The Impact of Technology on Work Practices

Lanu Kim

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2018

Reading Committee:

Katherine Stovel, Chair

Kyle Crowder

Jerald Herting

Jevin West

Program Authorized to Offer Degree:
Sociology

University of Washington

Abstract

The Impact of Technology on Work Practices

Lanu Kim

Chair of the Supervisory Committee:
Professor Katherine Stovel
Sociology

The recent development of computer technology has created the belief that it provides the most neutral and efficient solution for existing social problems; however, this dissertation shows that the use of technology and its impacts are inherently embedded in the social context. Thus, the impact of technology is more nuanced and complicated, and sometimes the opposite of what it intends to achieve depending on how people use it. Empirical chapters examine the impacts on two broad work practices. First, the study focuses on how academic search engines such as Google Scholar influences scholar's work behavior of searching and citing previous literature in the course of writing scientific articles based on quantitative approaches. The results suggest that the overall citation distribution has been stable or more unequal as time goes forward. However, for those scholars who have less expertise and thereby more affected by social influence, they rely more on new academic search engines providing the information of individual paper's previous citation count. Although the previous citation count has been more decisive in making a decision, journal's role as the credential system still remains firm. Secondly, the dissertation analyzes whether the advancement in information and communication technology influenced "the death of distance" by analyzing how much geography matters for occupations. So far, there is no evidence to support

this hypothesis; instead, occupations with higher technical skills are more interdependent on other occupations' location.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First of all, I would like to sincerely thank my advisor, Kate, for guiding me through every single hurdle of my Ph.D. journey. I could unlimitedly expand my ideas owing to her firm belief and support.

Also, I appreciate all data and emotional support from Jevin. 'Echo Chamber' project would not see the light without his positive energy.

This dissertation profoundly owes Joe's input in helping me structure and manage the data.

I want to thank my parents' endless support for me. I could finish the dissertation because I knew that I was free from any expectations and only supported by loving people.

Finally, this work is supported by the NSF award 1735194, awarded to Katherine Stovel and Jevin West.

Chapter 1
INTRODUCTION

Historically, technological advancement has led to maximizing the efficiency of human labor. Two of the largest technological transformations in history—the agricultural and industrial revolutions—dramatically increased the productivity of farming and manufacturing industries. The development of farming techniques including irrigation systems and the invention of new stone tools led to raised expectations regarding stability in food supply every year that laid the groundwork for sedentary societies. The industrial revolution also had equally substantial effects on human societies. Replacing manual labor with machines led to rapid progress in productivity that reduced the number of workers required to produce the same number of products.

Some people naively argue that the purpose behind technological development is to increase the efficiency of human labor, and by doing so, eased everyday life. In this sense, the effect of technology is neutral, because it cannot act independently or influence others but becomes useful only when people utilize it. People believe that technology is a neutral means to help them achieve their goals in a faster and easier way. In this process, it might be seen that technology does not attempt to alter the intention of users but limits itself to facilitating the choices that people initially wanted.

However, as is well acknowledged, these two huge transformations brought about major social changes rather than merely expediting productivity. The agricultural revolution changed human beings' lifestyles from hunter-gatherers to agricultural societies and increased

the size of the human population, which became the foundation of the rise of civilization. The industrial revolution contributed to produce surplus products that concentrated power and wealth to capitalists who owned the means of production, which caused the demise of a hierarchical society and the beginning of capitalism. Both revolutions initiated by the development of technology created new classes and social divisions by interacting with the existing social structure.

Mackenzie and Wajcman [97] summarize in their study that the adoption and consequences of technology can be only neutral when the new technology has never been used in history and no one has knowledge regarding its usage—a rare condition that can come into existence as far as it is employed in society. What they emphasize by setting up an unrealistic condition for the neutrality of technology is that the adoption of technology and its following effects are always embedded in social contexts such as norms, values, politics, and ethics. Moreover, technological development itself can be influenced by cultural norms. For example, Mackenzie [96], following a study of a piece of computer code, argued that the performance and operationality of technical products is influenced by social practices, and is the reason why technology cannot be an independent and formal object.

The increased efficiency in manufacturing industries by automated machines showed another example of how technological advancement responded to existing social structures. Automation of human labor in manufacturing industries not only brought about the fall of manufacturing population in Rust Belt cities, but also led to the reformulation of racial inequality in these cities. In Wilson's research regarding the increase of racial inequality in the process of deindustrialization [156], he points out that black male unemployment did not recover during the economic transition from the manufacturing to the service sector. This is because the emerging service sector mostly created high-skilled jobs that black people dismissed from manufacturing industries could not apply to due to their relative lack of education. The effect of deindustrialization on unemployment is supported by Alderson [4] and

Kollmeyer and Pichler [79] as well. This research demonstrates that unemployment rate is associated with the decline of employment opportunities in manufacturing industries, signifying that a large number of manufacturing workers who lost their jobs became unemployed instead of being absorbed into the service sector.

Although the history of technological development cannot be detached from how it has been incorporated into social contexts, we are again witnessing the proliferation of the belief that technological efficiency can solve most societal issues. The new computer technologies, such as Artificial Intelligence backed by the big data and machine learning algorithms, are considered the best solutions for existing social problems. The founders of leading software companies adhere to this idea in the belief that algorithms based on math can be applied to any situation while solutions processed by human beings need individual adjustments; therefore, the general solution needs to be based on using mathematical algorithms. Broussard calls this phenomenon "Technochauvinism" [24]. According to this ideology, the method based on mathematics is the most efficient and comprehensive answer and exists above the complexities of human society, including social norms or culture.

However, numerous real-world examples have illustrated that technology is indeed embedded in the social context instead of existing separate from or above the society. Recent research shows that computer algorithms, particularly recommendation algorithms based on the big data, do not limit their role in providing neutral suggestions and instead encourage reproducing existing bias toward racial minorities, women, and the poor. For example, Noble [115] claims that Google provides racially biased search results, which is evidence that it does not provide an equal playing field for all. Because search engines are the product of private companies and are designed to benefit advertisers, algorithms are likely to be biased in favor of whiteness and men. For example, in the initial stages of Amazon Prime's same day delivery service, non-white neighborhoods were largely excluded, even when they were no more distant that other white neighborhood. [72]. Amazon neither intended nor con-

sidered the demographic composition in the service; it just computed the most cost-efficient solution given the number of Amazon Prime members in the neighborhood. Similarly, many photo applications as well as verbal voice recognition system failed to recognize the face or voice of people of color because their algorithms were primarily fed with data compiled by white males [37]. Developers intend none of the racial and gender biases found in existing algorithms; however, the bias is realized in their actual services because the algorithm learns through the biased data created in the real world, which again confirms that technology cannot be separated from the perception and knowledge of people.

Besides, human beings are embedded in social relationships and are thus influenced by the choice of others. As technology creates tools to facilitate the communication between consumers through customer reviews, social networking services, or replies to YouTube videos, the opinion of preexisting consumers influences the process of forming one's own preferences. Previous research has documented the impact of technology on social interaction. For example, by conducting the experiment with and without revealing the preference of others, Salganik et al. [129] show that the musical preferences of consumers becomes unequal and unpredictable as compared to when they are informed of others' choices. The research of Kramer et al. [81] demonstrates that another person's emotion influences users of Facebook even without interpersonal interaction, but just by being exposed to it. These examples illustrate that technology is not a simple means to assist people to choose what they want; it can reformulate the thoughts, emotions, preferences, and choices of individuals by facilitating personal interaction.

As long as the use of technology and its consequences are embedded in social contexts, the existing social structure will respond to new technological developments. As the impact of technology on society differs according to social contexts, it remains a sociological as well as an empirical question pertaining to whether there have been any social changes driven by technology and, if so, the way it looks. When the existing social structure is resilient

and strong, technology might contribute to reproducing it instead of transforming it, and in such cases, society might not be disrupted by the new technology to the extent expected. Otherwise, technology might amplify or facilitate hidden conflicts that were not prominent before.

Along with this idea, this dissertation sees technology as a socially embedded object and aims to systematically analyze its social consequences. I argue that the impact of technology is more nuanced and complicated, and sometimes it is the opposite of what it intends to achieve and depends on how society absorbs it. In this dissertation, I examine two cases to focus on social consequences on work practices driven by technological development. There are two work practices that I concentrate on. The first case examines how researchers work on writing research papers, particularly focusing on the citation behavior which is essential in developing and situating their research questions. The way of being engaged with previous literature might have been impacted by academic search engines such as Google Scholar. Academic search engines are designed to help researchers efficiently locate relevant literature based on keywords. While there can be various ways to define how to "efficiently" find articles that users need, what Google Scholar weighs on is the previous citation count to show users which papers are broadly acknowledged by academic colleagues. I investigate how this feature started and spread and how Google Scholar transforms the literature-seeking behavior of researchers.

In the second case, I examine how advances made in information and communication technology, such as electronic mails and virtual communication tools, influence the necessity of general offline interaction at work. While these tools provide the foundation for people to expand their geographical boundary for personal interaction, whether it helps people less bounded by geographical areas remains an empirical question.

In my dissertation, I use the observed quantitative data to systematically examine the impact of technology on society. The analysis of the influence of technology mostly adopts

the experimental setting to directly examine the relationship between technology and its impact. In industrial fields, A/B testing is commonly employed; it compares two groups after randomizing all possible confounding variables except one key factor. This method is beneficial in charting the uninterrupted impact of one variable. However, because this dissertation aims to reveal the effect embedded in social contexts, I try to observe natural social consequences at the expense of the direct causal relationship. In the next section, I describe the data and methods used in this dissertation in detail and summarize central research questions and findings of the following chapters.

1.1 Data and method

This dissertation aims to analyze how newly developed technology intersects with changes in work practices. In the following four chapters, I investigate the empirical question using a quantitative approach and various sources of data. In chapter two, I examine the issues with current measures of inequality describing the citation distribution and suggest one way to develop them. In chapter three, based on the method devised in chapter two, I study the extent to which scholar's citation behavior is influenced by the new search technology depending on their familiarity with the field by comparing citations made within and between disciplines. In chapter four, I investigate whether the role of journals as the credential systems have changed by using the survival analysis and the hurdle model. In chapter five, I focus on studying the impact of communication technology on the necessity of offline interaction and how it relates to the concentration of cities based on the linear regression and zero-inflated regression model.

In the first two empirical chapters, I use the bibliographic data and the citation records from the Web of Science (WoS) database provided by Clarivate Analytics. The database includes journals indexed in the Science Citation Index Expanded, the Social Sciences Citation Index, and the Humanities Citation Index. While the database has changed the boundary of

disciplines or altered the format for author's name, it is by far the most comprehensive and continuous bibliographic database that exists. The database covers the period beginning from 1900 to now. As my analysis aims to compare the period before and after the rise of integrated academic search engines near the early 2000s, I have mostly used data from between 1990 and 2016.

In the third empirical chapter, I have opted to use the database from arXiv.org (arXiv) and Microsoft academic graph. The previous research studying the impact of journal status on subsequent citation counts is limited to separating the confounding effect of the paper's quality from the effect of a journal. As a solution to this problem, I have found a quasi-experimental situation from arXiv where citations are made without knowing the journal the paper has been published in. arXiv is a preprint service that has been in usage since 1992 and is now popularized mostly in science and engineering disciplines. The citation data gathered before journal publication is considered a proxy measure of the paper's quality in my analysis. By linking arXiv database to Microsoft academic graph, I can trace the performance of papers that are preprinted in arXiv and published in journals later. This data becomes the foundation for proposing a credible solution to previously unanswered research questions.

In the last empirical chapter, I have drawn data from Occupation Employment Statistics created by Bureau and Labor Statistics in 2006 and 2016. The data covers the number of employees and the estimated wages they provide for around 800 occupations. It consists of six semi-annual surveys of 1.2 million establishments in non-farm industries of the United States. In addition, I collected data to measure the Information and Communication Technology (ICT) level of occupations from the Occupational Information Network (O*NET) program supported by the U.S. Department of Labor/Employment and Training Administration. The database covers a wide range of occupation-related features such as required tasks, technological skills, as well as wages and employment trends since 1998. The data is based

on surveys of occupation holders and job analysts.

1.2 Dissertation outline

The dissertation consists of four empirical chapters investigating different research questions related to technology and its impact on work practices. The first three empirical chapters seek evidence for the effect of academic search engines on scholarly citation behavior. In the fourth empirical chapter, I have investigated how the development of communication technology is related to the "death of distance" hypothesis. In the last chapter, I have summarized my findings from empirical chapters and discuss their implications.

In the second chapter, I have sought an answer to the debate regrading whether internet-based technologies have led to greater concentration or dispersal of scientific citations [47, 84, 56]. I have replicated measures of inequality commonly applied to scholarly citation distributions using Web of Science data covering four broad scientific fields (Health, Humanities, Mathematics and Computational Sciences, and Social Sciences) during the period 1996–2016. Following this, I have adjusted the measures to create the same publishing environment in 1996 regarding the available journals, the number of papers per journal, and the number of citations per paper. This exercise revealed, but through weak and inconsistent evidence, that although most raw inequality trends decrease with time, once they are standardized to account for the changing publication environments and isolate the citation behavior of researchers, citations become concentrated across papers. Additionally, the results showed that the trends vary depending on the type of inequality measures and the length of the citation window while the main trend remains relatively consistent across broad categories.

In the third chapter, I have used the Web of Science database to study how the relationship between a scholar's expertise in the field and social influence has been changing along with the development of new academic search engines. In this chapter, I have assumed that within-disciplinary citations are made by scholars with high expertise while inter-disciplinary

citations are made by those with relatively low expertise in the field. Two specific research questions have been investigated. First, whether there have been any temporal changes in the level of inequality in either within- or inter-disciplinary citation distribution between 1999 and 2016 before and after the rise of academic search engines. Since the time trend alone does not necessarily indicate the impact of search engines, I have provided a second analyses that traces the influence of journal-level (Journal Influence Factor, JIF) and paper-level measure (previous citation count) on the number of citations received by a paper for citations both within and between disciplines, and I have examined whether the influence has changed between 1999 and 2016. This idea relies on the features of new search engines such as Google Scholar that use previous citation count in listing search results as well as in clearly displaying this information to users—features which have not been tried in traditional search engines such as JSTOR. If the effect of previous citation count has increased while the effect of JIF has remained stable, arguments can be made citing this group as more impacted by new academic search engines. The results show that recently scholars outside the field (inter-disciplinary citations) are more affected by the previous citation count, while the behavior of scholars within (within-disciplinary citations) has not changed. Moreover, the macro temporal pattern in inter-disciplinary citations has not changed over time.

The fourth chapter focuses on tracing the changing role of a journal as the credential system of its included papers along with the development of technology. Journals have been the primary tool in the scientific process by managing the quality of papers and providing a filter in the search process. However, the role of journals is a point of discussion since academic search engines partially replace its function, through enabling papers to be searched independently. This chapter examines whether the new search technology has transformed the role of journals as the credential system by answering two research questions: 1) Has the quality of papers that pursue journal publication increased or decreased? 2) Has the effect of journal status on subsequently received citations of an individual paper reduced over time? To an-

swer these research questions, I have created a proxy measure for the quality of an individual paper by combining the data between arXiv and Microsoft Academic Graph (MAG). By applying this measure to three disciplines (High energy physics—phenomenology, astrophysics, condensed matters), I separate the effect of arguably pure journal status independent from the paper's quality. According to the results, the effect of the paper's quality on journal publication has been reduced over time indicating that currently papers with higher quality are less likely to aim for journal publication. Additionally, while the overall effect of journal status on an individual paper's quality is emphasized more than it should be, I cannot find consistent evidence to support the claim that the impact of journals on a paper's received citations has declined.

