (ACM)	rho	.149	.066	.059	.280	.167	.165	-.228	-.135	-.106	-.195	-.109	-.151
	sig.	.000	.005	.011	.000	.000	.000	.000	.000	.000	.000	.000	.000
	n	1841	1841	1841	1609	1609	1609	1841	1841	1841	1841	1841	1841
CS (WoS)	rho	-.058	.017	-.014	-.020	.099	-.045	.134	.114	.054	-.178	-.090	-.147
	sig.	.000	.300	.388	.229	.000	.007	.000	.000	.001	.000	.000	.000
	n	3742	3742	3742	3588	3588	3588	3742	3742	3742	3742	3742	3742
Law	rho	.804	.777	-.476	.764	.742	-.497	.295	.308	.046	.498	.477	-.375
	sig.	.000	.000	.000	.000	.000	.000	.000	.000	.169	.000	.000	.000
	n	1096	1096	888	1300	1300	992	1096	1096	888	1425	1425	1047

Table 6.3: Correlation coefficient between impact and collaboration, with productivity factors controlled.

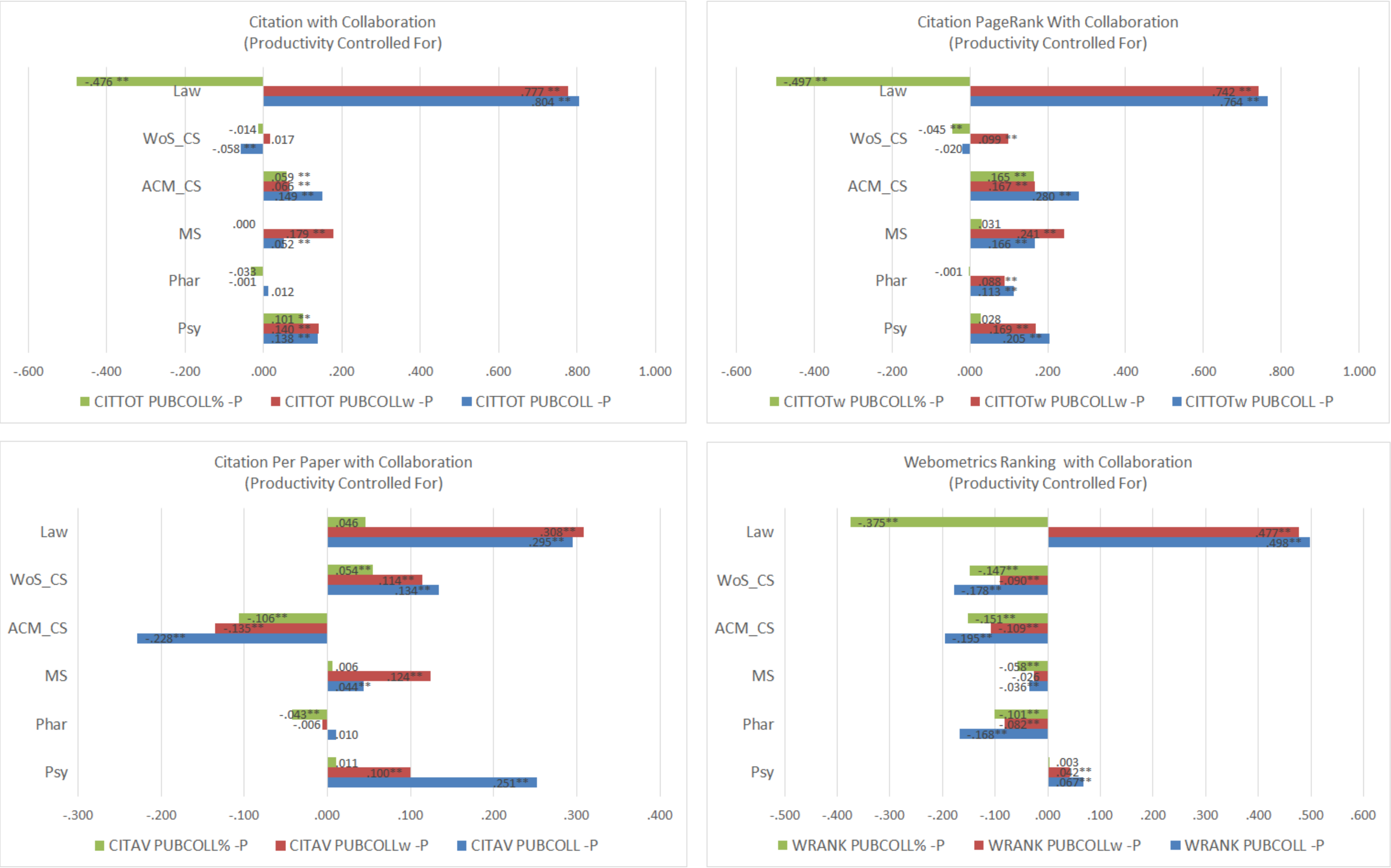


Figure 6.3: Correlation results of institution productivity and impact, with collaboration factors controlled (Visualisation of table 6.3) ** indicates $p < 0.01$

$CITAV$ has shown a significant positive correlation with $PUBCOLL$ and $PUBCOLLw$, the size of the correlation has reduced. $CITAV$ $PUBCOLL$ is now $r = 0.251$ and $CITAV$ $PUBCOLLw$ is $r = 0.100$.

The percent collaboration ($PUBCOLL\%$) and citations per institution ($CITTOT$) has turned positive, from $r = -0.198$ to $r = 0.101$; the percent collaboration with the other untransformed citation metrics ($CITTOTw$) has become insignificant, It was $r = -0.173$ before partialling the productivity.

The citation per paper ($CITAV$) and percent collaboration ($PUBCOLL\%$) has no significant correlation, same as before partialling the productivity.

$WRANK$ has shown significant and positive correlation with $PUBCOLLw$ and $PUBCOLL$. $PUBCOLLw$ and $PUBCOLL$ had positive correlation with $WRANK$, this correlation has reduced, with $r = 0.042$ and $r = 0.067$ respectively. $WRANK$ $PUBCOLL\%$ had negative correlation before and the correlation has become insignificant after the productivity partialling.

6.2.3.3 Pharmacology

The correlation of $CITTOT$ $PUBCOLL$ and $CITTOT$ $PUBCOLLw$ has become insignificant. It was a significant high positive correlation before the productivity partialling. The correlation of $CITTOTw$ $PUBCOLL$ and $CITTOTw$ $PUBCOLLw$ remained positive and significant, but the size has reduced. It was in the range of $r = 0.8$ while they are in the range of $r = 0.1$.

$CITAV$ has shown no significant correlation with $PUBCOLL$ and $PUBCOLLw$. They were significant and positive correlations in the range of $r = 0.4$.

The correlation of $PUBCOLL\%$ $CITTOT$ and $PUBCOLL\%$ $CITTOTw$ has become insignificant. They were significant and negative correlation in the range of $r = -0.4$ before partialling the productivity.

The correlation between $CITAV$ and $PUBCOLL\%$ has shown significant and negative correlation, same as before partialling the productivity, but the size has reduced to $r = -0.043$.

$WRANK$ has shown significant and negative correlation with $PUBCOLLw$, $PUBCOLL$ and $PUBCOLL\%$. $PUBCOLL$ and $PUBCOLLw$ had positive correlation with $WRANK$, now this correlation has turned negative, with $r = -0.168$ and $r = -0.082$ respectively. $WRANK$ $PUBCOLL\%$ had negative correlation before. The correlation was reduced after the productivity partialling.

6.2.3.4 Materials Science

The correlation between untransformed citation metrics $CITTOT$ and $CITTOTw$, and untransformed collaboration metrics $PUBCOLL$ and $PUBCOLLw$ has reduced from in range of $r = 0.9$ down to $r = 0.2$, but these pairs were positive and significant.

$CITAV$ has shown no significant correlation with $PUBCOLL$, which was significant and positive before. $CITAV$ and $PUBCOLLw$ had its correlation coefficient reduced to $r = 0.124$. It was $r = 0.420$ before.

The correlation of $PUBCOLL\%$ $CITTOT$ and $PUBCOLL\%$ $CITTOTw$ has become insignificant. They were significant and negative correlation in the range of $r = -0.4$ before partialling the productivity.

The correlation between $CITAV$ and $PUBCOLL\%$ has become insignificant, which was significant and negative before partialling the productivity.

$WRANK$ has shown no significant correlation with $PUBCOLLw$ and has shown negative correlation with $PUBCOLL$ and $PUBCOLL\%$. $PUBCOLL$ and $PUBCOLLw$ had positive correlation with $WRANK$ before, $WRANK$ $PUBCOLL$ has turned negative, with $r = -0.036$. $WRANK$ $PUBCOLL\%$ had negative correlation before. The correlation was reduced to $r = -0.058$ after the productivity partialling.