In the fifth chapter, I have studied whether the development of communication technology, which is supposed to replace offline interaction at work, at least partially, contributes to the dispersion of wealthy and talented people and helps prevent the concentration of wealth to a few cities. In between positive speculation on the role of technology in reducing the necessity of physical distance and the emphasis on the robust role of offline interaction, I have investigated this problem with a broader view through analyzing the distribution of occupations in the U.S. cities between 2006 and 2016. By using the data from Occupation Employment Statistics created by Bureau and Labor Statistics and the (O*NET), I have measured the importance of geographical location for occupations by employing the geographical dispersion and geographical interdependence of occupation dyad. With these two measures, I have studied whether the amount of required communication technology as an occupation is related to its geographical importance. From the results, there is no evidence to support the assertion that the required computer skills of an occupation decreased the geographical dispersion of occupations both in 2006 and 2016; however, it strengthens the geographical interdependence of occupations. To summarize, I could not find evidence in favor of the claim that information and communication technology has brought the death of

distance and is likely to bring it in the near future because the trend in the last ten years has barely changed. Instead, occupations with higher technical skills are more likely to exist at the same location and have geographical consistency with other occupations. In addition, the interdependent occupation pairs have higher income level than non-interdependent pairs, which is consistent with the concentration of wealth to a limited number of cities.

In the concluding chapter, I have summarized findings from empirical chapters and illustrated the limitations of the dissertation. In explaining limitations, I have pointed out the constraint of indirect evidence of technological change and possible data bias. Subsequently, I have found broad implications by expanding the meaning of findings and provide possible policy suggestions. Mainly, I have warned against the unquestioned use of search engines for literature search as well as bibliometrics for research evaluation. Also, I have discussed the limitation of technological determinism because the impact of technology needs to be interpreted in the social context. Finally, I have described future research ideas originated from this dissertation.

The dissertation contributes to broadening our understanding of how work practices have been changing as technology develops by focusing on the case of scholarly citation practices and the necessity of geographical locations at work. The development of the search engines and communication technology has originated from the private market competition, mostly led by a few companies. As a result, their algorithms and technologies are proprietary property that are neither under obligation to be explained to nor to be watched by the public. However, their impact on society is beyond what a single private company can achieve even if the initial intention of the company is limited to maximizing their profit. This dissertation focuses on discovering and elaborating on a few examples of these cases, especially in work practices. According to my findings, the impact of technology appears to be more subtle and complicated than universal and straightforward. On the one hand, the amount of impact is contingent on types of users and existing institutions in the work environment. On the

other, technology relates to social changes contrary to its primary purpose. The following four empirical chapters elaborate these points with the unique analytic strategies backed up by large-scale data and quantitative methodologies.

Chapter 2

GROSS INCREASE IN PUBLICATIONS AND CITATIONS IMPACTS MEASURES OF SCHOLARLY CITATION PATTERNS

2.1 Introduction

[1]While the structure of citations to scholarly papers has been studied since de Solla Price's seminal work [41], this line of research has been reinvigorated by the digitization of journals and the emergence of new integrated search engines. Understanding changing citation patterns is important because irrespective of whether new technologies intensify citation to a relatively small group of star papers or spread citation across a broader range of peer-reviewed articles, it has significant implications on scientific advancement [59, 84, 47, 10, 49, 50]. Studies querying whether search technologies concentrate or broaden exposure are not limited to scientific citation behavior but focus on consumers' decisions including online clothing markets [27], video rentals [161], and music consumption [129]. In the field of academic bibliographic studies, recent papers offer contradictory evidence on how citation patterns have changed, focusing on the rise of online journal access. Evans found evidence of increasingly concentrated citations [47], while Larivière et al.'s analysis of aggregate trends over time revealed more diversified citations in humanities, social sciences, natural sciences and engineering, and medical fields. [84].

In their subsequent discussion [56, 48], these authors conclude that the discrepancy in their findings is a consequence of posing slightly different research questions, different data

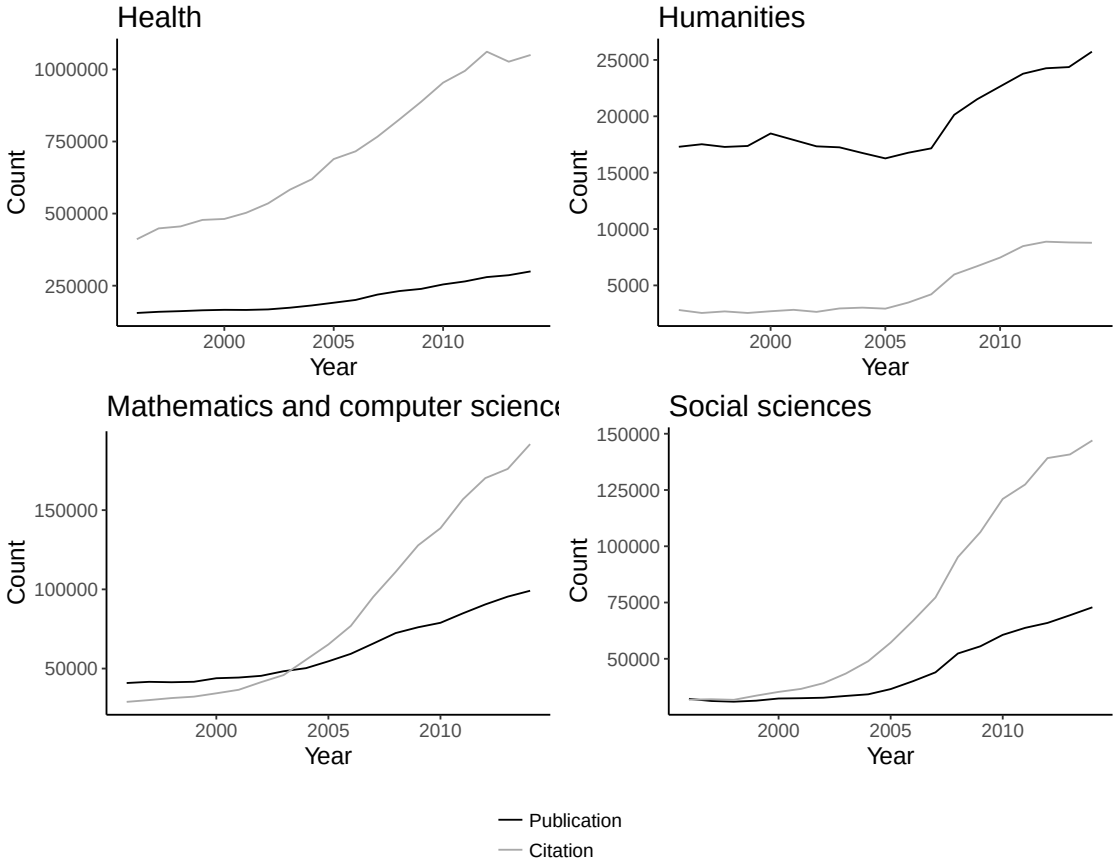
[1]This work is from a collaborative project undertaken with Christopher Adolph, Katherine Stovel, and Jevin West.

structures, and the application of control variables in the analysis. Specifically, Evans' original work estimated panel regression models that include fixed effects for subfield and year; he found that between 1965 and 2005, online availability of journals was associated with more recent and more concentrated citation of papers and journals. In contrast, Larivière et al. computing inequality measures[2] for the aggregate citation distribution each year between 1900 and 2007 reported that, in the observed period, a broader range of papers has been cited and there was less concentration towards top papers across all disciplines. Thus, while Evans concludes that citations have become more concentrated, he concedes that this pattern is only revealed after controlling the online availability of journals; without these controls, Evans found the same trend towards dispersal reported by Larivière et al.

I have contributed to this discussion by highlighting a fact that has important implications in this debate; over the past several decades there has been a dramatic increase in the number of academic publications and, in many fields, an even greater increase in the number of citations made. For example, Figure 2.1 shows changes in the number of publications and the frequency with which they were cited for four disciplines between 1996 and 2014. In each field, the number of papers published and citations to these papers has increased since 1996, in some cases drastically. This expansion in publications has been widely noted and studied in the context of understanding inflation in Journal Impact Factor [7], the aging of scientific literature [87], and the growing myopia of science [120]. However, current discussions, with respect to the inequality of citation patterns, pay less attention to the possible impact of the gross increase in publications and citations.

[2]Inequality measures indicate measures that are used to show the amount of concentration in the distribution, in this case, the citation distribution.

Figure 2.1: Journal Articles published between 1996-2014 and citations to these articles (two-year citation windows, 1998 - 2016); Black line - publication counts, grey line - citation counts



Assessing the change in the shape of a distribution requires a consistent measure that can be compared over time or place. It is well known that fully capturing the shape of a distribution with a single number is impossible, and for this reason, multiple approaches have been proposed to compare distributions. One strategy is to calculate the share of one value or entry in the total distribution [51, 161]; another approach is to summarize the shape of the distribution with respect to its total deviation from a normal or uniform distribution.

Commonly used measures include the Gini coefficient [27, 129] and the Herfindahl-Hirschman index (HHI) [47]. Each of these approaches has limitations, the most well-known being that different values of a measure can reflect quite differently shaped distributions. In spite of these problems, scholars studying the manner in which technology has changed behavior use these inequality measures to draw substantive conclusions about temporal changes (e.g., [70, 125, 159]).

Measuring distributions of academic citations introduces a series of less-appreciated problems: citations to papers are not divisible, the total number of citations is sometimes less than an order of magnitude greater than the number of citable papers, and in most fields, large fractions of papers are never cited (yielding many zeros in the distributions) [27, 84]. Due to this problem, the combination of the rapid increase in publications and references influences a large part of changing patterns of inequality in citation distribution [151]. Collectively, these problems suggest that comparisons based on inequality measures may be inadequate or even misleading when the increase in paper and citation counts confounds factors not associated with the scholar's citation behavior.

Thus, I have argued that it is essential to understand which factors drive the increase in the number of academic publications and separate these factors from the macro trend to understand a scholar's changing citation behavior. Among many reasons that are irrelevant to changes in researchers' citation behaviors but still influence the temporal pattern in inequality measures, I have identified three major scenarios that might confound understanding researchers' behavior. First, indexing of Web of Science (WoS) database might add different tiers of journals in different time periods. If low-tier journals are added intensively in a later period, citations sent from low-tier journals might have different patterns of citations different from high-tier journals. Even if indexing is not an issue, contemporary journals might have been placed in a different tier in comparison to traditional journals. Secondly, the number of papers per journal might have increased over time as more journals actively use

electronic publication tools. In this case, the proportional impact of each journal on citation patterns might have changed over time. The last scenario suggests that the citation count increase might be due to the increased number of references. With increased accumulation of scientific findings in recent years, the reference list tends to get extended.

In this study, I investigate how the distribution of scientific citation has transformed in the age of integrated academic search engines after removing possible confounding components. Using data gathered from the WoS, the trend is replicated and temporally extended towards the more dispersed citations found by Larivière et al. [84] and confirmed by Evans [48]. In comparison to prior works, this analyses includes more recent data spanning until 2016 and therefore captures the impact of integrated proprietary search engines like Google Scholar on academic citation behavior. The citation distributions are then adjusted such that they control possible scenarios assuming that all situations are congruent to the initial year of the observation—1996. The purpose of this exercise is to reveal how the distribution after separating confounding effects influences measures of inequality and whether this adjustment changes the interpretation of macro time trend; precisely, whether the adjusted citation distributions have become more concentrated or more dispersed. Following this, the time trend was reinterpreted with adjusted inequality measures by comparing two- and six-year citation windows. Finally, the impact of each component of adjustment on inequality measures by disciplines was ascertained. Based on the new adjusted measures, evidence of somewhat stable or less dispersed trend over time was found, though the strength of this trend varies according to discipline and inequality measures.

2.2 Data and method

Publication and citation data from the WoS, provided by Clarivate Analytics, was analyzed, and a data structure that generally follows Larivière's approach was constructed . WoS includes the Science Citation Index Expanded, the Social Sciences Citation Index, and the

Arts and Humanities Citation Index. The primary focus of this study are the research papers published in English language journals between 1996 and 2014[3] in four broad disciplinary fields (health, humanities, mathematics and computer sciences (Math & CS), and social sciences) and citations to these papers in the two and six years following publication.[4] Editorial comments, books, and other non-research articles from both citing and cited papers were excluded. To illustrate the data structure with a two-year time window for all papers published in the social sciences in 2014, citations made to these papers before the end of 2016 from papers in other disciplines have been registered. Therefore, published papers from 1996 to 2014 are followed by citations of these papers from 1996 and 2016. Table 2.1 provides the total number of papers and citations in each broad category. Details of how journals and disciplines are aggregated into the four fields are described in Appendix A.

Table 2.1: Total count of published papers and citations made to papers with two-year window by broad category between 1996 and 2014

Broad category	Papers	Citations
Health	4,904,174	15,804,637
Humanities	502,620	111,245
Mathematics and computer sciences	1,434,715	1,825,581
Social sciences	1,079,941	1,616,485

Using the data, the same yearly, field-specific measures of citation inequality used in

[3]Because of uneven coverage during much of the twentieth century, I limit the analyses to the period between 1996 and 2016. This would allow me to compare the results from 1996–2007 to those in [84] and then examine subsequent trends to 2016.

[4]While Larivière et al. use two- and five-year windows in counting citations, I replicate results by using a two-year window. Then, instead of using a five-year window, I have used a six-year window to to increase the citation period of these document. The result of the six-year window is attached in Appendix A.

Larivière et al. were calculated [84] i.e.: the percentage of papers published in a given year that received at least one citation; the percentage of papers needed to fill 20% and 80% of the total citations received by papers published in a given year; and the HHI.[5] In addition, the Gini coefficient was also computed—another standardized inequality measure frequently used to assess income or wealth inequality.

As discussed in detail in the following sections, these calculations reproduce Larivière et al.'s substantive results: across multiple disciplines, the share of papers cited is increasing, and the concentration of citations is declining. As part of this study, before adjusting the data, self-citations following Larivière et al were excluded. This step has been undertaken not only with the intention of replicating previous results, but also because self-citations are not necessary in the analysis since it is irrelevant with the author's literature search behavior.

In addition to the replicated results, the same inequality measures were calculated by using the data after adjusting it to have structural conditions similar to 1996 by broad category. The major goal of the adjustment exercise was to set up similar conditions for all years as 1996. As a means to achieve this goal, first, the list of journals that have published and cited articles between 1996 and 2014 was established. Both cited and citing articles that are published in this list during the observed period were left. Following this, the number of distinct papers per journal sending citations to papers published in 1996 was calculated, and the same number of papers per journal as 1996 was randomly selected with replacement for subsequent years. Lastly, the average number of citations per paper that cited articles published in 1996 was counted, and again, the randomly chosen citations with replacements

[5]The HHI is a commonly used measure of market concentration computed by summing the squared market share of each firm. In the context of this study, the market share is the citation count that one paper receives divided by the total citation count. Usually, when the HHI is smaller, it means the market is more decentralized; however, the HHI also tends to decrease when the number of participants is high. For example, when ten companies equally share a market, HHI is $(0.1)^2 * 10 = 0.1$, but when 100 companies share it, the HHI is $(0.01)^2 * 100$, or 0.01. As this illustration shows, ceteris paribus, HHI will decrease if publication counts increase. Due to its inherent limitations, the analyses of HHI is only included in Appendix A.

for the following years to each year had the same number of citations per paper in 1996.

2.3 Results

2.3.1 Comparison of adjusted and unadjusted data

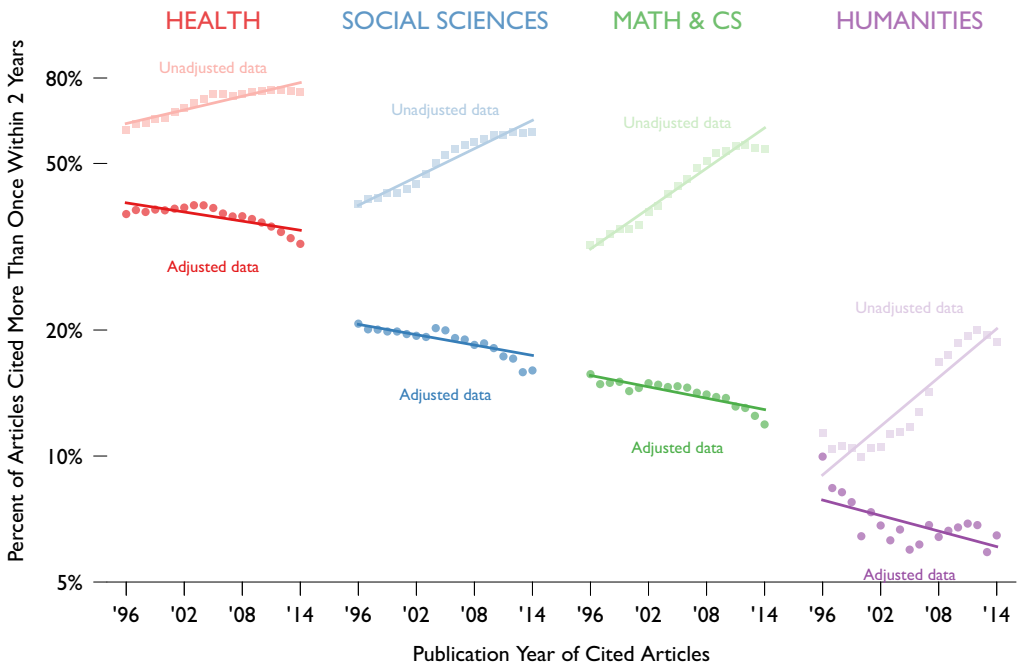
Figure 2.2 demonstrates the results for the direct measure of dispersion: the share of papers cited in a given year that maintain their rate of citation in the following two years. The figure shows this percentage between 1996 and 2014 in four broad domains with a two-year citation window. Opaque dots and lines denote the adjusted data and transparent dots and lines denote the unadjusted one. The lines across dots summarize the temporal trend computed based on a robust linear regression model. A solid line indicates that the temporal trend has changed statistically by a significant amount, whereas a dotted line implies that no statistically significant temporal trend has been detected. In Figure 2.2, the observed temporal trends are statistically significant in all instances.

The trend based on the unadjusted data in Figure 2.2 (transparent dots and lines) successfully replicates Larivière et al.'s analyses. While differences exist between broad disciplinary groups in the the share of papers typically cited (ranging from 10% to 80%), the observed temporal trends are generally upward, and confirm Larivière et al.'s analyses: the share of papers cited increases proportional to time. Focusing on the years 1996 and 2005 and the two fields (social sciences and humanities) that most closely resembles Larivière et al.'s analyses, the results roughly replicate the prior estimates. In the humanities, the share of ever cited papers between 1996 and 2005 hovers around 10%—a stable trend that is similar to their estimates—while in the social sciences the estimate increases from 44% in 1996 to 55% in 2005—almost identical to Larivière et al.'s increase from 42% to 54%.

In contrast, once the data was adjusted following the procedure explained above (opaque dots and lines), the temporal trend was found to be contradictory to the one based on the unadjusted data. In all four broad categories, the dispersion of citations across papers

shows significant decline with the adjusted data, which are inconsistent with Larivière et al.'s findings. In health, social sciences, and math & CS, the slope of declining percentage becomes steeper after 2005 than before. In humanities, the trend fluctuates more than other broad categories, possibly reflecting the decreased reliance on journal articles among researchers.

Figure 2.2: The temporal trend of percentage of ever cited papers between 1996 and 2014 by four broad categories (two-year citation window); Opaque dots – adjusted data, Transparent dots – unadjusted data; Solid line - statistically significant time trend & Dotted line – statistically insignificant time trend (a statistical test was conducted using a robust regression model)



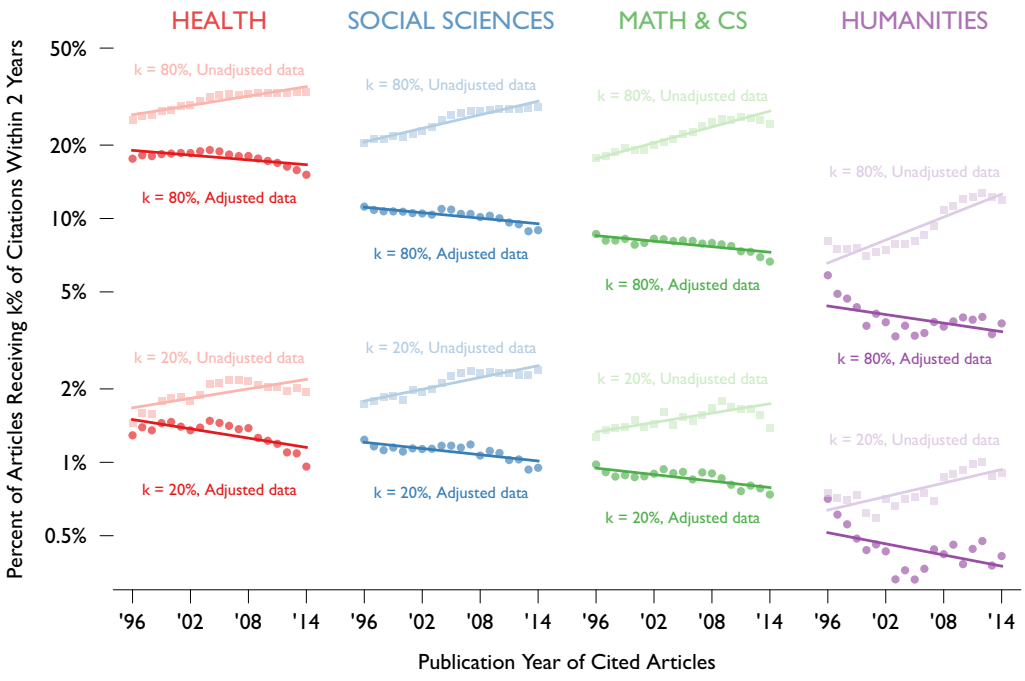
Similarly, the percentage of papers that account for 20% and 80% of citations in Figure 2.3

was examined.[6] When more papers are needed to account for each of the three percentages, citations are more equally distributed (less concentrated). Larivière et al.'s essential findings for humanities and social sciences were again replicated; it was found that approximately 10% of papers account for 80% of citations in the humanities from 1996–2005; about two percentage points higher than Larivière's result, but are similar in pattern. In the social sciences, this percentage increases from 24% to 28%, very similar to Larivière et al. Most critically, across all disciplines and levels of measures the 20% and 80% distributions are becoming more equitable over time—evidence Larivière et al. uses to support the claim that citations are becoming less concentrated.

However, as was the case for the percentage of ever cited papers, the trends in the 20% and 80% shares in the adjusted data showed the opposite and statistically significant temporal trend in comparison to the unadjusted data. The decreasing percentage implies that fewer papers account for most of the citations; evidence of increasing (rather than decreasing) concentration in citations in these fields.

[6]The percentage for papers that account for 50% of citations is omitted from Figure 2.3 to simplify the illustration of results.

Figure 2.3: The temporal trend of percentage of papers that needed to account for 20% and 80% of citations between 1996 and 2014 by four broad categories (two-year citation window); legends are the same as Figure 2.2

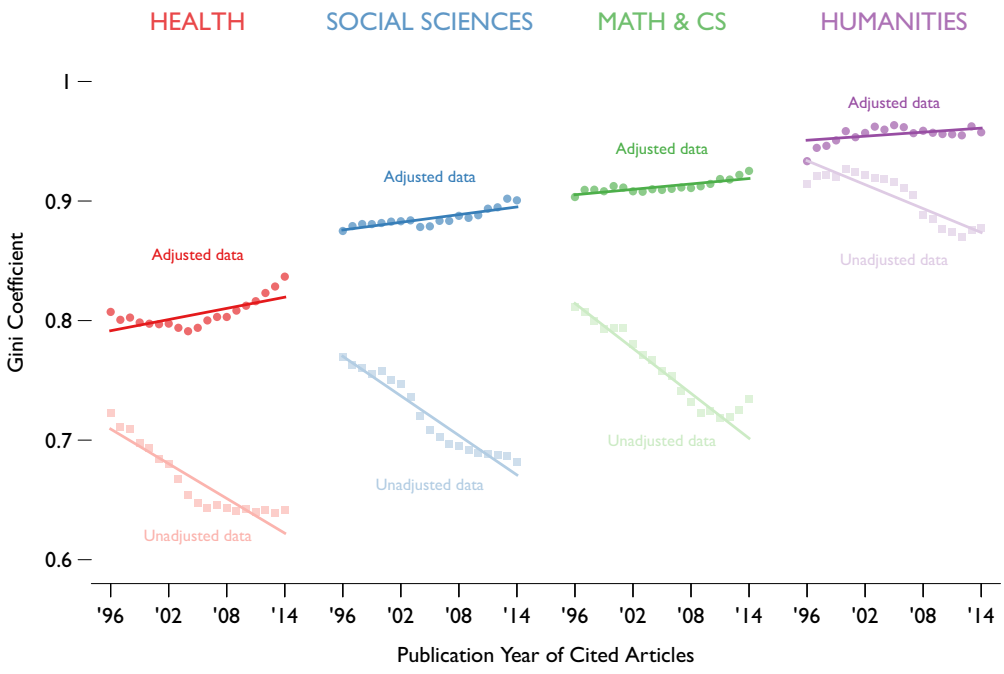


Lastly, in Figure 2.4, the same method was applied to the Gini coefficient. The Gini coefficient is another frequently used inequality measure (e.g., [129]) that attempts to standardize distributional differences by capturing the difference in area between an observed distribution and the line of equality. The Gini equals 1 for perfectly unequal distributions and 0 for perfectly equal distributions. However, like all scalar summaries of distributions, the interpretation of the Gini can be ambiguous because it is unclear as to which part of the distribution exacerbates or improves the Gini is unclear.

What Figure 2.4 shows is consistent with what was found in Figure 2.2 and Figure

2.3—while the citation distribution seems to become equal (close to 0) to the data before adjustments, it illustrates an opposite trend toward unequal distribution (close to 1) after adjustments. However, the amount of temporal change is distinct from the other two measures. In Figure 2.4, the amount of increase in the adjusted data is not as dramatic as the amount of decrease in the unadjusted data. To interpret, the distribution becomes dramatically democratized when analyzed with the unadjusted data; however, with the adjusted data, the distribution has been slowly becoming unequal over time. The degree of change of change in temporal trend caused by types of inequality measures implies that the adjustment variously influences the measures.

Figure 2.4: The temporal trend of the Gini coefficient between 1996 and 2014 by four broad categories (two-year citation window); legends are the same as Figure 2.2



2.3.2 Comparison of temporal trends with two- and six-year citation windows

In the previous section, the manner in which the temporal pattern of citation distribution changes with parallel adjustments made in the data to control external conditions that might confound scholar's citing behavior, were examined. In Figure 2.5 and Figure 2.6, the temporal trend of the citation distribution based on two- and six-year citation windows to examine whether there are any differences in a short-term and long-term citation distributions were compared.[7] The last observation made is 2014 for a two-year citation window, which means it is the distribution of citations that are made to papers published in 2014 where citations are collected from papers published in 2015 and 2016. For a six-year citation window, the last observation is 2010, which indicates the distribution of citations made to papers published in 2010 where citations are collected from papers published between 2011 and 2016. Thus, there are four fewer observations for a six-year citation window because it needs full six years after cited papers are published while a two-year window needs two years, and the last year of citation data is equally 2016 for both time windows. The results of a two-year citation window in Figure 2.5 is the same as the adjusted data in Figure 2.3.

Figure 2.5 summarizes the temporal trend of the percentage of papers that needed to account for 20% and 80% of citations in a two- and a six-year citation window. The percentages are in general higher for a six-year window than a two-year one. This indicates that as citations are accumulated longer, they are less concentrated on the top. Furthermore, temporal trends of two time periods were found to lead to a different conclusion. According to the test of a robust regression model, the concentration of the citation distribution has significantly increased with a two-year time window. However, in the analysis based on a six-year window, no statistically significant temporal changes are found except in humanities. There are two possible reasons for the presence of different temporal trends. First, because

[7] I compare the percentage needed to account for 20% and 80% of citations, and the Gini coefficient in this section. Plots comparing other measures are included in Appendix A.

the yearly difference of a six-year window is arithmetically smaller than that of a two-year window, the time trend might seem more stable with a longer time window. Another possible explanation is the lack of any considerable change in the scholar's behavior of citing papers aged older than two. Researchers might change their search behavior more dynamically for seeking recently published papers, which might be reflected in increasing concentration of the citation distribution with a two-year window.

Figure 2.5: The temporal trend of percentage of papers needed to account for 20% and 80% of citations by four broad categories (1996-2014 for two-year citation window, and 1996-2010 for six-year citation window); Circles – two-year citation window, Squares – six-year citation window; Solid line - statistically significant time trend & Dotted line – statistically insignificant time trend (a statistical test is done using a robust regression model)