6.2.3.5 Law

The correlation of $CITTOT$ $PUBCOLL$ and $CITTOT$ $PUBCOLLw$ had almost no change with the partialling of productivity. All four pairs of the correlation were still in the range of 0.8.

$CITAV$ has shown significant positive correlation with $PUBCOLL$ and $PUBCOLLw$. They were significant and positive correlation in the range of $r = 0.5$ before the partialling of the productivity. These correlations were reduced to $r = 0.295$ for $CITAV$ $PUBCOLL$ and $r = 0.308$ or $CITAV$ $PUBCOLLw$.

The correlation of $PUBCOLL\%$ $CITTOT$ and $PUBCOLL\%$ $CITTOTw$ has turned significant and negative, $PUBCOLL\%$ $CITTOT$ were $r = -0.476$ while $PUBCOLL\%$ $CITTOTw$ were $r = -0.497$. Both of these pairs were insignificant before partialling the productivity.

The correlation between $CITAV$ and $PUBCOLL\%$ has become insignificant, which was significant and positive with $r = 0.090$ before partialling the productivity.

$WRANK$ has shown significant and positive correlation with $PUBCOLLw$ and $PUBCOLL$, with $r = 0.477$ and $r = 0.498$ respectively. The correlation was in the range of 0.5 before. $WRANK$ $PUBCOLL\%$ had no correlation before the partialling, but the correlation is significant and negative, with $r = -0.375$.

6.2.3.6 Discussion of Collaborativity vs Impact with Productivity Controlled

Before the productivity partialling, $\boldsymbol{CITTOT}$ and $\boldsymbol{CITTOTw}$ were found to strongly correlate with $\boldsymbol{PUBCOLL}$ and $\boldsymbol{PUBCOLLw}$ in all five of the disciplines. With productivity factor controlled, Pharmacology and WoS Computer Science found to have *no correlation* left; ACM Computer Science, Materials Science and Psychology had very *small* correlations, though the correlations were significant. Law on the other hand, had all of the strong positive correlations almost*unchanged* with the productivity controlled.

This means that institutions publishing in Pharmacology and WoS Computer Science, the number of papers they have collaborated were not affected or affecting the total citations they received as an institution, after the productivity has been partialled out. In ACM Computer Science, Materials Science and Psychology, although significant positive correlation were found, the size of the correlation was very small. As a result, the partialled out variable – productivity – is more likely an interesting variable that have strong correlation with each of these two variables.

Law, however, is an exception, with these correlations almost unchanged(still significant and positive) after partialling. So productivity is not a major factor presented in either institutional impact or collaborativity. Law is not a discipline where collaboration is common, only about 1/10 of the papers were collaborated (while other disciplines have at least 1/3). When they do collaborate, the papers receive on average almost double number of citations (see Figure 4.5 and 4.6). So it could be the case that collaboration in law, despite its infrequency, is a very well recognised and cited when they do collaborate, leading to the observed positive correlation between collaboration and citation.

When the collaborative paper is normalised by the size of the institution's output ($PUBCOLL\%$), the correlation turned negative for four disciplines (except Law) in previous chapter. Now, with productivity partialled out from the impact metrics, we revealed a mixed results.

In Law, the higher the institution's $PUBCOLL\%$, the *lower* the impact was ($CITTOT$,$CITTOTw$), and no correlation was found with average paper impact ($CITAV$). Materials Science did not show correlation with ($CITTOT$,$CITTOTw$) at all, while keeping the significant negative correlation with $CITAV$.

Conflicting result was also found when impact was measured by $CITTOT$ or $WRANK$. In ACM Computer Science, the correlation with $CITTOT$ was positive while it was negative with $WRANK$. In Pharmacology, no correlation was found with $CITTOT$, but negative correlation was observed with $WRANK$. These disagreements could be rooted at the two different types of impact measurements. Citation is a measurement based on authors' judgement of a piece of research's relevance and influence in the domain. The citations to the papers approximate the institution's research influence.

On the other hand, Webometrics use the institution's web presence to determine its impact. The visibility, the popularity of the institution's research and website was measured. These differences may have strong impact on the correlation result.

The correlation changes after the productivity partial is quite dramatic. Disciplines change differently and had opposite result with each other. There is also no clear pattern split between SSH and NSE after productivity partial.

6.3 Chapter Summary

Following on the results obtained in the chapter 4, we performed a pairwise linear correlation again, but with the unwanted variables partialled out. We started with interpreting the variable partialling. We gave intuitive meanings to the remaining variables after partialling to make the interpretation of the results easier.

The results of the partial correlation were then presented. We previously found that an institution's number of collaborative papers was highly correlated with the institution's impact metrics (total citations, average citations and web rank), but when the productivity was partialled out, only Law and Psychology still showed the collaborative paper/impact correlation. The other three disciplines – Computer Science, Pharmacology and Materials Science were observed with the collaborative paper/impact correlation reduced or disappeared.

With the impact variable controlled, the pairwise correlation between the collaborativity and productivity still presented in four of the five disciplines, but disappears completely in the case of Law.

The productivity as measured by total paper count was positively correlated with all impact metrics. The more productive an institution, the higher its impact metrics in all the disciplines studied (though the effect was weaker in Law). With the collaboration variables controlled, the total paper count still significantly correlates with total citation and average citations in 4 out of 5 disciplines, but 3 out of 5 are positive with the impact measured by rank and 2 out of 5 with impact measured by weighted citation.

Institutional collaborativity, as measured by percentage of collaborative papers, has shown disciplinary differences after the partialling. Law, which had weak correlations between collaboration metrics and productivity metrics before partialling out productivity, becomes the only discipline that has a very large correlations. Apart from Law, only Psychology and Computer Science still have sizable positive correlation. Percent collaboration, which was all negatively correlated with impact metrics, didn't appear to be related to citation metrics (citation, weighted citation and average citation) for disciplines except Law once the productivity is partialled out. But the ranking of the institutions is showing negative correlations in disciplines, same as before.

We have also noticed that the number of authors involved in each individual collaboration has little affect with institutional productivity, nor the impact metrics.

Chapter 7

Conclusions

This thesis examined the empirical evidence on the interrelations among research productivity, impact and collaborativity.

There were a number of reasons to conduct this research project. For example, institutions experience substantial competition for the best researchers, research fundings and reputation. Policy makers have assumed that to publish more research and to publish collaboratively should have a direct influence on the impact of the research output. Funders such as JISC and European Framework Programme have placed increasing emphasis on collaborative research, their funding calls often include collaboration as one of the requirements. Yet there was little large scale empirical evidence on the impact of collaborative research.

In this thesis, I explored inter-institutional collaboration and its effect on the impact and the productivity of institutions. I first classified nearly half a century's worth of article publication data across two datasets – ACM and Web of Science – in terms of their authors' institutions. I then used attributes of these articles as metrics for the institutions' productivity, impact and collaborativity:

- institutional productivity – measured by publications count.
- institutional impact – measured by citation count, Pagerank weighted citations and average citations per paper.

- institutional collaborativity – measured by collaborative paper count, size-weighted collaboration and percent collaboration.

In addition to these derived publication-based impact metrics, I used a published world university ranking as a further measure of institutional quality (Webometrics ranking – July 2010 version).

Pairwise linear correlations were calculated among the three factors for institutions across the globe. All three were highly intercorrelated positively: Institutions with high paper productivity were also highly collaborative and of high impact. At the same time, these results represented a circular correlation, making it difficult to interpret the result. Partial correlation was used to control the third variable and calculate the independent correlation between the remaining two variables. The results were compared with previous studies as well as across disciplines. A network visualisation was also used to visualise the relationships among the variables.

7.1 Overview of Research Findings

7.1.1 Measuring institutional collaborativity, impact and productivity

There is no agreed standard measure of research collaboration, impact or productivity in the literature. The first objective was to devise a reliable, reproducible measure for these research activities at the institution level.

7.1.1.1 Research collaborativity

Collaboration is a complex human interaction, and depending on the interpretation, sometime the entire research community can be counted as one collaboration, which making measuring collaboration challenging. A phenomenon in research publication – co-authorship – has been studied intensively in the literature [20, 159] and in more recent years, it has been used as a proxy to collaboration[15, 118]. The use of co-authorship was mainly for studying interindividual collaborativity, but the present study builds on them and uses interinstitutional co-authorships as measures of interinstitutional collaboration. Three measures of collaborativity based on co-authorship were used for correlational analysis in this study.

7.1.1.2 Institutional impact

The quality of an institution is difficult to measure due to its multidimensional nature. An institution's quality is an aggregation of its facilities, infrastructures, resources, research projects, research outcomes, researchers and so on, but such data is inaccessible at the global scale. To obtain a reproducible quality measurement, impact based on bibliometric as well as webometrics methods was considered. Citation as a paper impact measure has been debated for its validity, but it was also the earliest, widely used and generally carries more authority than alternative metrics. The citation was adapted to use as an impact measure in this study. In addition, a university ranking based on webometrics measures were also incorporated in this study as the forth quality measurements, providing an second perspective of institutional quality.