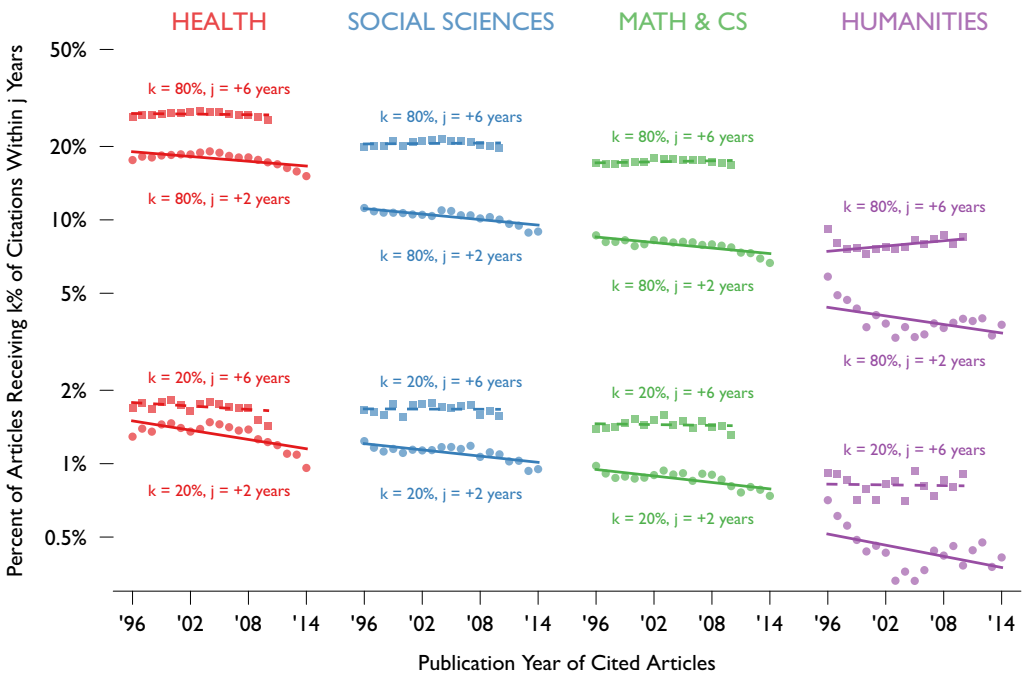
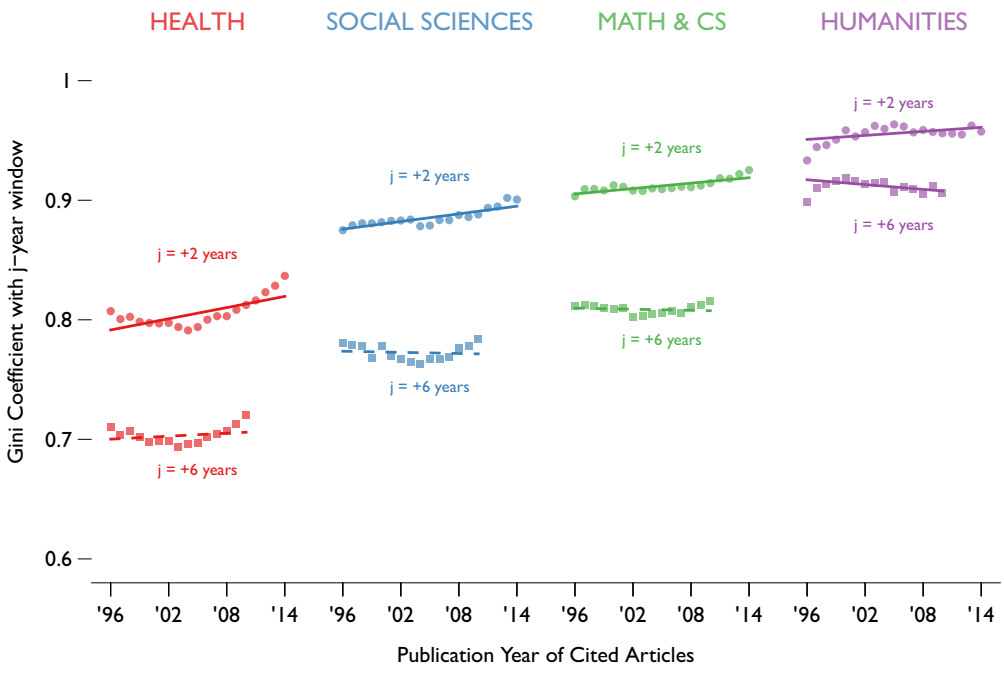


Figure 2.6 summarizes the result of the Gini coefficient analyzed in the same manner as Figure 2.5. It displays findings similar to the above figure: the distribution is more democratized, and the temporal change is more stable in a six-year window than in a two-year one. Although the amount of concentration in the citation distribution with a six-year window seems to decrease until the early 2000s and begins to increase after, which might cancel out the effect of temporal trend measured in a linear specification, more evidence from recent years is required to confirm increasing concentration trend.

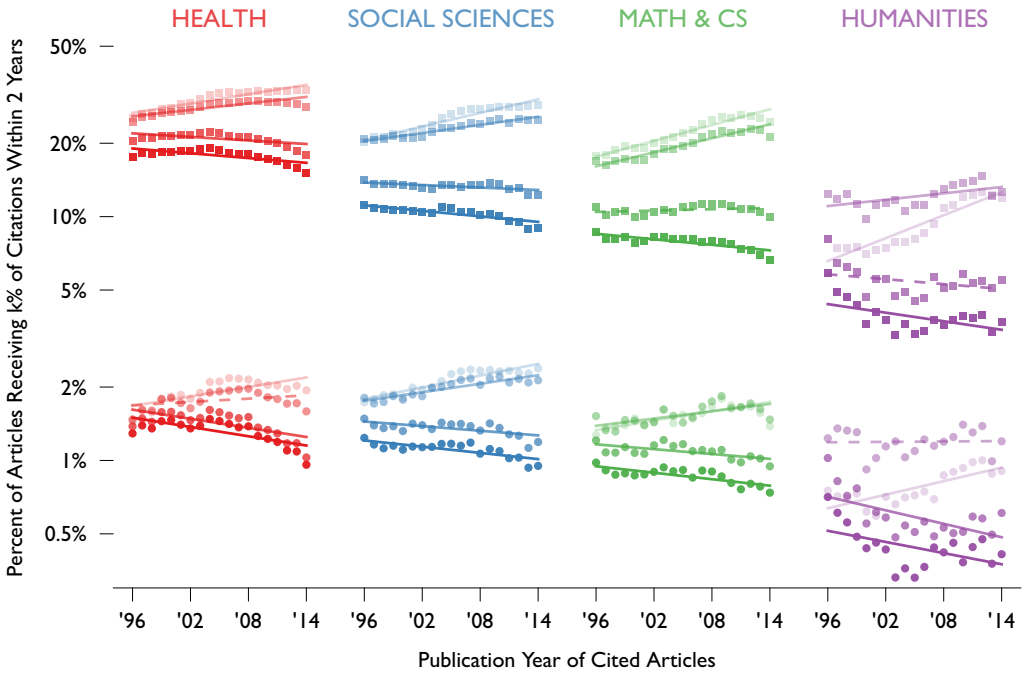
Figure 2.6: The temporal trend of the Gini coefficient by four broad categories (1996–2014 for two-year citation window, and 1996–2010 for six-year citation window); legends are the same as Figure 2.5



2.3.3 Analysis by adjustment component

In this section, the adjustments are broken down to determine the effect of each component step-by-step. Figure 2.7 summarizes the results. The most transparent dots indicate the pattern of unadjusted data and the most opaque dots show the fully adjusted data, which replicates the results from Figure 2.3. In between the two extreme colors, the less transparent dots signify the results that control the list of journals, and the next transparent dots indicate the results controlling the number of papers per journal in addition to the list of journals. The major change in the percent of papers receiving either 20% or 80% of citations was found when the number of papers per journal for citing articles was adjusted. The same trends are found in other inequality measures, which are attached in Appendix A.

Figure 2.7: The temporal trend of the percentage of papers that had to account for 20% and 80% of citations between 1996 and 2014 by four broad categories (two-year citation window); Circles – 20% of citations, Squares – 80% of citations; Solid line - statistically significant time trend & Dotted line – statistically insignificant time trend (a statistical test is done using a robust regression model); From the brightest to the darkest color: unadjusted data, the list of journals adjusted, the list of journals + the number of papers per journal adjusted, fully adjusted



2.4 Discussion and conclusion

Since changes in inequality measures have been used to infer changes in how scholars behave [70, 125, 159], it is important to ensure that measures are comparable year-over-year. Measures used in previous bibliographical studies were replicated and it was confirmed that the

time trends observed in these studies were partly an effect of external factors not related to scholar's citing behavior. The external factors include the expansion of journal lists in WoS database, the increase in the number of published papers per journal, and an extended list of references. Not only the percentage of ever cited papers and percentage of papers accounting for 20% and 80% of citations, but also a standardized measure—the Gini coefficient—is influenced by these factors. Thus, while it is true that in absolute terms more and more papers are being cited, the trend behind this finding is predominantly driven by structural factors other than scholar's search behavior. After adjusting measures of inequality to control these confounding effects, little evidence was found of a dominant trend towards decentralization. Instead, the time trend in the new measures suggests that if anything, citations are becoming even more concentrated than in earlier eras. However, the increasing concentration trend was found only in citation distributions with a two-year time window, which captures relatively recent responses to papers. For a six-year time window, it is hard to find evidence against the concentration and decentralization of the trend. Moreover, the most influential external factor was found to be the number of papers per journal of citing papers.

These results are relevant to any research on the impact of access to more information or options on decisions. Marketing studies consistently show that online technologies increase overall sales (e.g., [40]), but as in citation research, whether technology encourages idiosyncratic consumption or amplifies blockbusters is in dispute. One stream of research argues that online search technology increases sales of niche commodities because it makes them more accessible[28, 27, 161, 152]; however, another stream of research reveals a contradictory pattern: convergence on a smaller number of massively popular products [66, 91, 51]. Salganik et al. [129] and Elberse [46] find that this latter phenomenon is amplified when users know others' preferences, particularly for goods for social consumption like music or movies. Given that scientists rely on prestige signals to gauge the quality of papers [94] and prioritize their attention on this basis [102], the results of adjustment exercises are consistent with

the idea that new search technologies amplify existing concentration in citations, perhaps by providing status cues that scholars use to navigate the ever-expanding sea of scientific literature.

A limitation of this research is its exclusion of citations made between two different broad categories in order to control external factors. As a means of restricting the list of journals, only the same list of journals that were available in 1996 for both cited and citing articles were considered, which subsequently deleted the citations made outside of the broad category of cited articles. Thus, if a different pattern exists for citations made between two different categories, it has not been accounted in these results.

Chapter 3

EXPERTISE, SOCIAL INFLUENCE, AND TECHNOLOGY

3.1 Introduction

[1] Accessing prior scientific knowledge is a critical component of modern scientific practice, and technological developments over the past several decades have revolutionized how scientists discover and cite previous research [47]. More and more scientists now access their literature through online search engines and digital libraries, and rare is the scientist who walks into the library and peruses the journal shelves for new papers [114] [143] [117]. Among these technological developments, the emergence of academic search engines such as Google Scholar are of particular interest because of scholars' increased reliance on them in seeking literature, and also because these services are powered by black-box, proprietary algorithms that attempt to actively anticipate the users' needs rather than simply listing papers indexed by keyword. Research suggests that these new technologies have implications on the breadth and depth of the users' knowledge of their fields [91] [133] [149].

In the process of searching and citing articles in a paper, researchers often seek signals such as journal or author prestige and previous citation count that are supposed to reflect the quality of articles. Because the challenge of searching for literature is growing along with the size of the scientific corpus, which is doubled roughly in every nine years [20], positional cues of articles can help reduce the search cost for researchers by narrowing down the pool of papers that are worth investing time and energy over. Moreover, citing articles with clear signals can easily persuade the audience to accept the significance of the research or the logic

[1] This work is from the collaborative project undertaken with Katherine Stovel and Jevin West.

of arguments. The reliance on social cues is especially stronger for scholars citing outside of their own field, not only because outsiders in general have higher search costs, but also because they are in need of convincing the audience by citing influential articles [94]. This process is similar to the diffusion of innovations where a late adopter only complies with new technology when a certain proportion of early adopters already use it [160].

The changing relationship between expertise and susceptibility to social influence and its transformation alongside the advancement of search technology and its subsequent impacts on the scope of science, have been studied relatively less. The introduction of academic search engines might provide an opportunity and at the same time act as a limitation in decreasing the gap of citation behavior between researchers with and without expertise. Since almost every scientific literature can now be accessed online and since the size of total literature is endlessly growing, it becomes increasingly difficult for scholars to keep up with even their own fields. The new search engines reduce the cost of literature search by facilitating the navigation of this huge corpus. Search based on keywords increases the visibility of articles even in low-tier journals if papers contain matching keywords, which helps outsiders to easily locate papers in their various interests.

However, the possibility also exists of scholars with a lack of expertise being more impacted by social influence through using academic search engines than scholars with expertise. Most of these services likely use received citation count as an input in the algorithm that determines what results to display [14], and presents viewers with this quantified information along with search results. In general, the order of search results strongly affects how users allocate attention due to human cognitive biases [91], and thus, displaying the measure of popularity of an item might cause the consumption of only items that are already popular[67, 129]. Scholars citing outside of their fields might rely more on the search engine's algorithm and choose papers that top search results because the massive amount of incoming information might put new pressure on the searching and filtering processes, particularly for

non-experts.

Depending on how researchers use search engines, citation pattern might be more democratized as new tools contribute to making scientific literature more accessible, or vice versa. If it equalizes the citation pattern, it would certainly have implications on the prospects of scientific discovery, since it implies that scientists are drawing on all relevant prior works—a critical ingredient for high-quality scientific activity. On the other hand, the availability of enormous amounts of information might increase the search costs and make researchers rely on top search results of academic search engines. This could mean that scientists are increasingly reading and citing a more concentrated subset of papers, and are therefore at risk of being trapped in a scientific echo chamber.

Through studying citations made in academic research articles, this study seeks to investigate the following research questions: does the pattern of scientists engaging with prior scientific research of the field differ by the amount of expertise scholars have, as the academic search engines are popularized? Specifically, do the search engines influence scholars to be more susceptible or immune to social influence depending on their knowledge distance to the field they are citing? Then, subsequently, has scholars' citation behavior expanded to a wide range of potentially relevant prior work, thereby democratizing science, or concentrated into an ever smaller set of "star" papers?

In this work, these questions will be answered in the reverse order. First, the temporal trend of insiders and outsiders in the field between 1999 and 2016 that covers the period before and after the emergence of academic search engines was traced to understand whether scientific citations are becoming more concentrated or diversified across a wide range of disciplines. As a means of comparing the behavior of insiders and outsiders, the distribution of inter- and within-disciplinary citations was examined. While the scholars' use of search engines in seeking literature, particularly Google and Google Scholar, has dramatically increased since the early 2000s across places [150, 117, 143], the time trend alone does not

necessarily indicate the impact of search engines. Therefore, a second set of analyses has been provided that compares the influence of journal-level (Journal Influence Factor, JIF) and paper-level measures (previous citation count) on the number of received citations of a paper for both inter- and within-disciplinary citations, and see whether the influence has changed between 1999 and 2016. This analysis leads to two observations. First, tracing changes in the effect of previous citation count of within-disciplinary (insiders) and inter-disciplinary (outsiders) citations reveals the group which has become more susceptible to social influence along with the use of search engines. Additionally, if the impact of previous citation count changes while the effect of JIF is relatively stable over time, it serves as evidence that the group of scholars in the analysis is arguably more influenced by search engines. This is because only the new systems illustrate paper-specific information such as previous citation count and hyperlinks from and towards a paper, which were provided in traditional search engines such as JSTOR. If evidence is found as part of this dissertation in favor of the increased influence of the popularity of a paper, it would imply that the use of search engines is changing the behavior of researchers, and contributing to the observed macro patterns of citation distributions.

3.2 Literature review / Background

3.2.1 Expertise, social influence, and technology

When researchers filter papers to read and cite, they are influenced by the status of papers as signaled by a journal of published papers, authors themselves, and authors' affiliated institutions, and the significance of papers as perceived by their colleagues. Researchers especially rely on status signals when they are not confident about their judgment of its quality or due to lack of time and energy for searching, under the belief that the status reflects the quality of papers [123]. Although status is supposed to reflect the quality of a paper, evidence has shown that the coupling between status and quality is likely to be

weakened when social interaction influences the process of status formation [31, 57, 95]. Gould [57] argues that socially influenced judgments are likely to exaggerate the quality of those who are objectively positioned above the average in terms of their 'actual quality' and underestimate those positioned below the average in comparison to the situation where there is no social interaction, because social actors seeking signals for their judgments are influenced by interactions suggesting the actual quality. In this context, the researcher's expertise in the field helps maintain a tight relationship between status and quality as the expertise lessens the reliance of researchers on others in judging the quality of articles. This expertise can be gained by training graduate students within a certain boundary of academic discipline to discern the robustness of scientific methodology and the significance of studies, or by engaging in academic discussions with colleagues sharing similar research interests. Based on the systematized scientific training and academic communication in the specific field, scholars with expertise can reduce the time spent seeking quality signals and learn evaluating quality from the article itself.

However, disciplines are not isolated segments, but resemble an interconnected web that allows knowledge to diffuse [73], implying there can be audience with varying levels of expertise in the field [94]. Depending on their partners in major academic conversations, scholars in one group might have a different perception regarding a paper's status in comparison to another group since status is the quality perceived by observers [123] and the local consensus about status can be different from global status hierarchies [80]. Status construction theory supports the possibility that scholars might have varying notions pertaining to status hierarchies of articles [80]. The theory explains how status beliefs are updated and spread through social interaction of groups [126], and through this social process, sometimes the observed status hierarchies do not reflect the real quality difference [127, 98]. By emphasizing the importance of social interaction in creating and diffusing status hierarchies, this theory suggests that scholars with different backgrounds might not reach the consensus in evaluating

the status of papers.

The recent development of search technology contributes to the decrease in disunity around the perception of status hierarchies. Academic search engines, particularly the proliferating Google Scholar, create an equal ground for everyone to have access to papers regardless of their level of expertise. Before academic search engines, the researcher's perception of the paper's popularity varied depending on the academic communities they were affiliated with. However, as search engines begin providing the same order of papers for the same keywords' search and information on the previous citation count gets updated every moment, scholars access the same result pages, which can help reach a global consensus regarding the paper's status. In this case, the standardized result pages impacting the user's perception of the paper's significance generated by proprietary algorithms is not a carefully planned work of institution (e.g., [135]), but rather a by-product of technological transformation that is still changing in practice depending on the components of the algorithms (e.g., [119]).

The next section goes into a detailed description of the background of academic search engines, mostly Google Scholar, and summarizes previous literature about its effect on citations.

3.2.2 Background

Google Scholar, one of the most popularly used search engines, was launched with the audacious goal of creating a single efficient search engine where scholars as well as the general public could access scholarly information from all disciplines and languages [55]. Due to the appeal of its brand, accessible interface design and broad coverage, this search engine has quickly become widely used by scholars for research purposes; nearly three quarters of PhD students use or have used Google Scholar [35], though when compared to other web based services, there is variation in its usage by discipline and age [117, 150, 143, 21]. Despite

Google's dominance, there has been an explosion of academic search engines over the last couple of years. In 2016, the Allen Institute of Artificial Intelligence (AI2) released a new search engine called Semantic Scholar. Microsoft Academic search resurfaced with new recommendations and visualizations for exploring the literature. JSTOR is developing a new search interface for their large corpus of papers. The Web of Science (WoS) was recently purchased by Clarivate, with plans to incorporate new recommendation algorithms. PubMed recently added paper recommendations. Mendeley now provides recommendations based on researchers' bibliographies. During the last several years, we have seen the golden age of academic search and recommendation, but there is not yet a consensus on the most efficient and reliable search algorithms, nor on the best way to evaluate the performance of search results (e.g., [155, 153, 13]). These are exciting times for researchers trying to manage an ever-expanding literature, but with new technologies, there is a need for examining its effects on the curation and navigation of literature.

Despite the rising significance of new search engines' role in scholars' research, little is publicly known about how they rank papers, leaving researchers to try to reverse engineer their algorithms [14]. Results suggest that Google Scholar weighs title words and citation counts, skewing results towards informative titles and frequently cited papers. Based on this finding, Beel and Gipp [14] concluded that Google Scholar is more suitable "when searching for standard literature rather than gems, the latest trends, or articles by authors advancing a different view from the mainstream," and likely to produce a Matthew effect in citation distributions (p.6). However, search outcomes may change in response to algorithmic updates or A-B testing, and thus supply little micro-level stability in this ecology. In addition, the study of Google Scholar shows that while it has aggressively expanded its coverage and constructed a strong database, it fails to have consistent quality control and maintain clear indexing guidelines [58].

While there has been a lively discussion about how to construct effective and stable

search algorithms and how to measure the significance of search outcome in the Web 2.0 era, the impact of technological development on researchers' behavior and the aggregate effect on science have been relatively neglected. Evans [47] argues that the digitization of journals has accelerated scientific consensus through quick communication, resulting in increased convergence in citation patterns; in contrast, Larivière et al. [84] argue that the dominance of top papers has declined in aggregate-levels since the introduction of modern scientific publications. Separated from the digitization of journals, search engines might roughly reduce the scope of papers that are presented; if researchers locate (and subsequently cite) the same papers from the same search engines, this could have adverse effects on what is the primary source of the literature. Recent studies have pointed to the potential myopia of science [120]. Second- and third- tier journals may be accessed with less frequency with sleeping beauties never waking up [76] (although the negative effects could be balanced with search engines uncovering the less cited papers). In addition, people's careers depend on whether academic engines show their papers on the first page or on the tenth page of search results. Those in the first couple of pages have a much greater chance of being cited, which could lead to promotions. Two other studies directly examining Google Scholar [149, 133] similarly conclude that search engines contribute to higher inequality in citation; however, these studies only include classic papers and do not appropriately address the numerous possible confounding variables that could influence researchers' citing behavior.

3.2.3 Research questions and hypotheses

The primary focus of this chapter is the examination of the manner in which the impact of social influence varies with the level of expertise in the rapidly changing academic search environments. If new academic search engines broaden the gaze of scholars and facilitate seeking relevant literature, scholars without expertise exhibit lowered reliance on on status signals. Consequently, the distribution of inter-disciplinary citations would be less concen-

trated and more dispersed in recent periods under the assumption that researchers become more reliant on search engines when the distribution of within-disciplinary citations would be relatively stable. Also, the signalling effect of peer opinion (social influence) on citing behavior would be stable in all observed years. In contrast, outsiders of the field might be more dependent on status signals because they do not have enough background to check the quality of papers. In this case, they might be more susceptible to academic search engines' new algorithms that use and present the colleague's recognition of an individual paper. If it is true, the distribution of inter-disciplinary citations would become increasingly unequal, and the effect of social influence would increase as time progresses.

3.3 Data and method

3.3.1 Data source and coverage

For the purposes of this study, the Clarivate Analytics' Web of Science (WoS) data was used for calculating citation counts to/from papers and to/from journals. The WoS data includes the Science Citation Index Expanded, the Social Sciences Citation Index and the Arts and Humanities Citation Index, while the full data set includes more than 100 million publications and over 1 billion links between papers from 1900 to 2017. From this database, only data between 1989 and 2016 was applied.The discipline categories of journals indexed by WoS were utilized in defining the journal's discipline. Since many journals are classified into more than one category, the first two categories were chosen and journals that had matching disciplines either in the first or second category were extracted. Citations made in the journals categorized in the same discipline are defined as within-disciplinary, and out of them as inter-disciplinary citations. Original research articles from journals have been selected while editorial reviews or book reviews have been excluded as part of this paper. In addition, journals in each discipline that have published papers at a rate of at least 10 papers per year, for more than five years out of 28 years of the observed period have been used to

minimize the coverage issue in the database. While this strategy proves disadvantageous in excluding recently started journals—which has increased recently—this study argues that it is more important to compare all periods under similar environments as much as possible to isolate temporal changes. Self-citations are removed from the data.

Since the main research question revolves around the effect of technology on patterns of citation and how it varies by within- or inter-disciplinary citations, it is necessary to control other factors such as citation norms [60] by limiting the analysis to a solid discipline. I present six disciplines, sociology, political science, statistics, mathematics, microbiology, and cardiology, relatively established, and traditional disciplines that experience arguably small amount of dynamic changes in terms of the scope of the discipline's boundary or citations' norms during the observed period (1999-2016), but disciplines with different levels of incorporating technology in doing research.

3.3.2 Data structure

There are several ways of organizing bibliographic data for analysis. Two of the most frequently used data structures are bibliography-based data structures and citation-based data structures. A bibliography-based data structure focuses on outgoing ties, in the sense that it identifies all papers contained in the bibliographies of a set or sample of papers; this analysis typically focuses on the distribution and characteristics of the papers cited in these bibliographies. A citation-based data structure is organized around in-coming ties, and selects a set of papers and the papers that cite them. In both cases, the challenge is to identify the appropriate pool of potential targets (in the bibliographic case) or senders (in the citation case). For example, Lynn [94] circumvents this problem by identifying articles published in well-known journals in select disciplines between 1985 and 1986, and counting the citations these papers received in the 20 years following their publication.

In order to investigate whether researchers' citations have become more expansive or more

concentrated since the rise of academic recommender systems, the data structure must meet two criteria: it must include an appropriate pool of papers that could have been cited by scholars working at a particular moment in time, and it must allow one to compare citation behavior over time. Thus, although Lynn's approach specifies a target pool of papers that could be potentially cited by scholars, it is not suitable for the research question because it does not allow the easy examination of changes in researchers' behavior over time.

Subsequently, this work proceeded to create a data structure to identify a pool of papers that could have been cited to determine if there are temporal changes in researchers' citation behavior. For the analyses, a data structure defined by the complete list of articles published in year t (where t ranges from 1999 to 2016) and their out-going citations was used. This data structure also included, for each year t, a target corpus containing all articles published in the same discipline in the prior ten years (and indexed in WoS). A ten-year window was used because most papers are cited approximately 10 years after publication, and because the possibility of older papers cited is relatively low. For example, for articles published in sociology in 1999, the target corpus contained articles published in sociology journals between 1989 and 1998; for sociology articles published in 2016, the target corpus was articles published in sociology journals between 2006 and 2015. These target corpora thus specified defined pools of arguably relevant articles that could have been cited by the articles published in a given year.

Linking outgoing citations from the lists of anchor articles to the associated target corpora provides the network of within-disciplinary out-going citations for a given publication year.[2] In addition to within-disciplinary links, inter-disciplinary citations were also identified by

[2]The percentage of citations made to the applicable target corpora is stable and low, for example, ranging between 7–8 percent in sociology. Briefly, about 70% of all citations go to sources such as news articles, datasets, books, and internet sources that are not indexed by the WoS database. Among citations to sources that are indexed in WoS, about one third of citations are made to sociology papers. In addition to these conditions, this study also limits its target corpora to the article format (which does not include conference proceedings or book reviews) written in English, and in the same discipline.

expanding anchoring papers published in year t to all papers cited outside of the discipline of this work's interest[3] that make at least one citation to papers published in the preceding ten years and in the discipline of our interest. For example, for non-sociology papers published in 2016 that have cited at least one sociology paper, the target corpus consists of articles from sociology journals between 2006 and 2015.

3.3.3 Methodology

This study commences by investigating whether the concentration rates of citation patterns become more- or less- concentrated, and whether changes in these patterns vary between within- and inter-disciplinary citations. Using the data structure for each year and discipline, measures were adopted to explicitly direct attention to the tips and tails of the distributions of counts: the concentration of citations refers to the extent to which a small number of papers absorb a large fraction of the incoming citations (the share of citations made by the top 1%), while dispersion refers to the fraction of papers that are ever cited (the percentage of papers cited at least once). Thereupon, the steps pursued in the second chapter to control possible confounding factors that might influence the measures of inequality for citation distributions were followed. The time of the out-going citation links was adjusted to time t to ensure the availability of the same publication environments as in year 1999 by controlling three conditions: the list of journals, the number of papers per journal, and the number of citations per paper. With the adjusted data, two measures of inequality were computed for each year between 1999 and 2016 for within- and inter-disciplinary citations. If the attention of more scholars concentrates to popular papers located on top search results with the increased use of academic search engines, the concentration of citation distribution will rise, and by broadening the scope of search, the dispersion of papers will also multiply. In many contexts, greater concentration implies less dispersion, though it is statistically possible

[3]This is a common way of defining inter-disciplinary citations.[83, 148]

for these tendencies to run in opposite directions.

Next, in an effort to better understand factors that impact citation behavior, a statistical model that predicts the citation counts of papers was designed and it was examined whether the effects of factors have changed over time. The primary interest here lay in whether, in the wake of new search technologies, there has been a decline in the impact of the journal a paper is published in and an increase in the impact of the papers' previous record of citations on the number of citations that it was predicted to receive in a given year. Based on our corpus[4], the number of citations received by paper j in year t was predicted by using a quantile regression with tau of 0.8. A quantile regression model is similar to the linear regression, but it estimates .80 quantile of papers cited at least once. For each year between 1999 and 2016, the model for two response variables was conducted, one for a received *within*-disciplinary citation count, and one for *inter*-disciplinary. The two main explanatory variables are Journal Influence Factor (JIF) for journal impact (the average citation count of a journal in recent two years) and the cumulative prior citations received by paper j in years prior to year t. This analytic strategy was loosely based on the assumption that temporal trend would be absent in coefficients of previous citation count estimated each year if scholars do not respond to the new signal of social influence. For standard errors, clustered bootstrap method was used with 1000 repetitions, because articles are nested in journals.

In the analyses, both response variables, the within- and inter-disciplinary citation count received in year t, was logged with a natural log after adding 1. The JIF and cumulative prior citation count were employed as two key explanatory variables in the model. While JIF represents the traditional status signal of paper's quality, the cumulative prior citation count is the relatively new status signal that becomes more noticeable along with the rise of search engines. JIF is recalculated for each year, though empirically they are relatively stable year over year. For the purposes of this study, the year t of JIF, when paper j was

[4]I use the original corpus that has not been adjusted.

cited (rather than paper j's publication year), was used because factors that influence the behavior of scholars making decisions in year t about what literature to cite was modeled. The cumulative prior citation is the total number of within- and inter-disciplinary citations of paper j through time *t-1*. For example, for papers aged three in 2016, the total number of citations that they received until age two i.e. between 2013 (age 0) and 2015 (age 2), is the cumulative previous citation. Both variables use a natural log after adding 1. Taking a log in key variables is essential in this analysis for two reasons. First, citation distributions are always highly skewed to the right, and hence the extreme values need to be adjusted. Second, with the constant growth in size of publications over time, citation count has seen a parallel rise, which might inflate the size of coefficients. By taking a log in both explanatory and response variables, coefficients were freed from the impact of inflation in citation count, and thus made comparable across time, because each additional unit of variables represents a percentage change after being logged. In addition, following Evans' [47] and Lynn [94]'s studies, three control variables were included , all that were measured on paper j: age (the number of years since published), page count, and the number of references in the paper's bibliography.

3.4 Results

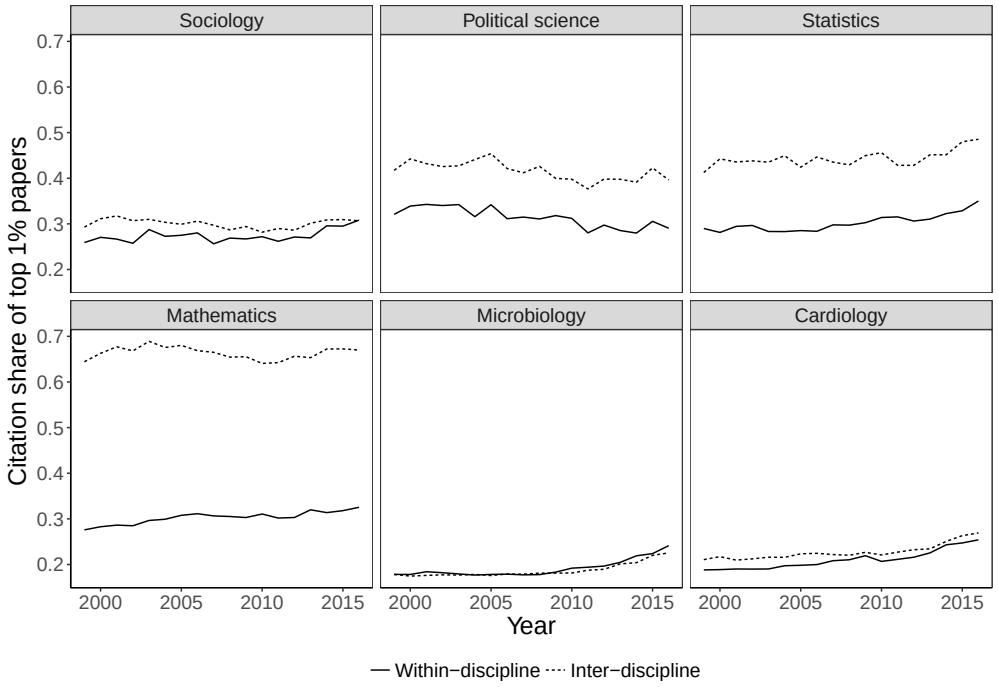
3.4.1 Macro time trend in citation distribution

The concentration and dispersion trend was analyzed to evaluate whether there is evidence of change in aggregated distributions, and whether this seems to be associated with temporal change in the citing habits of scholars. Figure 3.1 summarizes the concentration trend between 1999 and 2016 computed from the citation share of top 1% papers in each 10-year window corpus by six disciplines. Solid lines show the temporal trend of within-disciplinary citations and dotted lines represent inter-disciplinary citations. The citation share of top 1% papers differs by disciplines, but at least 20% of the whole citations were found to be

dominated by the top 1% papers, which again confirmed the nature of high concentration of star papers [41]. Moreover, in all disciplines except microbiology, the concentration was higher in inter-disciplinary citations. While in the case of sociology and cardiology, the difference between within- and inter-disciplinary citations was relatively small, political science and statistics showed a larger gap between the two types of citations. In mathematics, the concentration of inter-disciplinary citations turned out to be much higher than in within-disciplinary ones.