7.1.1.3 Research productivity

Research productivity – as a ratio of input and output – is also problematic to measure and quantify due to lack of data. The input (funding, facility, equipment, the number of researchers, researcher skill level) of the institution, which resulted the research output (publication and awards) maybe absent: some are not collected and not obtainable (*e.g.* the number of researchers, researcher skill level), some are not released by the authority (*e.g.* the amount of funding). These factors can impact strongly on the research output of any institution. To measure productivity as a ratio of output over input, although ideal, is unrealistic in large scale study such as this one.

Previous studies[33, 55] used publication counts as an estimate of productivity; publication counts have also been used as the main metrics in recent UK and Australian research assessment exercises. We have accordingly adopted publication counts in this study for measuring institutional productivity.

Although many efforts have been taken into account while interpreting the results, It should be noticed that different discipline put different emphasis on various types of publication (e.g. journal articles, monographs). A biased dataset can leave certain disciplines disadvantaged in analysis.

7.1.2 Institutional collaboration percentage

The world's most productive institutions publish about half of their papers independently, without any co-authors from other institutions. This is true in natural sciences and engineering disciplines as well as social science disciplines we have tested (except Law). For example, the top institutions from Computer Science published as many as 80% of their papers independently. On the other hand, for many of the least productive institutions – those that only publish a few papers each year – the majority of their papers were found to be collaborative. With this observation, an institutional policy that encourages researchers to publish interinstitutional papers instead of single-institution papers is not a good strategy for increasing the institutional productivity. There can be many reasons why some institutions are less productive, such as that the institution's focus is on teaching; or there is not enough funding for researches; or there is no equipment to undertake research. Institution would do better to identify the specific cause for their low productivity, rather than assuming that collaboration is the driver for productivity.

7.1.3 Collaboration structure

Previous research has shown that country collaborations form core-peripheral structures[46], where highly collaborative countries collaborate mostly with one another, but rarely with other countries. We did not find such core-peripheral structures at the institution level. Those institutions that published many collaborative papers have proportionally higher numbers of collaborating institutions. This reveals that the collaborative papers were not a result of frequent collaborations between a few institutions, however. Rather, they were the result of collaboration with a wide range of institutions. The average collaborative papers per collaborator at the institution level varies slightly across disciplines. Computer Science has the highest ratio of 1.8 papers per collaborator while Pharmacology has the lowest of ratio of 1.3 papers per collaborator.

7.1.4 Institutional impact and productivity

We found that institutional research impact and institutional research productivity correlate positively with each other, which contradicts previous findings on patent impact and productivity. (No direct work was conducted on the relationship between an institution's productivity and impact). At the institution level, with collaboration factors partialled out, the average citations received by each paper correlate positively with the number of papers the institutions published. With this result, it seems that the more papers the institutions publishes, the higher impact each paper gets. However, this does not necessarily imply that publishing more having caused the higher citation counts. It may be more plausible that researchers who publish the higher impact papers publish more frequently, which in turn leads to the institution level effect.

7.1.5 Individual impact and productivity

The investigation of paper impact and paper productivity at the individual researcher level revealed a disciplinary difference. In ACM Computer Science and Psychology, the more papers the researchers published, the more cited the papers were; whereas in Materials Science, Pharmacology and WoS Computer Science the opposite was found. No such relationship was found in Law.

We further split the authors into three productivity tiers: high, medium and low for Materials Science, Pharmacology, WoS Computer Science and Law that showed inconsistency between institution level and individual level effects. In the lower productivity author groups in these disciplines, we found positive correlations between productivity and impact in all these disciplines. In contrast, in the high productivity author groups, a negative correlations was found in Pharmacology and WoS Computer Science. This indicates an interesting phenomenon: for a less productive author, each extra paper attracts a more-than-average number of citations while for a more productive author, the extra paper attracts a less-than-average number of citations. (The former is true in all of the disciplines investigated while the latter is true in Pharmacology and WoS Computer Science). For less productive authors, (e.g. those that publish only a few papers each year) publishing a few extra papers can not only give them the potential extra citations to these papers, but an increased number of citation to all of their papers. Highly productive authors in Pharmacology and WoS Computer Science, on the other hand, may see their average citations per paper reducing when they publish extra papers. It almost appears as if they have attracted all the citations possible, and extra papers would not attract any more, but only divides their existing citations. This individual productivity and impact correlation in fact gives evidence to a seemingly known fact: most researchers at any institution publish very little, while only a small proportion of them publish heavily.

7.1.6 Paper count

Out of all the variables covering three factors investigated in this study, raw paper count carried the most information. Partialling paper count from the other two produces the largest change in their correlation.

In Computer Science, Pharmacology, Materials Science and Psychology, little or no correlation was found between collaborativity and institutional impact after partialling out productivity, although the correlations had been significantly positive before partialling. No correlation means that papers published by the highly collaborative institutions are not cited more frequently. Law, on the other hand, was found to have strong, positive correlations, almost as high as before the partialling. The correlation between collaborativity and institutional impact was very weakly affected by productivity in Law, but productivity carried the most information regarding impact and collaborativity in all other disciplines. Despite the current climate favouring high collaborativity, our analyses failed to detect strong associated improvements in institutions' research profiles.

7.1.7 Collaboration visualisation

Finally, a case study of applying social network analysis to institutional collaboration was successfully conducted. Several graphs were plotted to visualise the patterns of institutional collaboration. The relationship between the impact and collaboration can be clearly observed in the graph (Figure 4.15) for the ACM dataset. An interesting observation is that the erroneous data can be picked out by eye immediately, which can then feed back to correct the source data. Research collaboration as seen in ACM Computer Science was dominated by US, where almost every corner of the graph has a US institution presented. Institutions from Korea, Japan and Brazil also showed a homophily effect, where many institutions from these countries collaborate almost exclusively with institutions coming from the same country. On the other hand, EU countries do not exhibit homophily; institutions from UK, Spain, France and Germany have been shown to collaborate with each other. Perhaps this is due to the EU funding bodies, which explicitly encourage collaborations between the member countries. Despite the homophily effect, the majority of institutions are connected to each other within very short steps, so knowledge is channelled through these institutions quickly and efficiently.

7.2 Answers to Research Questions

We now answer the research questions posed in chapter 3. All questions were based on our central question:

> *What are the relationships, at the institution aggregation level, among collaboration, productivity and impact?*

This question addresses many aspects, to answer it more clearly, we split into the sub-questions below.

7.2.1 Relationships between collaborativity and impact

> *Are higher impact institutions more collaborative?*

Without removing the productivity effects, institutions that published more collaborative papers had higher impact metrics in terms of total citations, weighted citations and Webometrics rank in all disciplines studied. However, this relationship was found to be strongly affected by the productivity of the institutions and the effect was different in various disciplines. With productivity controlled, Pharmacology and WoS Computer Science were found to have *no correlation* between impact and collaborativity; ACM Computer Science, Materials Science and Psychology had very *small correlations*, though the correlation was still significant. Law, on the other hand, retained a similar positive correlation coefficient before and after partialling out productivity.

If we use institutions' average citations per paper as impact indicator, institutions which published more collaborative papers also received more average citations to each of their published papers. Further analysis found that citations and number of collaborative papers were found to be affected directly by institution's productivity for certain disciplines. With productivity partialled out, the correlation disappeared in Pharmacology and became negative in ACM Computer Science. That is, in Pharmacology, the institutions that published high number of collaborative paper did not publish more highly cited papers, while in Computer Science as indexed by ACM, institutions publishing more collaborative papers received on average fewer citations per paper. For the rest of the investigated disciplines – Materials Science, Psychology, Law and WoS Computer Science, a significant positive correlation was found, yielding a positive answer to this question.

Do higher impact institutions emphasize on collaborative research?

We have assumed that an institution's emphasis on collaborative research was reflected by the proportion of its total papers that were collaborative (percent collaboration).

The answer is no for institutions' publication in Computer science, Materials Science, Pharmacology and Psychology, while yes in Law.

Although the percent collaboration showed a negative correlation with citations, weighted citations and Webometrics rank in all disciplines (except Law), with productivity effects partialled out, the same correlation disappeared in WoS Computer Science, Materials Science, Pharmacology and Psychology, while turned strongly negative in Law.

7.2.2 Relationships between collaborativity and productivity

> *Do institutions that publish more papers also collaborate more? (and emphasize on collaborative research?)*

Except for Law, the answer to the former question is positive, and the answer to the latter question is negative for all the disciplines studied here. Paper counts were found to have high correlation with collaborative paper counts, but a negative correlation with percent collaboration, both before and after partialling out impact. Paper counts in Law showed a positive correlation with collaborative paper and a negative correlation with the percent collaboration before partialling out impact, but none the effect of impact was removed.