However, no changes were observed in temporal trends from Figure 3.1. The gap between within- and inter-disciplinary citations was stable or showed no clear patterns during the observed period.

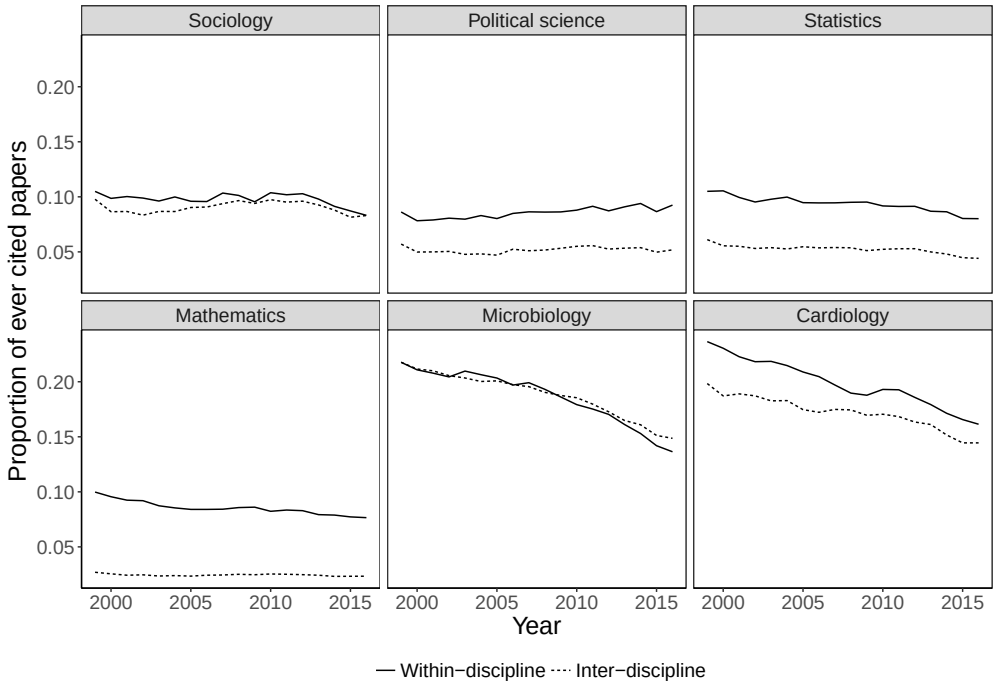
Figure 3.1: The citation share of top 1% papers between 1999 and 2016 by six disciplines; solid line is for within-disciplinary citations while the dotted line is for inter-disciplinary citations.



While Figure 3.1 focuses on the top of the distribution, Figure 3.2 shows the tail of the distribution by illustrating the temporal trend of the proportion of ever cited papers between 1999 and 2016. As exhibited, the proportion of ever cited papers was as high as .20 and as low as .02, which indicates that, overall, only the limited number of papers in the corpus are cited in each year. Again, with the exception of microbiology, inter-disciplinary citations had a lower proportion of ever cited papers than within-disciplinary ones. Also, similar to the pattern of concentration, there were no distinguishable changes in temporal pattern in Figure 3.2. Combining the two findings, while inter-disciplinary citations were likely to reach at the

smaller subset of papers than within-disciplinary citations, its relative ratio has remained stable over time.

Figure 3.2: The proportion of ever cited papers between 1999 and 2016 by six disciplines; solid line is for within-disciplinary citations while the dotted line is for inter-disciplinary citations.



So far, this study has discovered that although inter-disciplinary citations are more concentrated to the top and are less dispersed overall in comparison to within-disciplinary citations, the macro temporal trend has not been changed before and after the popularized use of academic search engines. This result is not consistent with previous observations about search engines and following changes in distributions. For example, multiple studies have shown a 'Google Scholar effect', where more citations go to old and popular articles

[149, 133], which implies that citations concentrate on a few popular papers since people rely on search engines; however, analysis conducted as part of this work, does not reach similar conclusions.

3.4.2 The influence of previous citation count and JIF

Figure 3.3 shows coefficients of the previous citation count from 1999 to 2016 of six disciplines with 95% confidence interval. As an example of interpreting the model, coefficient 0.3 means that each percentage increment in previous citation count increases the .80 quantile of the predicted received citation count by 0.3%. Red lines show the coefficient from models predicting within-disciplinary received citations, and green lines are for inter-disciplinary ones. First, all coefficients and their confidence intervals are above 0, which means that previous citation count has a positive impact across disciplines and time, and is statistically significant with .05 p-value. For sociology, political science, statistics, and microbiology, coefficients of within-disciplinary citations remain stable or slowly increase over time, while those of inter-disciplinary citations change more dynamically. It indicates that inter-disciplinary citations become more responsive to previous citation count than within-disciplinary citations roughly since the mid-2000s. In mathematics and cardiology, no significant change in the difference between within- and inter-disciplinary citations was found.

Figure 3.3: Coefficients of previous citation count by discipline between 1999 and 2016; red line is for within- and green line is for inter-disciplinary citations; The analysis of only the positive citation count.

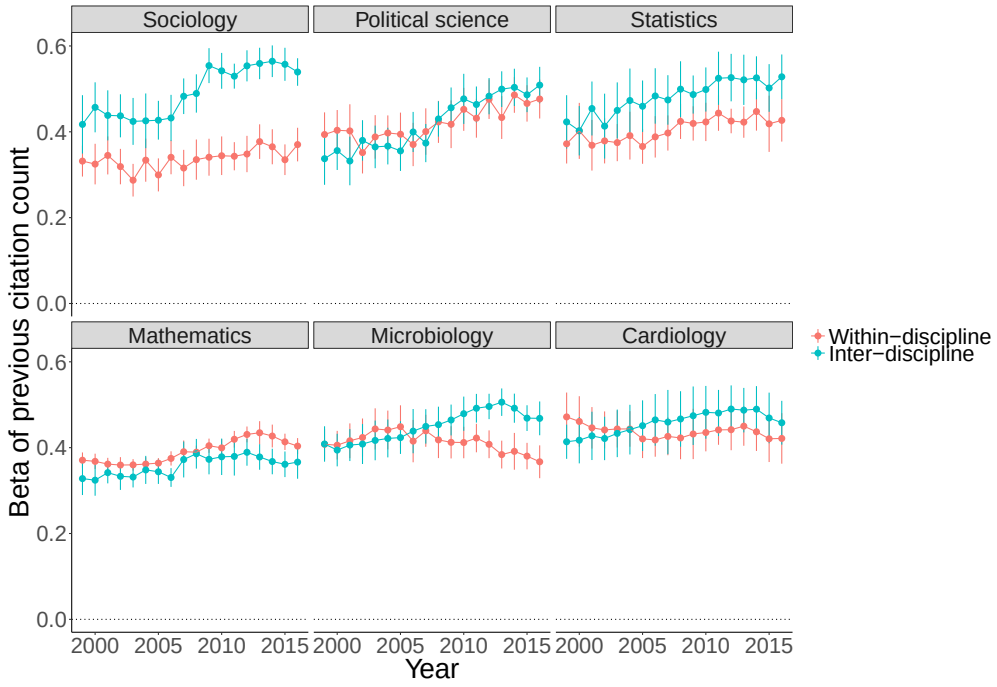
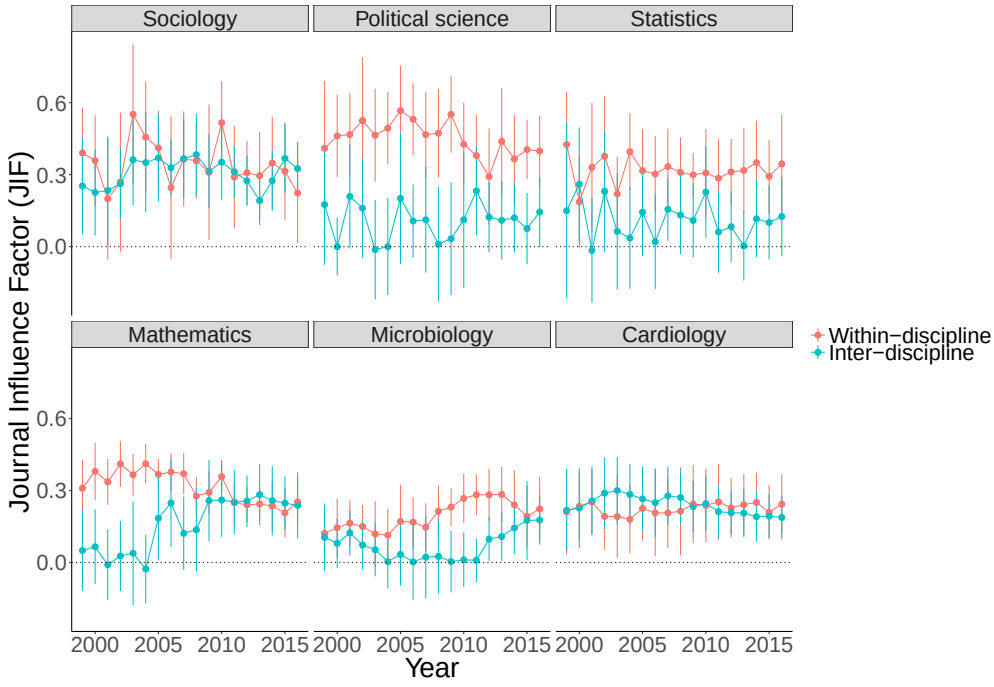


Figure 3.4 presents the coefficients of JIF for within- and inter-disciplinary citations over time. Confidence interval of JIF coefficients is generally wider than coefficients of previous citation count, which reconfirms the large variance of the quality of papers published in the same journal (e.g., [132, 85]). However, enough evidence was not found to support the decreasing impact of JIF in all six disciplines in digital age as has been argued by previous research [93] [88] [2].

Figure 3.4: Coefficients of JIF by discipline between 1999 and 2016; red line is for within- and green line is for inter-disciplinary citations; The analysis of only positive citation count.



3.5 Discussion and conclusion

This research focused on studying how scholars' expertise of the field relates to their sensitivity to social influence in the context of the development of literature search technology, and aimed to examine its following impacts on the scope of science. To investigate these questions, macro citation distributions over time in six arguably stable and well-established disciplines was examined. It was found that less literature is being cited (i.e., more zero cited papers) and there exists a higher concentration of citations of the star papers for inter-disciplinary citations in comparison to within-disciplinary citations. However, these results

do not contain evidence to determine whether there has been any temporal change in the amount of gap between insiders and outsiders of the field.

In the second set of results, I found that the impact of previous citation count on inter-disciplinary citations has increased in sociology, political science, statistics, and microbiology, while the impact on within-disciplinary citations stayed relatively stable during the observed period. This finding suggests that more scholars citing outside of their fields are influenced by the previous citation count, the information that was explicitly provided by new academic search engines such as Google Scholar, which is consistent with the idea that scholars without expertise rely more on the search engines and thus become more vulnerable to the influence of social recognition. However, its impacts differ by disciplines. One possible explanation for disciplinary differences is the broad disciplinary norm regarding which search engines to use. For example, in cardiology, it is likely that non-expert scholars citing cardiology are also from the health category, and the major search engine used in health is known to be PubMed instead of Google Scholar, which does not explicitly provide information on previous citation count.

Interestingly, outsiders' higher reliance on previous citation count does not necessarily relate to higher concentration of citation distribution. Although the impact of previous citation count has increased for inter-disciplinary citations as time progresses, the inequality of inter-disciplinary citation distribution has not increased faster than the within-disciplinary one. I suggest two possible explanations on these intuitively contradictory findings. First, the increased impact of previous citation count might not be significant enough to create the distinctive temporal change in the macro trend. Another possibility is that search engines might foreground highly cited papers, but at the same time, by expanding the scope of search results, it might offset the inequality of citation distribution.

Some might argue that increased influence of prior citation count might be due to specialization or concentration of the field rather than the effect of search engines. However, if

one field has received more attention than another, all papers published in a journal on this field will have an equal probability to be spotlighted rather than a few highly cited ones. Thus, instead of saying that the specialization of the field drives behavioral changes, it would be more reasonable to argue that search engines might facilitate specialization of the field or concentration on a dominant theory, because few highly cited papers will dominate search engines and have a higher chance to be seen, and therefore unconsciously influence scholars when they conduct research.

While I have mainly argued that the technological development in search engines drives transformation of researchers' behavior which it is suggested by changing citation distributions over time, this change cannot be solely attributed to the effect of technology. There are other possible scenarios that might lead to researchers' change in behavior. For example, conferences, twitter, facebook, and related technologies may highlight individual papers and might detach them from journals. While I cannot completely distinguish this possibility from the effect of search engines, the increased influence of previous citation count still suggests an underlying impact of increased use of search engines because those social network services do not provide or use citation count in searching or ordering results.

Also, classic papers in the analysis such as works written by Karl Marx or Thomas Kuhn that have explosive influential power and argued to be cited much more in 'Google Scholar era' [133, 149] have been ignored by this study. The preferential attachment mechanism might be a better explanation if I only look at papers that have received several thousand of citations. However, I have decided to exclude them for two reasons. First, as I explained in the data structures, it is necessary to identify a pool of papers that would have been cited; if I decided to include classic papers, a pool of papers would have covered publications from early 20th century. Second, the intention of researchers citing recent papers and classic, foundational ones might be different; researchers might cite more fundamental papers when they need to bring authority in their research to persuade readers and to emphasize the

significance of the study [60]. Thus, I believe that it is more important to separate the two different kinds of citations and isolate the effect of technology than to include star papers to answer my main research questions.

I reserve the judgment on whether the transforming search behavior of non-expert scholars is beneficial in sustaining a healthy academic environment or vice versa. The literature roughly doubles every 20 years, and as this expansion continues, it will be increasingly difficult for scholars to keep up with even their own fields. While search algorithms will be needed for assisting scholars to guide literature searches, there is also a risk that built-in, algorithmic biases, as well as human behavioral biases, would impact what science is actually found, read, cited, and communicated to the public. Thus, I have argued that it is important to understand and monitor how search engines change the manner of conducting research, particularly, as these recommendation algorithms become more common for every day research. Based on the findings so far, the new technology does not passively assist researchers' job in searching literature, but may actively interfere in researchers' evaluation of which papers are more important to be cited in their scientific work.

Chapter 4

IS JOURNAL STILL MEANINGFUL AS THE CREDENTIAL SYSTEM IN NEW SEARCH ENVIRONMENTS?