7.2.3 Relationships between productivity and impact

> *Do institutions that publish large number of papers have higher impact? Are there disciplinary differences?*

Using the publication data without partialling out collaboration, all institutions showed large correlations between published papers and their citations as well as ranking, which answers the question affirmatively. However, this relationship was also affected by the collaborativity of the institutions, especially in Materials Science. With the effect of collaborativity partialled out, the size of this relationship reduced across disciplines, with Materials Science no longer significant. The relationship between the number of papers institutions published and the institution's impact was discipline and collaborativity dependent. If we ignored the effect of collaborativity (*i.e.*, without partialling), this relationship seemed to be positive in all disciplines; but with the collaborativity effect removed, only Law, Psychology, Pharmacology and Computer Science still showed a positive relationship.

> *Are papers published by high paper-output institutions cited more often than papers published by low paper-output institutions?*

Both before and after partialling out collaborativity, disciplines including Law, Psychology, Pharmacology, Materials Science and ACM Computer Science showed significant and positive correlations. This means that for institutions published more papers, their papers' average citation counts were also higher in these disciplines. It should be noted that publishing more papers would not cause more citations and vice versa. In fact, it was more plausible that authors who published highly cited papers publishing more frequently (instead of the other way round). WoS Computer Science did not show this relationship, either positively or negatively.

7.3 Implications of the Findings for Universities and Scientists

A number of insights regarding publication, citation and collaboration arose from the findings in this study, but it must be recalled that these findings were based on analyses of Web of Science and ACM data for only five disciplines. Before these results can be generalised to scientific research as a whole, more disciplines would have to be examined. With this in mind, we looked at some possible implications.

7.3.1 Collaborativity does not enhance productivity

Collaborativity correlated positively with productivity, but it did not cause productivity to increase. Research collaboration can be encouraged by institutions, but not if the objective was higher productivity. In fact, to do more collaborative research in place of less non-collaborative research did not improve productivity at all. Institutions that publish mostly on collaborative work were lower in productivity as well as impact. Institutions pressuring scientists to collaborate in the hope of improving their research profile should look at other factors, such as equipment, facilities and resources, which may prove to be more effective.

7.3.2 Discipline differences

Different disciplines have different research cultures. Institutions should develop discipline-specific strategies to improve their activities in research. In social science disciplines, the institutions leading in research (either published the most papers or been cited the most) were found to publish many papers collaboratively too. On the other hand, in engineering and natural science disciplines, leading research institutions were found to publish many papers non-collaboratively.

Although we were unable to confirm whether the collaborativity was the cause here, but never the less, it can be used as an indicator. So if institutions with low research activities find that most of their engineering and natural science papers are collaborative, they should investigate why their own researchers collaborated for those publications and what prevent them to publish independently.

7.3.3 More is not always better

Our result have shown that for top productive scientists (top third-tier) in Pharmacology and Computer Science, any extra papers they publish would receive below average citations than their existing papers. As a result, publishing further papers would in fact make their average citations *lower*. For the top scientists in these two disciplines, publishing more is not always better.

On the other hand, we do want to stress that for all the disciplines we studied, for low productive scientists (bottom third-tier), publishing extra paper gave above average citations (average citations for their existing papers), hence making their average citations *higher*. From a different perspective, low productive scientists were creating new audiences for each additional publication, perhaps due to the slightly increased visibility with the additional paper; while top productive scientists may not necessarily attract enough new audiences to match up their prestige with the additional paper they publish.

Scientists publishing in the range of below 5 papers per year should try their best to publish more. It is well worth putting in the extra effort.

7.4 Limitations and recommendations for future studies

This study was limited on the selection of disciplines. There were only two representing social science and humanity and three representing engineering and nature science. To have a comprehensive view of the disciplinary differences, more disciplines should to be studied and compared. When more disciplines involved, it is also a good idea to use only a sample the entire discipline's data to reduce the amount of data to be processed. In addition, the sampling gives the possibility of eye balling the data to remove non-sense item that may be presented in the source data.

This study used as much as available years of data as possible. However, from several aspects it was not the best decision made. Firstly, the amount of data to be processed is massive, it took significant effort to clean the data and the process may introduce programming error. 37 years is long and research paradigm shift may have occurred during this period for certain disciplines. Mash the entire period together would obscure the result and does not help comparison across disciplines. In addition, windowed citation should also be used instead of the citation to date used by this study. Citation to date gives bias towards older publication and may potentially skew the correlation.

A longitudinal analysis can be performed using the same dataset to understand how the relationship of the three variables evolved over time at the institution level; and are there highs and lows of these relationships across disciplines, or whether the relationships are stable over the time.

Bibliography

[1] Giovanni Abramo, Ciriaco Andrea DAngelo, and Flavia Di Costa. Research collaboration and productivity: is there correlation? *Higher Education*, 57(2):155–171, 2009.

[2] James D Adams, Grant C Black, J Roger Clemmons, and Paula E Stephan. Scientific teams and institutional collaborations: Evidence from us universities, 1981–1999. *Research Policy*, 34(3):259–285, 2005.

[3] Robert Adler, John Ewing, and Peter Taylor. Citation statistics. *Statistical Science*, 24(1):1, 2009.

[4] Anurag A Agrawal. Corruption of journal impact factors. *TRENDS in Ecology and Evolution*, 20(4):157, 2005.

[5] I.F. Aguillo, J.L. Ortega, and M. Fernandez. Webometric ranking of world universities: Introduction, methodology, and future developments. *Higher education in Europe*, 33(2):233–244, 2008.

[6] DagW. Aksnes. A macro study of self-citation. *Scientometrics*, 56:235–246, 2003. ISSN 0138-9130. doi: 10.1023/A:1021919228368. URL `http://dx.doi.org/10.1023/A%3A1021919228368`.

[7] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74(1):47–97, Jan 2002. doi: 10.1103/RevModPhys.74.47.

[8] Juan A. Almendral, J.G. Oliveira, L. Lpez, J.F.F. Mendes, and Miguel A.F. Sanjun. The network of scientific collaborations within the european framework programme. *Physica A: Statistical Mechanics and its Applications*, 384

(2):675 – 683, 2007. ISSN 0378-4371. doi: 10.1016/j.physa.2007.05.049. URL `http://www.sciencedirect.com/science/article/pii/S0378437107006073`.

[9] LAN Amaral, A. Scala, M. Barthelemy, and HE Stanley. Classes of small-world networks. *Proceedings of the National Academy of Sciences of the United States of America*, 97(21):11149, 2000.

[10] ARC. Australian research council, 2010. URL `http://www.arc.gov.au/era/era_journal_list.htm[Accessed2/9/2010]`.

[11] Douglas N Arnold and Kristine K Fowler. Nefarious numbers. *Notices of the AMS*, 58(3):434–437, 2011.

[12] Necmi K Avkiran. Scientific collaboration in finance does not lead to better quality research. *Scientometrics*, 39(2):173–184, 1997.

[13] David Bakewell. Publish in english, or perish? *Nature*, 356:648, 1992.

[14] Stephane Baldi and Lowell L Hargens. Reassessing the n-rays reference network: The role of self citations and negative citations. *Scientometrics*, 34(2):239–253, 1995.

[15] A. L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, 311(3-4):590 – 614, 2002. ISSN 0378-4371. doi: 10.1016/S0378-4371(02)00736-7.

[16] A.L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509, 1999.

[17] M. Barthélémy and L.A.N. Amaral. Small-world networks: Evidence for a crossover picture. *Physical Review Letters*, 82(15):3180–3183, 1999.

[18] Alan E Bayer and John Folger. Some correlates of a citation measure of productivity in science. *Sociology of education*, pages 381–390, 1966.

[19] D.B. Beaver. Does collaborative research have greater epistemic authority? *Scientometrics*, 60(3):399–408, 2004.

[20] D.deB Beaver and R. Rosen. Studies in scientific collaboration. *Scientometrics*, 1: 65–84, 1978. ISSN 0138-9130. doi: 10.1007/BF02016840. URL `http://dx.doi.org/10.1007/BF02016840`.

[21] Donald D Bergh, John Perry, et al. Some predictors of smj article impact. *Strategic Management Journal*, 27(1):81–100, 2006.

[22] J.C. Billaut, D. Bouyssou, and P. Vincke. Should you believe in the shanghai ranking? 2009.

[23] J. Bollen, H. Van de Sompel, J.A. Smith, and R. Luce. Toward alternative metrics of journal impact: A comparison of download and citation data. *Information Processing & Management*, 41(6):1419–1440, 2005.

[24] Johan Bollen, Marko A Rodriquez, and Herbert Van de Sompel. Journal status. *Scientometrics*, 69(3):669–687, 2006.

[25] Johan Bollen, Herbert Van de Sompel, and Marko A Rodriguez. Towards usage-based impact metrics: first results from the mesur project. In *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*, pages 231–240. ACM, 2008.

[26] B. Bollobás. The diameter of random graphs. *Transactions of the American Mathematical Society*, 267(1):41–52, 1981.

[27] Susan Bonzi and Herbert W Snyder. Motivations for citation: A comparison of self citation and citation to others. *Scientometrics*, 21(2):245–254, 1991.

[28] María Bordons, Javier Aparicio, and Rodrigo Costas. Trends in the collaborative structure of the spanish pharmacological scientific production and its influence over research impact. In *Proceedings of STI 2012. 17th international conference on science and technology indicators*, volume 1, 2012.

[29] K. Borner, C. Chen, and K.W. Boyack. Visualizing knowledge domains. *Annual review of information science and technology*, 37(1):179–255, 2003.

[30] L. Bornmann and H.-D. Daniel. What do citation counts measure? a review of studies on citing behavior. *Journal of Documentation*, 64(1):45–80, 2008. ISSN 0022-0418.

[31] Lutz Bornmann. Towards an ideal method of measuring research performance: Some comments to the opthof and leydesdorff (2010) paper. *Journal of Informetrics*, 4(3):441–443, 2010.

[32] Lutz Bornmann, Moritz Stefaner, Felix de Moya Anegón, and Rüdiger Mutz. Ranking and mapping of universities and research-focused institutions worldwide based on highly-cited papers: A visualization of results from multi-level models. *arXiv preprint arXiv:1212.0304*, 2012.

[33] Barry Bozeman and Sooho Lee. The impact of research collaboration on scientific productivity. 2003.

[34] Tony Brinn, Michael John Jones, and Maurice Pendlebury. Measuring research quality: peer review 1, citation indices 0. *Omega*, 28(2):237 – 239, 2000. ISSN 0305-0483. doi: 10.1016/S0305-0483(99)00048-1. URL `http://www.sciencedirect.com/science/article/pii/S0305048399000481`.

[35] T. Brody, S. Harnad, and L. Carr. Earlier web usage statistics as predictors of later citation impact. *Journal of the American Society for Information Science and Technology*, 57(8):1060–1072, 2006.

[36] T. Brody, L. Carr, Y. Gingras, C. Hajjem, S. Harnad, and A. Swan. Incentivizing the open access research web: Publication-archiving, data-archiving and scientometrics. *CTWatch Quarterly*, 3(3), 2007.

[37] Alison Callahan, Stephen Hockema, and Gunther Eysenbach. Contextual cocitation: Augmenting cocitation analysis and its applications. *Journal of the American Society for Information Science and Technology*, 61(6):1130–1143, 2010. ISSN 1532-2890. doi: 10.1002/asi.21313.

[38] Juan Miguel Campanario. Empirical study of journal impact factors obtained using the classical two-year citation window versus a five-year citation window. *Scientometrics*, 87(1):189–204, 2011.

[39] P Campbell. Not-so-deep impact. *Nature*, 435(77045):1003–1004, 2005.

[40] AE Cawkell. Citations, obsolescence, enduring articles, and multiple authorships. *Journal of Documentation*, 32(1), 1976.

[41] C. Chen. Visualising semantic spaces and author co-citation networks in digital libraries. *Information Processing and Management*, 35(401):420, 1999.

[42] C. Chen and L. Carr. Trailblazing the literature of hypertext: author co-citation analysis (1989–1998). In *Proceedings of the tenth ACM Conference on Hypertext and hypermedia: returning to our diverse roots: returning to our diverse roots*, pages 51–60. ACM New York, NY, USA, 1999.

[43] Chaomei Chen. Generalised similarity analysis and pathfinder network scaling. *Interacting with Computers*, 10(2):107 – 128, 1998. ISSN 0953-5438. doi: DOI: 10.1016/S0953-5438(98)00015-0.

[44] Chaomei Chen. *Mapping scientific frontiers: The quest for knowledge visualization*. Springer Verlag, 2003.

[45] Chaomei Chen and Steven Morris. Visualizing evolving networks: Minimum spanning trees versus pathfinder networks. *Information Visualization, IEEE Symposium on*, 0:9, 2003.

[46] Sujin Choi. Core-periphery, new clusters, or rising stars?: international scientific collaboration among advanced countries in the era of globalization. *Scientometrics*, 90:25–41, 2012. ISSN 0138-9130. URL `http://dx.doi.org/10.1007/s11192-011-0509-4`. 10.1007/s11192-011-0509-4.

[47] F. Chung and L. Lu. Connected components in random graphs with given expected degree sequences. *Annals of combinatorics*, 6(2):125–145, 2002.

[48] F. Chung and L. Lu. The average distance in a random graph with given expected degrees. *Internet Mathematics*, 1(1):91–113, 2004.

[49] J.R. Cole. A short history of the use of citations as a measure of the impact of scientific and scholarly work. *The Web of Knowledge: A Festschri in Honor of Eugene Garfield*, pages 281–300, 2000.

[50] B. Cronin and L. Meho. Using the h-index to rank influential information scientistss. *Journal of the American Society for Information Science and Technology*, 57(9):1275–1278, 2006.

[51] Blaise Cronin and Kara Overfelt. Citation-based auditing of academic performance. *Journal of the American Society for Information Science*, 45(2):61–72, 1994.

[52] J. Davidson Frame and M.P. Carpenter. International research collaboration. *Social Studies of Science*, 9(4):481–497, 1979.

[53] R. De Castro and J.W. Grossman. Famous trails to paul erdős. *The Mathematical Intelligencer*, 21(3):51–53, 1999.

[54] M.A. De Menezes, C. Moukarzel, and TJP Penna. First-order transition in small-world networks. *Arxiv preprint cond-mat/9903426*, 1999.

[55] Daniela Defazio, Andy Lockett, and Mike Wright. Funding incentives, collaborative dynamics and scientific productivity: Evidence from the eu framework program. *Research Policy*, 38(2):293 – 305, 2009. ISSN 0048-7333. doi: 10.1016/j.respol.2008.11.008. URL `http://www.sciencedirect.com/science/article/pii/S0048733308002709`.

[56] Fereshteh Didegah and Mike Thelwall. Which factors help authors produce the highest impact research? collaboration, journal and document properties. *Journal of Informetrics*, 7(4):861 – 873, 2013. ISSN 1751-1577. doi: http://dx.doi.org/10.1016/j.joi.2013.08.006. URL `http://www.sciencedirect.com/science/article/pii/S1751157713000709`.

[57] Fereshteh Didegah and Mike Thelwall. Determinants of research citation impact in nanoscience and nanotechnology. *Journal of the American Society for Information Science and Technology*, 64(5):1055–1064, 2013.

[58] SN Dorogovtsev, JFF Mendes, and AN Samukhin. Metric structure of random networks. *Nuclear Physics B*, 653(3):307–338, 2003.

[59] David Easley and Jon Kleinberg. The small-world phenomenon. 2007.

[60] John P. Eaton, James C. Ward, Ajith Kumar, and Peter H. Reingen. Structural analysis of co-author relationships and author productivity in selected outlets for consumer behavior research. *Journal of Consumer Psychology*, 8(1):39 – 59, 1999. ISSN 1057-7408. doi: http://dx.doi.org/10.1207/s15327663jcp0801_02. URL http://www.sciencedirect.com/science/article/pii/S1057740899703438.

[61] Leo Egghe, Ronald Rousseau, and Guido Van Hooydonk. Methods for accrediting publications to authors or countries: Consequences for evaluation studies. *Journal of the American Society for Information Science*, 51(2):145–157, 2000.

[62] Ergin Elmacioglu and Dongwon Lee. On six degrees of separation in dblp-db and more. *SIGMOD Rec.*, 34(2):33–40, 2005. ISSN 0163-5808. doi: http://doi.acm.org/10.1145/1083784.1083791.

[63] P. Erdős and A. Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5:17–61, 1960.

[64] P. Erdős and A. Rényi. On the strength of connectedness of a random graph. *Acta Mathematica Hungarica*, 12(1):261–267, 1961.

[65] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication*, page 262. ACM, 1999.

[66] A Fassoulaki, A Paraskeva, K Papilas, and G Karabinis. Self-citations in six anaesthesia journals and their significance in determining the impact factor. *British Journal of Anaesthesia*, 84(2):266–269, 2000.

[67] R.V. Florian. Irreproducibility of the results of the shanghai academic ranking of world universities. *Scientometrics*, 72(1):25–32, 2007.

[68] JamesH. Fowler and DagW. Aksnes. Does self-citation pay? *Scientometrics*, 72: 427–437, 2007. ISSN 0138-9130. doi: 10.1007/s11192-007-1777-2. URL http://dx.doi.org/10.1007/s11192-007-1777-2.

[69] Mary Frank Fox. Publication productivity among scientists: A critical review. *Social Studies of Science*, 13(2):285–305, 1983.

[70] Mary Frank Fox. Research, teaching, and publication productivity: Mutuality versus competition in academia. *Sociology of education*, pages 293–305, 1992.

[71] J.D. Frame. Quantitative indicators for evaluation of basic research programs/projects. *Engineering Management, IEEE Transactions on*, EM-30(3):106–112, 1983. ISSN 0018-9391. doi: 10.1109/TEM.1983.6448601.