4.1 Introduction

[1]Prior to the advent of digitization, journals played a central role in the scientific process, both evaluating research (through peer-review and the editorial process) and serving as an efficient filter for the search process. Before the use of integrated academic search engines, both the physical location of journal archives on library shelves and personal subscriptions facilitated individual scholars to be familiar with the research published in a particular set of journals—all the while increasing the difficulty of even learning about, let alone gaining access to, scholarship published in other outlets. Digitization made the process of accessing a known paper far easier, but it initially did little to improve search, and the curatorial role of journals remained. A consequence of the pre-academic search engine is that papers published in high profile or well-distributed journals were likely to be seen, and hence cited, more than papers published in journals with smaller subscription bases or lower reputations.

Besides being broadly read and cited, papers published in high status journals sent a positive signal regarding its quality and the significance of its research to readers. Because the size of scientific output is roughly doubled in every nine years since the beginning of modern science [20], readers are always short of time to review all published articles, especially when they lack expertise in the field. Therefore, scholars needed the credential system to sort out articles that are worth investing their limited time into to read and understand, which is

[1]This work is from the collaborative project undertaken with Jason Portenoy, Katherine Stovel, and Jevin West.

similar to the example of information asymmetry in the labor market [139, 34, 18]. The most frequently used credentials that readers relied on is the status of journal, generally measured by Journal Influence Factor (JIF), a simple metric reflecting the average number of citations to recently published articles. As consumers of research articles screened which papers to concentrate on based on journal status, more authors, particularly young generation scholars who have not achieved significant career progress to prove themselves, strove to publish in high status journals and signal the quality of their research.

However, the role of the journal as the credential system of the quality of the papers included in it has been in question as the rise of integrated academic search engines begins to change the way of searching and screening papers to read. With the new technology, papers' positions in various electronic archives (and the algorithms used to access these archives) may be decoupled from the journal they are published in, while paper relevant features (such as prior citation, or authorship) may play a role in increasing their visibility to readers. As the introduction of new search technology diversifies the manner of sending signals about the quality of papers to readers, previous research has attempted to examine changes in the impact of the journal as the credential system [93, 88, 2].

The main methodological problem in estimating the effect of a journal is that published academic articles are inherently nested in journals, and thus it is impossible to separate the influence of its journal status from an individual paper's quality. Once papers are published in a journal, readers always evaluate the quality of articles under the influence of the journal status. One solution to this problem is to set a counterfactual experiment that enables readers to evaluate articles with and without a journal name. However, it is especially difficult for setting up an experiment with academic articles, even with the development of online tools, because reproducing the complex scientific arguments and its nuanced contribution is almost impossible.

As a solution to this problem, this study suggests a creative approach to measure the pure

impact of journal status after controlling individual paper's quality in a quasi-experimental situation, then trace the temporal changes in the role of journals as the credential system along with the emergence of integrated academic search engines. I found that 'arXiv.org' (arXiv), the preprint service for academic articles founded in 1991, is close to a quasi-experimental setting where researchers read and make decisions to cite papers without knowing the journal name. I counted the received citations made to pre-published articles in arXiv and used them as a proxy measure of the paper's quality. This measure is arguably independent of journal status since it is based on the data that precedes the information regarding the publishing journal. Moreover, the citation count can be considered a socially driven proxy measure for the paper's quality based on the assumption that a paper is worth citing if it has already been cited [94].

To create the proxy measure, I first collected the information of all papers pre-published in arXiv in a certain period, then found out citations made to these papers. When these papers were later published in a journal and cited with the printed journal name, this information was collected by linking a paper's ID in arXiv data to the one in Microsoft Academic database. I chose Microsoft Academic Search out of other search engines that have bibliographic information owing to it being the only academic search engine that allows public access. I extracted papers uploaded in arXiv and citations made towards these papers between 1997 and 2016. The disciplines that I analyzed are high energy physics (phenomenology), astrophysics, condensed matters because papers in these disciplines were consistently uploaded in arXiv since 1997 according to the statistics provided by arXiv.org.

The journal's role as the credential system was investigated in two ways. First, I study changes, if any, in the quality of papers that pursue journal publication. Because arXiv partly replaces the journal's role as the only platform to present research results, researchers might feel less necessity to publish papers and postpone going through the long review process of a journal. Particularly, I hypothesize that this tendency would be stronger for papers that

are read and cited many times in arXiv, and thus already acknowledged their contribution to the field. If authors of articles that are cited many times in arXiv are less likely to pursue journal publication, it can be argued that said journal is no longer the only means of receiving credentials about quality. Previous literature supports this hypothesis by suggesting that there might be a "quality bias" toward arXiv publications meaning that better papers and high impact authors are more likely to appear in arXiv [39, 54]. I further aim to examine this hypothesis by using a more systematized statistical methodology.

The second research question asks what the real impact of journal status on an individual paper's subsequent citation count is post journal publication, and whether it has changed over time in the new search environment. While the distribution of received citation count of papers published in the same journal is very dispersed [85], Journal Influence Factor (JIF) has always been the best measure to predict the received citation count (e.g., [44, 118, 19]). However, the previous literature failed to separate the confounded effect of individual paper's quality from the journal effect. By reexamining the journal impact after controlling a paper's quality and tracing its temporal changes, I explore whether the journal's credential role is as persistent as before the academic search engines were popularized.

4.2 Literature review

4.2.1 The influence of journal status on the citation count

Previous research has confirmed that the status of journals, usually represented by JIF has been the primary criterion in judging the quality of individual papers. JIF turned out to be an essential factor in various contexts predicting the received citation count [44, 118, 142, 19]. Particularly, for recently published papers that have not had ample time to be fairly evaluated, JIF performs better in predicting their following citation count [1]. Thus, the early citation is correlated with JIF, and the early citation count explains more than half of the variation in cumulative citations received over a more extended period, which supports

the decisive role of JIF in a paper's received citation count [137].

So far, there is evidence that the attachment of papers to journals has reduced since the wake of integrated search engines. Lozano et al. [93] show the weakening relationship between the impact factor and papers' citation count. They compare R-square of the model predicting the citation count based on JIF over time and find that R-square has decreased, indicating that JIF's power to explain citation count decreases. As a complementary analysis to Lozano et al. [93], Larivière et al. [88] supports the hypothesis of the decline of journals by demonstrating that the percentage of citation share made by elite journals such as Science or Nature is declining. The research team in Google also supports this finding [2]. According to research they conducted by computing the fraction of the top-cited articles published in non-elite journals and the fraction of the total citations to non-elite journals the impact of non-elite journals has increased. Specifically, "the percentage of citations to articles in non-elite journals went up from 27% in 1995 to 47% in 2013. Six out of nine broad areas had at least 50% of citations going to articles published in non-elite journals in 2013." [2] Although journal's function of evaluating research as a gatekeeper of checking the quality of papers based on peer-reviews remains [113], and it is not yet replaceable by other tools such as arXiv [33], its role as a screening filter for search processes has been threatened since the popularization of academic search engines.

However, due to JIF's simplicity in evaluating a paper's quality, it still yields power on making career decisions and distributing grants or perks [162], though JIF was initially developed for librarians to decide which journals to subscribe [9]. The often used two-year citation window for computing JIF is considered especially harmful in hindering the rise of innovative articles, which usually take more time to be evaluated and cited [136]. Regardless of the warnings toward using JIF [132, 99, 136] and regardless of the manifestos [62, 85], it seems it is not yet easy to stop using JIF in evaluating research.

Regarding methodology, the previous attempts to measure the impact of journal influence

on subsequent citations of an individual paper have not succeeded in sorting out the impact of paper's quality. Lariviére and Gingras [82] provide one solution to this problem by discovering a quasi-experimental situation. They find cases of publications that submitted the same manuscript to two different journals and compare whether the received citation count differs depending on the journal's status. In the results, Lariviére and Gingras show that papers published in journals with higher JIF are cited more than ones in lower JIF even though they have the same manuscripts. Although this research suggests a possible way to separate the effect of journal status from the paper's quality, the finding is limited to a few unethical outliers, which is hard to expand in general.

4.2.2 The beginning of arXiv and its use in scholarly communication

arXiv was first founded in 1991 mainly for High Energy Physics researchers to facilitate scholarly interaction through the unified online-based depository without access fee. While the purpose of arXiv has been shared by researchers from other disciplines as well such as condensed matter physics, astrophysics, mathematics, computer sciences, statistics, economics, etc., the extent of the use of arXiv varies by disciplines [86]. Overall, the total number of pre-published papers and citations towards these papers has continuously increased [103, 116]. Moreover, preprint services similar to arXiv such as bioRxiv for biology researchers have been launched. Scholars upload research papers to the preprint services to circulate their results earlier without waiting for the journal review process that usually takes several weeks and also to reach a broader group of audience who do not have access to journals with a subscription fee.

Since arXiv has evolved as a favorite tool for scholarly communication from an archiving tool for a small number of scholars [86], citing preprint version of papers in arXiv has also become a common practice.[2]. While it is difficult to measure the quality of papers because

[2]However, in some of the earlier stage of pre-published server such as chemistry preprint server, editors

there has been no clear consensus on defining their quality[71, 142], I choose to use the citation data accumulated of preprint version of papers as a way of seeking a proxy measure that is independent from journal status. By using this measure, I seek to find out the real impact of a journal on citation count after controlling the individual paper's quality.

Owing to the launching of arXiv, now scholars have one more conduit to present their research with minimal review process and still reach a wide range of audience either with or without enough resources for journal subscription. Also, integrated academic search engines foster the growth of preprint services by counting papers from them into their search results, which expands these papers' visibility to the audience. While scholars before-arXiv relied on academic journals as a way to present research and get certification of the research quality via peer-reviews, now the role of journals in the current academic environment has been limited to maintain the quality of research. However, only ambivalent evidence has been found so far in support of the decreasing role of journals in comparison to arXiv. Although the role of preprint archives in developing scientific discourse has been stronger [54], particularly in discussing developing topics [69], people cite journal publications as soon as the article is published, which still sustains the purpose of journals [64, 86].

4.3 Data and method

4.3.1 Data

For the purposes of this dissertation, all papers that have been uploaded in arXiv between 1996 and 2016 were collected and categorized as high energy physics - phenomenology (Hep-ph), astrophysics, and condensed matter. These three disciplines were specifically used because they have actively and continuously used arXiv during the observed period and for these reasons, studied in previous research [54, 103, 105]. Subsequently, I linked these papers

in high-status journals did not allow authors to cite preprint version of papers because they have not been peer-reviewed [25, 26]

to Microsoft Academic Graph (MAG), the database of Microsoft Academic Search. Microsoft Academic Search is an integrated academic search engine similar to Google Scholar that helps scholars search relevant literature by searching keywords. The search results include necessary bibliographic information as well as incoming and outgoing links to citations. Unlike Google Scholar, since Microsoft database opens its APIs to users, bibliographic details of papers and citation links in the database could be utilized. Linking the database of arXiv to Microsoft enables to trace the citation information of papers during the period between when it was first uploaded in arXiv and after it was published in a journal. Especially, using MAG citation data provides a more comprehensive coverage than the previously used Web of Science (WoS) database (e.g., [86]) because it includes citations in between preprint versions of papers.

I used the following strategy to match arXiv metadata to papers in the MAG: I first matched as many records as I could for which the arXiv metadata included a journal publication Document Object Identifier (DOI). For the remaining arXiv articles, the matched were based on title, querying the Academic Knowledge API for MAG.[3] [4] arXiv papers which could not be matched were excluded at this point. I also excluded MAG records that have no associated references. Most cases without reference lists that was hand-checked had incomplete information regarding title, author, published year, and/or source, which made it impossible to link these lists to available bibliographic information. In this study, such papers are not qualified as proper research articles. In addition, I removed records for which the publication date in MAG is before the arXiv upload date because I only aimed to analyze papers uploaded in arXiv first and later published in a journal. Finally, when there is more than one matched record for an arXiv paper, I went through systematized steps to finally match one paper in MAG to one paper in arXiv. These steps are summarized in Appendix

[3]https://labs.cognitive.microsoft.com/en-us/project-academic-knowledge

[4]The validity of this approach is verified in Thelwall (2018)[144].

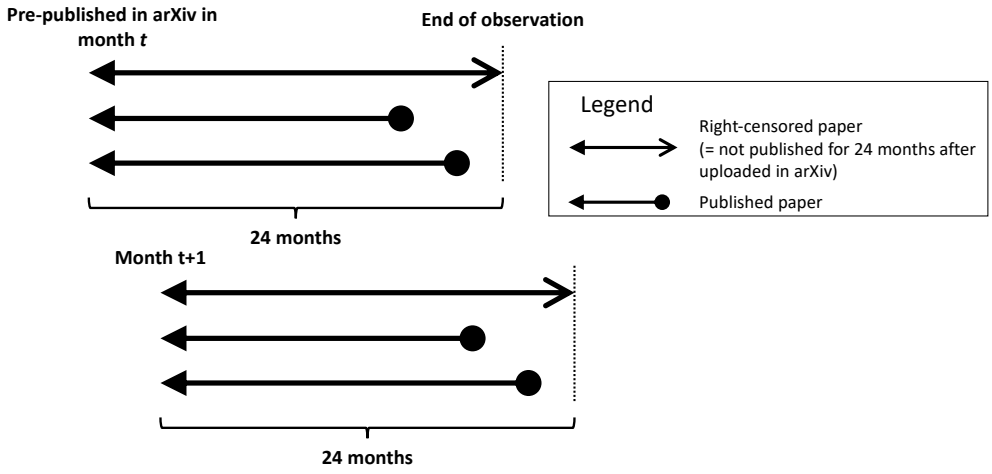
C.

4.3.2 Data structure and method

Based on the extracted data, I created two separate data structures to answer two research questions. The first data structure is to examine whether there have been any changes in the quality of papers that pursue journal publication. If papers are read and cited enough before they are published in journals after being uploaded in arXiv, do these papers still focus on being published in a journal?

To examine the research question, I collected all papers uploaded in arXiv at month t and traced these papers' status change for the following 24 months. Once a paper is published in a journal, then this observation concludes. If a paper is not published after 24 months, then I treat these papers as right-censored, and I stop the observation. I repeat this process for all the months between 1997 and 2012. Figure 4.1 summarizes the data structure.

Figure 4.1: Data structure of survival analysis

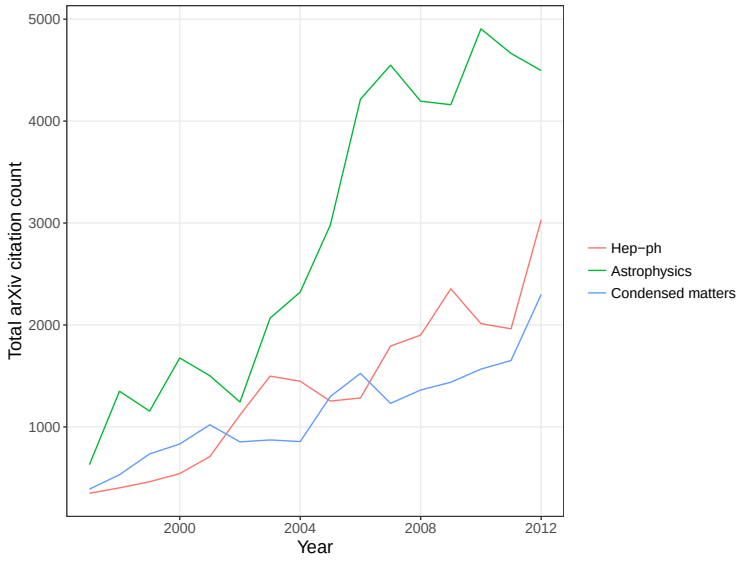


I used Cox's proportional hazards regression model (survival analysis) to analyze the data set. Because this model is able to associate the survival time (time-to-event) with other covariates, I applied the method to understand how the cumulative citation count of papers influences the hazard ratio depending on the amount of time in arXiv before journal publication. Thus, the response variable of the analysis is the survival time, and the primary explanatory variable is the cumulative citation count in arXiv in month t. The cumulative citation count is, therefore, a time-varying variable, and it is logged in the analysis. I adjusted the inflation[5] of the cumulative citation count to let the same citation count in different years mean the corresponding indicator of paper's quality. To compute the inflation rate, I counted the total number of papers aged 0-4 and the total citation count made to these papers, then calculated the average citation count per paper. The average citation count per paper in each year is used to adjust the yearly inflation rate.