[72] Massimo Franceschet and Antonio Costantini. The effect of scholar collaboration on impact and quality of academic papers. *Journal of informetrics*, 4(4):540–553, 2010.

[73] O. Frank and D. Strauss. Markov graphs. *Journal of the American Statistical Association*, 81(395):832–842, 1986.

[74] A. Fronczak, P. Fronczak, and J.A. Holyst. Exact solution for average path length in random graphs. *Arxiv preprint cond-mat/0212230*, 2002.

[75] Emili García-Berthou and Carles Alcaraz. Incongruence between test statistics and p values in medical papers. *BMC Medical Research Methodology*, 4(1):13, 2004.

[76] Eugene Garfield. Citation indexing for studying science. 1970.

[77] Eugene Garfield. Citation analysis as a tool in journal evaluation. *Science*, 178 (4060):471–479, 1972.

[78] Jerry Gaston. The reward system in british science. *American Sociological Review*, pages 718–732, 1970.

[79] Marianne Gauffriau, PederOlesen Larsen, Isabelle Maye, Anne Roulin-Perriard, and Markus Ins. Comparisons of results of publication counting using different methods. *Scientometrics*, 77(1):147–176, 2008. ISSN 0138-9130. doi: 10.1007/s11192-007-1934-2. URL `http://dx.doi.org/10.1007/s11192-007-1934-2`.

[80] Ali Gazni and Fereshteh Didegah. Investigating different types of research collaboration and citation impact: a case study of harvard universitys publications. *Scientometrics*, 87(2):251–265, 2011. ISSN 0138-9130. doi: 10.1007/s11192-011-0343-8. URL `http://dx.doi.org/10.1007/s11192-011-0343-8`.

[81] Ali Gazni, Cassidy R. Sugimoto, and Fereshteh Didegah. Mapping world scientific collaboration: Authors, institutions, and countries. *Journal of the American Society for Information Science and Technology*, 63(2):323–335, 2012. ISSN 1532-2890. doi: 10.1002/asi.21688. URL `http://dx.doi.org/10.1002/asi.21688`.

[82] Russell Gersten, Scott Baker, and John Wills Lloyd. Designing high-quality research in special education group experimental design. *The Journal of Special Education*, 34(1):2–18, 2000.

[83] Yves Gingras and Vincent Larivière. There are neither king nor crown in scientometrics: Comments on a supposed alternative method of normalization. *Journal of Informetrics*, 5(1):226–227, 2011.

[84] M. Girvan and MEJ Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(12):7821, 2002.

[85] W. Glänzel and A. Schubert. Analyzing scientific networks through co-authorship. 2004.

[86] Wolfgang Glänzel and Urs Schoepflin. A bibliometric study of reference literature in the sciences and social sciences. *Information processing & management*, 35(1): 31–44, 1999.

[87] Shaun Goldfinch, Tony Dale, and Jr. DeRouen, Karl. Science from the periphery: Collaboration, networks and 'periphery effects' in the citation of new zealand crown research institutes articles, 1995-2000. *Scientometrics*, 57(3): 321–337, 2003. ISSN 0138-9130. doi: 10.1023/A:1025048516769. URL `http://dx.doi.org/10.1023/A%3A1025048516769`.

[88] Trisha Greenhalgh. Assessing the methodological quality of published papers. *BMJ: British Medical Journal*, 315(7103):305, 1997.

[89] S. Harnad. Open access scientometrics and the uk research assessment exercise. *Scientometrics*, 79(1):147–156, 2009.

[90] S. Harnad, T. Brody, F. Vallieres, L. Carr, S. Hitchcock, Y. Gingras, C. Oppenheim, H. Stamerjohanns, and E.R. Hilf. The access/impact problem and the green and gold roads to open access. *Serials review*, 30(4):310–314, 2004.

[91] Stevan Harnad. Validating research performance metrics against peer rankings. *Ethics in Science and Environmental Politics*, 8(11), 2008.

[92] Richard L Hart. Collaboration and article quality in the literature of academic librarianship. *The journal of academic librarianship*, 33(2):190–195, 2007.

[93] Zi-Lin He, Xue-Song Geng, and Colin Campbell-Hunt. Research collaboration and research output: A longitudinal study of 65 biomedical scientists in a new zealand university. *Research Policy*, 38(2):306 – 317, 2009. ISSN 0048-7333. doi: 10.1016/j.respol.2008.11.011. URL `http://www.sciencedirect.com/science/article/pii/S0048733308002813`.

[94] A.G. Heffner. Funded research, multiple authorship, and subauthorship collaboration in four disciplines. *Scientometrics*, 3:5–12, 1981. ISSN 0138-9130. doi: 10.1007/BF02021860. URL `http://dx.doi.org/10.1007/BF02021860`.

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

[96] P.W. Holland and S. Leinhardt. An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, 76(373):33–50, 1981.

[97] EdwardJ. Huth. Stealing into print: Fraud, plagiarism, and misconduct in scientific publishing. *Publishing Research Quarterly*, 9(2):78–79, 1993. ISSN 1053-8801. doi: 10.1007/BF02680404. URL `http://dx.doi.org/10.1007/BF02680404`.

[98] Paul Jeffrey. Smoothing the waters: Observations on the process of cross-disciplinary research collaboration. *Social Studies of Science*, 33(4):pp. 539–562, 2003. ISSN 03063127. URL `http://www.jstor.org/stable/3182968`.

[99] In-Su Kang, Seung-Hoon Na, Seungwoo Lee, Hanmin Jung, Pyung Kim, Won-Kyung Sung, and Jong-Hyeok Lee. On co-authorship for author disambiguation.

Information Processing & Management, 45(1):84–97, January 2009. ISSN 0306-4573.

[100] Miray Kas, Kathleen Carley, and L. Carley. Trends in science networks: understanding structures and statistics of scientific networks. *Social Network Analysis and Mining*, pages 1–19, 2012. ISSN 1869-5450. URL http://dx.doi.org/10.1007/s13278-011-0044-6. 10.1007/s13278-011-0044-6.

[101] David A. Katz. Faculty salaries, promotions, and productivity at a large university. *The American Economic Review*, 63(3):pp. 469–477, 1973. ISSN 00028282. URL http://www.jstor.org/stable/1914379.

[102] J. Katz. Geographical proximity and scientific collaboration. *Scientometrics*, 31: 31–43, 1994. ISSN 0138-9130. URL http://dx.doi.org/10.1007/BF02018100. 10.1007/BF02018100.

[103] J. Katz and Diana Hicks. How much is a collaboration worth? a calibrated bibliometric model. *Scientometrics*, 40:541–554, 1997. ISSN 0138-9130. URL http://dx.doi.org/10.1007/BF02459299. 10.1007/BF02459299.

[104] J. Sylvan Katz and Ben R. Martin. What is research collaboration? *Research Policy*, 26(1):1 – 18, 1997. ISSN 0048-7333. doi: DOI:10.1016/S0048-7333(96)00917-1. URL http://www.sciencedirect.com/science/article/B6V77-3SWTPF2-1/2/a048c55fe69c0245af070846dd619a24.

[105] J.S. Katz. *Bibliometric assessment of intranational University-University collaboration*. PhD thesis, University of Sussex, 1992.

[106] Jon Kleinberg. The small-world phenomenon: an algorithm perspective. In *STOC '00: Proceedings of the thirty-second annual ACM symposium on Theory of computing*, pages 163–170, New York, NY, USA, 2000. ACM. ISBN 1-58113-184-4. doi: http://doi.acm.org/10.1145/335305.335325.

[107] S. Kyvik and M. Teigen. Child care, research collaboration, and gender differences in scientific productivity. *Science, Technology & Human Values*, 21(1):54–71, 1996.

[108] Jean O. Lanjouw and Mark Schankerman. Patent quality and research productivity: Measuring innovation with multiple indicators*. *The Economic Journal*, 114(495):441–465, 2004. ISSN 1468-0297. doi: 10.1111/j.1468-0297.2004.00216.x. URL `http://dx.doi.org/10.1111/j.1468-0297.2004.00216.x`.

[109] Vincent Larivière, Éric Archambault, Yves Gingras, and Étienne Vignola-Gagné. The place of serials in referencing practices: Comparing natural sciences and engineering with social sciences and humanities. *Journal of the American Society for Information Science and Technology*, 57(8):997–1004, 2006.

[110] Vincent Larivière, Yves Gingras, Cassidy R Sugimoto, and Andrew Tsou. Team size matters: Collaboration and scientific impact since 1900. *Journal of the Association for Information Science and Technology*, 2014.

[111] G. LaRowe, R. Ichise, and K. Borner. Analysis of japanese information systems co-authorship data. In *2007 11th International Conference on Information Visualization*, pages 433–8, Piscataway, NJ, USA, 2007. IEEE. ISBN 0-7695-2900-3.

[112] STEPHEN M Lawani. Citation analysis and the quality of scientific productivity. *BioScience*, pages 26–31, 1977.

[113] Stephen M. Lawani and Alan E. Bayer. Validity of citation criteria for assessing the influence of scientific publications: New evidence with peer assessment. *Journal of the American Society for Information Science*, 34(1):59–66, 1983. ISSN 1097-4571. doi: 10.1002/asi.4630340109. URL `http://dx.doi.org/10.1002/asi.4630340109`.

[114] Sooho Lee and Barry Bozeman. The impact of research collaboration on scientific productivity. *Social Studies of Science*, 35(5):pp. 673–702, 2005. ISSN 03063127. URL `http://www.jstor.org/stable/25046667`.

[115] Roosa Leimu and Julia Koricheva. Does scientific collaboration increase the impact of ecological articles? *BioScience*, 55(5):438–443, 2005.

[116] Jonathan M Levitt and Mike Thelwall. Citation levels and collaboration within library and information science. *Journal of the American Society for Information Science and Technology*, 60(3):434–442, 2009.

[117] Jonathan M Levitt and Mike Thelwall. Does the higher citation of collaborative research differ from region to region? a case study of economics. *Scientometrics*, 85(1):171–183, 2010.

[118] Xiaoming Liu, Johan Bollen, Michael L. Nelson, and Herbert Van de Sompel. Co-authorship networks in the digital library research community. *Information Processing & Management*, 41(6):1462 – 1480, 2005. ISSN 0306-4573. doi: DOI: 10.1016/j.ipm.2005.03.012. Special Issue on Infometrics.

[119] E.L. Logan and M. Lee Pao. Analytic and empirical measures of key authors in schistosomiasis. In D. Henderson, editor, *ASIS '90. Proceedings of the 53rd ASIS Annual Meeting*, pages 213–19, Medford, NJ, USA, 1990. ASIS, American Soc. Inf. Sci. ISBN 0 938734 48 2.

[120] Kathleen N Lohr. Rating the strength of scientific evidence: relevance for quality improvement programs. *International Journal for Quality in Health Care*, 16(1): 9–18, 2004.

[121] A.J Lotka. The frequency distribution of scientific productivity. *Journal of Washington Academy Sciences*, 16:317–323, 1926.

[122] Terttu Luukkonen, Olle Persson, and Gunnar Sivertsen. Understanding patterns of international scientific collaboration. *Science, Technology, & Human Values*, 17(1):pp. 101–126, 1992. ISSN 01622439. URL `http://www.jstor.org/stable/689852`.

[123] PJ Macdonald, E. Almaas, and A.L. Barabási. Minimum spanning trees of weighted scale-free networks. *EPL (Europhysics Letters)*, 72:308, 2005.

[124] P. Mählck and O. Persson. Socio-bibliometric mapping of intra-departmental networks. *Scientometrics*, 49(1):81–91, 2000.

[125] Gunther Maier. Impact factors and peer judgment: The case of regional science journals. *Scientometrics*, 69(3):651–667, 2006.

[126] Ben R Martin. The use of multiple indicators in the assessment of basic research. *Scientometrics*, 36(3):343–362, 1996.

[127] W.S. Martins, M.A. Gonçalves, A.H.F. Laender, and N. Ziviani. Assessing the quality of scientific conferences based on bibliographic citations. *Scientometrics*, 83(1):133–155, 2010.

[128] Yutaka Matsuo, Junichiro Mori, Masahiro Hamasaki, Takuichi Nishimura, Hideaki Takeda, Koiti Hasida, and Mitsuru Ishizuka. Polyphonet: An advanced social network extraction system from the web. *Web Semantics: Science, Services and Agents on the World Wide Web*, 5(4):262 – 278, 2007. ISSN 1570-8268. doi: DOI:10.1016/j.websem.2007.09.002.

[129] Gran Melin. Pragmatism and self-organization: Research collaboration on the individual level. *Research Policy*, 29(1):31 – 40, 2000. ISSN 0048-7333. doi: 10.1016/S0048-7333(99)00031-1. URL `http://www.sciencedirect.com/science/article/pii/S0048733399000311`.

[130] Henk F Moed. Cwts crown indicator measures citation impact of a research group's publication oeuvre. *arXiv preprint arXiv:1003.5884*, 2010.

[131] H.F. Moed. Bibliometric rankings of world universities. *Centre for Science and Technology Studies report 2006*, 1, 2006.

[132] M. Molloy and B.A. Reed. A critical point for random graphs with a given degree sequence. *Random Structures and Algorithms*, 6(2/3):161–180, 1995.

[133] Richard Monastersky. The number that's devouring science. *The Chronicle*, 52: A12, 2005.

[134] James Moody. The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999. *American Sociological Review*, 69(2):213–238, 2004. ISSN 00031224.

[135] C.F. Moukarzel. Spreading and shortest paths in systems with sparse long-range connections. *Physical Review E*, 60(6):6263–6266, 1999.

[136] G.L. Naber. The geometry of minkowski spacetime. *Washington DC American Geophysical Union Geophysical Monograph Series*, 1, 1992.

[137] F. Narin, K. Stevens, and E.S. Whitlow. Scientific co-operation in europe and the citation of multinationally authored papers. *Scientometrics*, 21(3):313–323, 1991.

[138] Anthony J Nederhof, Rolf A Zwaan, Renger E De Bruin, and PJ Dekker. Assessing the usefulness of bibliometric indicators for the humanities and the social and beha vioural sciences: A comparative study. *Scientometrics*, 15(5):423–435, 1989.

[139] M. E. J. Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2):404, 2001.

[140] M. E. J. Newman. Scientific collaboration networks. i. network construction and fundamental results. *Physical Review E*, 64(1):16131, 2001.

[141] M. E. J. Newman. Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Physical Review E*, 64(1):16132, 2001.

[142] M. E. J. Newman. Clustering and preferential attachment in growing networks. *Phys. Rev. E*, 64(2):025102, Jul 2001. doi: 10.1103/PhysRevE.64.025102.

[143] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167, 2003.

[144] M. E. J. Newman. Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences*, 101(90001):5200–5205, 2004.

[145] M. E. J. Newman and D. J. Watts. Renormalization group analysis of the small-world network model. *Physics Letters A*, 263(4-6):341 – 346, 1999. ISSN 0375-9601. doi: DOI:10.1016/S0375-9601(99)00757-4.

[146] M. E. J. Newman and D. J. Watts. Scaling and percolation in the small-world network model. *Phys. Rev. E*, 60(6):7332–7342, Dec 1999. doi: 10.1103/PhysRevE.60.7332.

[147] Byung-Won On. Social network analysis on name disambiguation and more. In *2008 Third International Conference on Convergence and Hybrid Information Technology (ICCIT)*, volume 2, pages 1081–8, Los Alamitos, CA, USA, 2008. IEEE Computer Society. ISBN 978-0-7695-3407-7.

[148] Tobias Opthof and Loet Leydesdorff. Caveats for the journal and field normalizations in the {CWTS} (leiden) evaluations of research performance. *Journal of Informetrics*, 4(3):423 – 430, 2010. ISSN 1751-1577. doi: http://dx.doi.org/10.1016/j.joi.2010.02.003. URL http://www.sciencedirect.com/science/article/pii/S1751157710000106.

[149] Tobias Opthof and Loet Leydesdorff. A comment to the paper by waltman et al., scientometrics, 87, 467481, 2011. *Scientometrics*, 88(3):1011–1016, 2011. ISSN 0138-9130. doi: 10.1007/s11192-011-0424-8. URL http://dx.doi.org/10.1007/s11192-011-0424-8.

[150] M. Ozana. Incipient spanning cluster on small-world networks. *EPL (Europhysics Letters)*, 55:762, 2001.

[151] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999. URL http://ilpubs.stanford.edu:8090/422/. Previous number = SIDL-WP-1999-0120.

[152] Ray J Paul. Measuring research quality: the united kingdom government's research assessment exercise. *European Journal of Information Systems*, 17(4):324–329, 2008.

[153] O. Persson. All author citations versus first author citations. *Scientometrics*, 50 (2):339–344, 2001.

[154] Olle Persson, Göran Melin, Rickard Danell, and A Kaloudis. Research collaboration at nordic universities. *Scientometrics*, 39(2):209–223, 1997.

[155] N. Pravdić and V. Oluić-Vuković. Dual approach to multiple authorship in the study of collaboration/scientific output relationship. *Scientometrics*, 10(5):259–280, 1986.

[156] S. Presser. Collaboration and the quality of research. *Social Studies of Science*, 10(1):95, 1980.

[157] D.J.S. Price. Networks of scientific papers. *Nuovo Cimento*, 5:199, 1957.

[158] D.J.S. Price. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, 27(5):292–306, 1976.

[159] D.J.S. Price and D. Beaver. Collaboration in an invisible college. *American Psychologist*, 21(11):1011, 1966.

[160] S. Redner. How popular is your paper? an empirical study of the citation distribution. *The European Physical Journal B*, 4(2):131–134, 1998.

[161] THOMSON REUTERS. Thomson reuters master journal list, 2013. URL `http://science.thomsonreuters.com/cgi-bin/jrnlst/jlresults.cgi?PC=MASTER`. Retrived on Aug 2013.

[162] J. Rigby and J. Edler. Peering inside research networks: Some observations on the effect of the intensity of collaboration on the variability of research quality. *Research Policy*, 34(6):784 – 794, 2005. ISSN 0048-7333. doi: 10.1016/j.respol.2005.02.004. URL `http://www.sciencedirect.com/science/article/pii/S0048733305000703`.

[163] Ronald Rousseau. Comments on the modified collaborative coefficient. *Scientometrics*, 87(1):171–174, 2011. ISSN 0138-9130. doi: 10.1007/s11192-010-0300-y. URL `http://dx.doi.org/10.1007/s11192-010-0300-y`.

[164] G. Rueda, P. Gerdsri, and D.F. Kocaoglu. Bibliometrics and social network analysis of the nanotechnology field. In *Management of Engineering and Technology, Portland International Center for*, pages 2905–2911. IEEE, 2007.

[165] R. Santama and R. Theron. Overlapping clustered graphs: co-authorship networks visualization. In *Smart Graphics. 9th International Symposium, SG 2008*, pages 190–9, Berlin, Germany, 2008 2008. Springer-Verlag. ISBN 978-3-540-85410-4. Smart Graphics. 9th International Symposium, SG 2008, 27-29 August 2008, Rennes, France.

[166] Michael Schreiber. Self-citation corrections for the hirsch index. *EPL (Europhysics Letters)*, 78(3):30002, 2007.

[167] András Schubert and T Braun. International collaboration in the sciences 1981–1985. *Scientometrics*, 19(1):3–10, 1990.

[168] R.W. Schvaneveldt. *Pathfinder associative networks: Studies in knowledge organization.* Ablex Publishing, 1990.

[169] R.W. Schvaneveldt, F.T. Durso, and D.W. Dearholt. Network structures in proximity data. *The psychology of learning and motivation: Advances in research and theory*, 24:249–284, 1989.

[170] Monica Sharma and Shalini R. Urs. Network dynamics of scholarship: a social network analysis of digital library community. In *Proceeding of the 2nd PhD workshop on Information and knowledge management*, pages 101–104, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-257-3. doi: 10.1145/1458550.1458570.

[171] R.J. Silverman. Higher education as a maturing field? evidence from referencing practices. *Research in Higher Education*, 23(2):150–183, 1985.

[172] M. Smith. The trend toward multiple authorship in psychology. *American Psychologist*, 13(10):596, 1958.

[173] Radhamany Sooryamoorthy. Do types of collaboration change citation? collaboration and citation patterns of south african science publications. *Scientometrics*, 81(1):177–193, 2009.

[174] M.E. Soper. Characteristics and use of personal collections. *The Library Quarterly*, 46(4):397–415, 1976.

[175] Maria Souto, Mariusa Warpechowski, and José de Oliveira. An ontological approach for the quality assessment of computer science conferences. *Advances in Conceptual Modeling–Foundations and Applications*, pages 202–212, 2007.

[176] D. Strauss. On a general class of models for interaction. *SIAM review*, 28(4):513–527, 1986.

[177] Y.M. Su, S.C. Yang, P.Y. Hsu, and W.L. Shiau. Extending co-citation analysis to discover authors with multiple expertise. *Expert Systems with Applications*, 36(3):4287–4295, 2009.

[178] K. Subramanyam. Bibliometric studies of research collaboration: A review. *Journal of information Science*, 6(1):33–38, 1983.

[179] James Edwin Swiss. *Public management systems: Monitoring and managing government performance.* Simon & Schuster Trade, 1990.

[180] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. Arnetminer: extraction and mining of academic social networks. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 990–998, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-193-4.

[181] Jim Taylor. The assessment of research quality in uk universities: Peer review or metrics? *British Journal of Management*, 22(2):202–217, 2011. ISSN 1467-8551. doi: 10.1111/j.1467-8551.2010.00722.x. URL `http://dx.doi.org/10.1111/j.1467-8551.2010.00722.x`.

[182] NWB Team. Network workbench tool, 2006. URL `http://nwb.slis.indiana.edu`. Indiana University, Northeastern University, and University of Michigan.

[183] Robert Tijssen, Martijn Visser, and Thed van Leeuwen. Benchmarking international scientific excellence: Are highly cited research papers an appropriate frame of reference? *Scientometrics*, 54:381–397, 2002. ISSN 0138-9130. URL `http://dx.doi.org/10.1023/A:1016082432660`. 10.1023/A:1016082432660.

[184] J. Travers and S. Milgram. An experimental study of the small world problem. *Sociometry*, pages 425–443, 1969.

[185] A.F.J. van Raan. Measuring science. capita selecta of current main issues. *Handbook of quantitative science and technology research*, pages 19–50, 2004.

[186] A.F.J. van Raan. Challenges in ranking of universities. *Invited paper for the First International*, 2005.

[187] A.F.J. van Raan. Comparison of the hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups. *Scientometrics*, 67(3):491–502, 2006.

[188] A.F.J. van Raan. Bibliometric statistical properties of the 100 largest european research universities: Prevalent scaling rules in the science system. *Journal of the American Society for Information Science and Technology*, 59(3):461–475, 2008.

[189] T. van Raan, T. van Leeuwen, and M. Visser. Non-english papers decrease rankings. *Nature*, 469(7328):34–34, 2011.

[190] L. Waltman, N.J. van Eck, T.N. van Leeuwen, M.S. Visser, and A.F.J. van Raan. Towards a new crown indicator: An empirical analysis. *Scientometrics*, 87(3): 467–481, 2011.

[191] Ludo Waltman, Nees van Eck, Thed van Leeuwen, Martijn Visser, and Anthony van Raan. On the correlation between bibliometric indicators and peer review: reply to opthof and leydesdorff. *Scientometrics*, pages 1–6, 2011. ISSN 0138-9130. URL `http://dx.doi.org/10.1007/s11192-011-0425-7`. 10.1007/s11192-011-0425-7.

[192] Jian Wang. Citation time window choice for research impact evaluation. *Scientometrics*, 94:851–872, 2013. ISSN 0138-9130. doi: 10.1007/s11192-012-0775-9. URL `http://dx.doi.org/10.1007/s11192-012-0775-9`.

[193] D.J. Watts. Networks, dynamics, and the small-world phenomenon 1. *American Journal of Sociology*, 105(2):493–527, 1999.

[194] D.J. Watts and S.H. Strogatz. Collective dynamics of small-world networks. *Nature*, page 301, 1998.

[195] H.D. White. Authors as citers over time. *Journal of the American Society for Information Science and Technology*, 52(2):87–108, 2001.

[196] H.D. White and B.C. Griffith. Author cocitation: A literature measure of intellectual structure. *Journal of the American Society for Information Science*, 32(3): 163–171, 1981.

[197] H.D. White and K.W. McCain. Visualizing a discipline: An author co-citation analysis of information science, 1972-1995. *Journal of the American Society for Information Science*, 49(4):327–355, 1998.

[198] S. Woolgar. Beyond the citation debate: Towards a sociology of measurement technologies and their use in science policy. *Science and Public Policy*, 18(5): 319–326, 1991.

[199] Su Yan and Dongwon Lee. Toward alternative measures for ranking venues: a case of database research community. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pages 235–244, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-644-8. doi: 10.1145/1255175.1255221. URL `http://portal.acm.org/citation.cfm?id=1255175.1255221`.

[200] Jiadi Yao, Les Carr, and Stevan Harnad. Understand institutional collaboration network: a comparison of computer science and psychology. In *COLLNET and WIS (Webometrics, Informetrics, Scientometrics)*, 2013.

[201] Qi Ye, Bin Wu, and Bai Wang. Visual analysis of a co-authorship network and its underlying structure. In *2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, number vol.4, pages 689–93, Piscataway, NJ, USA, October 2008. IEEE. ISBN 978-0-7695-3305-6. 2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 18-20 October 2008, Jinan Shandong, China.

[202] D. Zhao. Going beyond counting first authors in author co-citation analysis. *Proceedings of the American Society for Information Science and Technology*, 42(1), 2005.

[203] Ziming Zhuang, Ergin Elmacioglu, Dongwon Lee, and C Lee Giles. Measuring conference quality by mining program committee characteristics. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pages 225–234. ACM, 2007.

[204] Mountaz Zizi and Michel Beaudouin-Lafon. Accessing hyperdocuments through interactive dynamic maps. In *Proceedings of the 1994 ACM European conference on Hypermedia technology*, ECHT '94, pages 126–135, New York, NY, USA, 1994. ACM. ISBN 0-89791-640-9. doi: doi.acm.org/10.1145/192757.192786.

[205] Harriet Zuckerman. Nobel laureates in science: Patterns of productivity, collaboration, and authorship. *American Sociological Review*, pages 391–403, 1967.