I also included the interaction term of temporal trend and the cumulative citation count to see whether the effect of the cumulative citation count has increased or decreased over time. The temporal pattern is measured by the number of total citations made to papers uploaded in arXiv and not published in a journal, and aged between 0 and 4. Figure 4.2 shows the temporal trend of total citation count between 1997 and 2012 in three disciplines of interest. While the trend fluctuates a little, the total citation count has continuously increased between 1997 and 2012. The increase in astrophysics is relatively faster than the other two disciplines. I used the total citation count instead of the number of years that have passed since the beginning year of arXiv because the temporal trend might vary depending on how much disciplines are actively involved in utilizing arXiv in research. As a measure of the size of their research activities, I included the total citation count in the model.

[5]Because arXiv has been growing since 1997, the number of published papers and citations has increased. Thus, the meaning of receiving one citation in 1997 is not the same as one in 2016. To make each citation have the same significance, I decided to adjust the inflation.

Figure 4.2: Total citation count made to papers uploaded in arXiv and not yet published in a journal, and aged between 0 and 4



The second data structure answers the research question regarding the affects of the status of the journal on a paper's performance, measured by citation count. While it is evident that papers published in high-status journals are more likely to be cited than ones in low-status journals, I mainly focus on whether the effect of journal status on paper's performance has reduced in recent years.

As I stated in the introduction, the main contribution of our approach is its use of the citation count made between the date the paper was first uploaded on arXiv and published in a journal, which might signal the quality of the paper. Following this idea, the response variable is the total received citations of a paper for 36 months after publication, which is a measure of the paper's performance.

The main explanatory variable here is the measure of journal influence. I used ArticleInfluence score developed by West et al. [154] as a measure of journal influence. ArticleInflu-

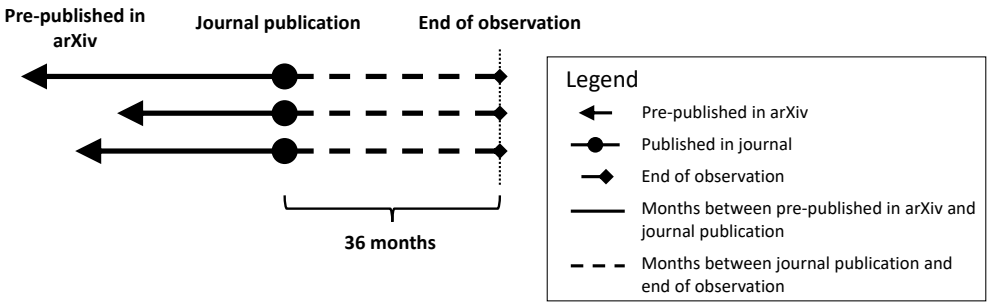
ence is a network-based journal-level method for ranking journals based on the Eigenfactor algorithm and normalized by the size of the citing journal. [6]

Moreover, there is an important control variable, the citation count while in arXiv—a measure of a paper's quality separated from journal status. Because papers have a different amount of time existing in arXiv before journal publication, I weigh the citation count by the square root of the number of months in arXiv.

Among all papers uploaded in arXiv, I only chose papers that have been published in journals . Then, I traced their citation records for 36 months after being published. Subsequently, I collected papers published in journals between 1998 and 2013. Figure 4.3 below summarizes the data structure.

[6]Specifically, there are three advantages to using ArticleInfluence instead of JIF as a better representation of how people perceive journal status [154]. First, ArticleInfluence uses the Eigenfactor algorithm which values citations differently depending on cited journals. A citation from a high-status journal is valued higher than one from an anonymous journal, while JIF equally counts the two citations. Second, ArticleInfluence weighs a citation by the number of references in a source paper; a citation out of ten references is considered heavier than a citation out of fifty references. Third, ArticleInfluence uses a five-year citation window, while JIF usually uses a two-year citation window. However, the correlation between ArticleInfluence and JIF still turn out to be high.

Figure 4.3: Data structure of hurdle model



I divided the data set into two: the first data set includes papers that stayed in arXiv for more than six months and less than or equal to 12 months before journal publication, and the second data set includes papers that stayed for more than 12 months. I excluded papers in arXiv which were less than or equal to six months due to the few citation counts for those papers. For example, in condensed matter, 92% of papers staying six months received no citations. This percentage is 85% for astrophysics and 76% for high energy physics. Among papers that stayed relatively long enough in arXiv to be evaluated by readers, I again divided the data set to distinguish papers that have collected different amounts of information regarding its quality. For papers that stayed less than 12 months, the signal effect of citation count made during arXiv would be weaker in comparison to papers that

stayed for more than 12 months. Another reason for separating the data set is to control the different aging trajectories of papers. The distribution of received citation counts after being published has a pattern that usually peaked within a few years and then decreased [23, 122]. My strategy to divide the dataset helps gather papers with similar stage of aging to minimize its effect. The approximate ratio of observations between the first data to the second one is four to one.

For the statistical analysis, I used a hurdle regression model with negative binomial distribution. I chose a hurdle model because the process of generating none of the citation counts after being published consists of two steps. If a paper is considered to possess a good enough quality to be cited, it will be cited. Then, among those papers that satisfy the basic paper quality standard, there will be papers with many citations or papers with few citations. Because the distribution of citations in the second step shows a positively skewed distribution, I chose a negative binomial distribution which allows more dispersion than Poisson. I used clustered standard error by journal computed using the bootstrapping method with 1,000 repetitions.

With a hurdle regression model, I tested three nested models. The first model included an explanatory variable, journal influence[7], and two basic control variables, the number of months stayed in arXiv and the total citation count in arXiv for a given year t. The second model added the key control variable, citation count while in arXiv. In the third model, the interaction effect of journal influence and the total citation count in arXiv which increases as time goes forward was added to trace the changing impact of journal status over time. All explanatory and control variables were logged.

Both data structures in Figure 4.1 and Figure 4.3 have the advantage of controlling other effects such as the prestige of author or author's affiliation [142] that might influence the outcome variables. Because I trace the citation records of the same paper before and af-

[7]While I use ArticleInfluence as a measure of journal influence, I use the term "journal influence" from now on to avoid possible confusion regarding this measure as an article-level metric.

ter journal publication, its author- or paper-related information remains the same as citing researchers during the complete observed period. Thus, assuming that the effect of possible alternative explanations is relatively stable for preprints and journal publications, the data structures have the same effect with controlling the alternative explanations of the changes in response variables in the model, which helps focus on interpreting the effect of key explanatory variables.

4.4 Results

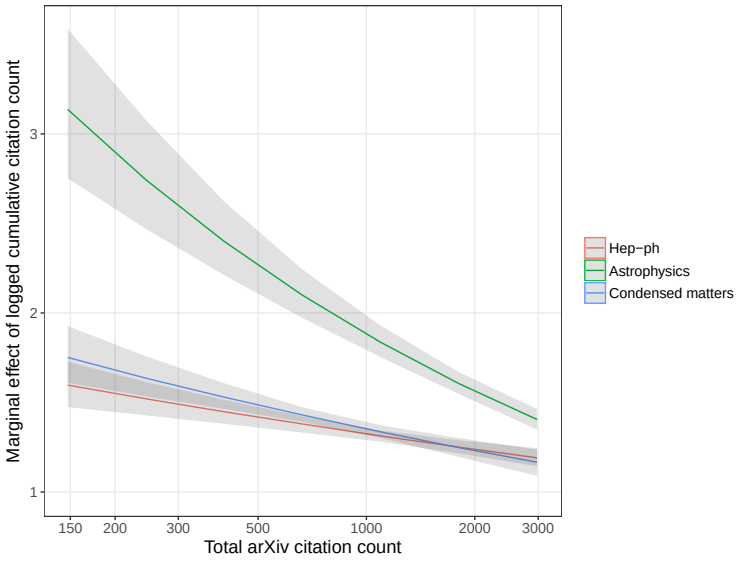
4.4.1 Survival analysis

Figure 4.4 presents the marginal effect of logged cumulative citation count by total arXiv citation count in three disciplines. The marginal coefficients are from the interaction term included in the results of Cox's proportional hazards regression model. The details of the full models are included in Appendix C. The marginal effect indicates the effect of the logged cumulative citation count on the hazard rate (the probability of being published in a journal) when the logged cumulative citation count increases one unit after controlling other variables.

In all three disciplines, the marginal effect of the citation count decreases as the total arXiv citation count increases.[8] The decreasing marginal effect implies that the influence of cumulative citation count has been weaker as the use of arXiv has expanded. Among three disciplines, the decrease is the most dramatic in astrophysics in comparison to the other two disciplines that see a relatively smooth decline. Considering the use of arXiv has been popularized in astrophysics faster than in the other two disciplines (Figure 4.2), researchers in astrophysics might feel less need to pursue journal publication and rely on arXiv as a channel to present their research.

[8]The finding is robust when I use the number of years passed after 1997 instead of total arXiv citation count to measure the temporal trend. It also remains robust with the cumulative citation count without inflation adjustment.

Figure 4.4: Marginal effect of logged cumulative citation count by disciplines from the survival analysis in Appendix C. The shaded areas represent 95% confidence interval.



4.4.2 Hurdle model

Figure 4.5 summarizes the results of the count part of hurdle regression analysis by three disciplines. The figure compares the coefficients of journal influence before and after controlling the citation count in arXiv. In each panel, the left set of coefficients analyzes the set of papers that have stayed in arXiv between 6 and 12 months, and the right one examines the set longer than 12 months. Lines on the top of the bars show the 95% confidence interval of coefficients with robust standard errors.

In all models across disciplines and data sets, the coefficients of journal influence are above 0. It indicates that when papers are published in a high-status journal, it is more likely to be cited in the next three years. The impact of journal status as a signal of the paper's quality turns out to be stable in all situations. However, when I compare the coefficient in Model 1 (red bars) and Model 2 (blue bars), the one in Model 2 is always smaller than Model 1.

This finding indicates that the impact of journal status is overstated if paper's quality is not controlled in explaining the positive citation count. In other words, the previous attempts to measure the effect of journal status automatically over-predicts it as their study design lets journal status appropriate the quality of individual papers. After separating the two factors, the impact of journal influence is significantly lower than before.

Also, papers that have been in arXiv longer show a more significant decrease from Model 1 to Model 2 in all three disciplines. Under the assumption that papers can be more accurately evaluated before journal publication when they stay in arXiv longer, the accurate estimation of a paper's quality reduces the effect of journal influence even greater.

Figure 4.5: Coefficients of journal influence from the count part of hurdle regression analysis by three disciplines. Y-axis shows the coefficient of journal influence before (Model 1) and after (Model 2) controlling the citation count in arXiv. Red bar represents the result of Model 1, and Blue bar for Model 2. X-axis shows the type of data set. Lines on the top of the bar indicate the 95% confidence interval with robust standard error.

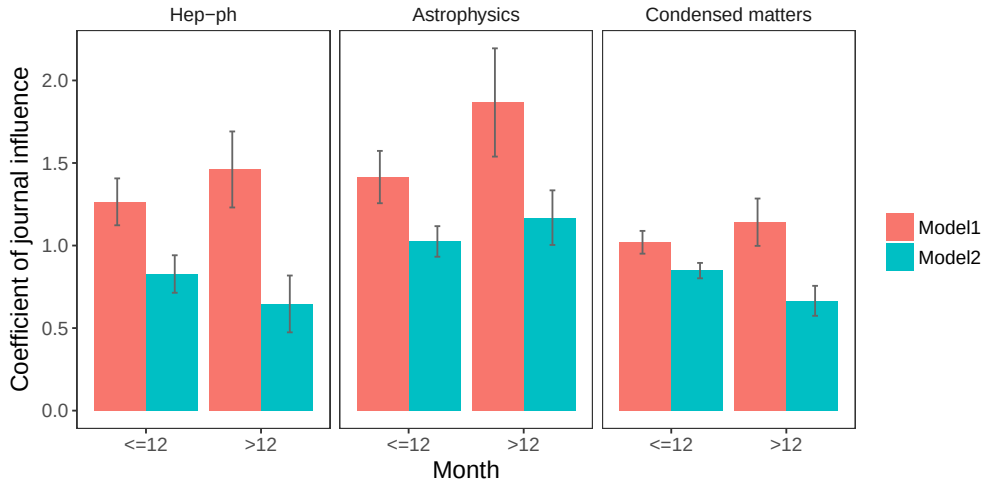


Figure 4.6 summarizes all the results of the same analysis, but the hurdle part that

estimates the probability of being ever cited or not in the first three years after publication. Similar to Figure 4.5, all coefficients across models, disciplines, and data sets by months are above 0. This finding indicates that papers in high-status journals are more likely to be cited at least once, and it is statistically significant with a p-value .05. However, the differences in coefficients between Model 1 and Model 2 is small, which is the main difference in comparison to the results of count part (Figure 4.5). The slight difference in coefficients of journal influence after controlling individual paper's quality indicates that the impact of journal influence on whether its paper is cited or not is less confounded with a paper's quality. In other words, regardless of the paper's quality, the name of the journal might give some papers a higher chance of being cited at least once.

Figure 4.6: Coefficients of journal influence from the zero part of hurdle regression analysis by three disciplines. Y-axis shows the coefficient of journal influence before (Model 1) and after (Model 2) controlling the citation count in arXiv. Red bar represents the result of Model 1, and Blue bar for Model 2. X-axis shows the type of data set. The line at the top of the bar is the 95% confidence interval with robust standard error.

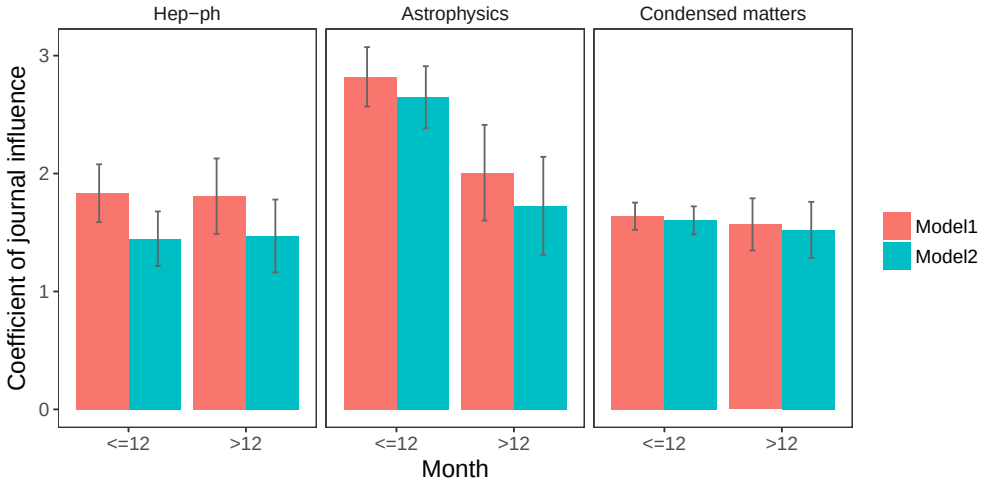
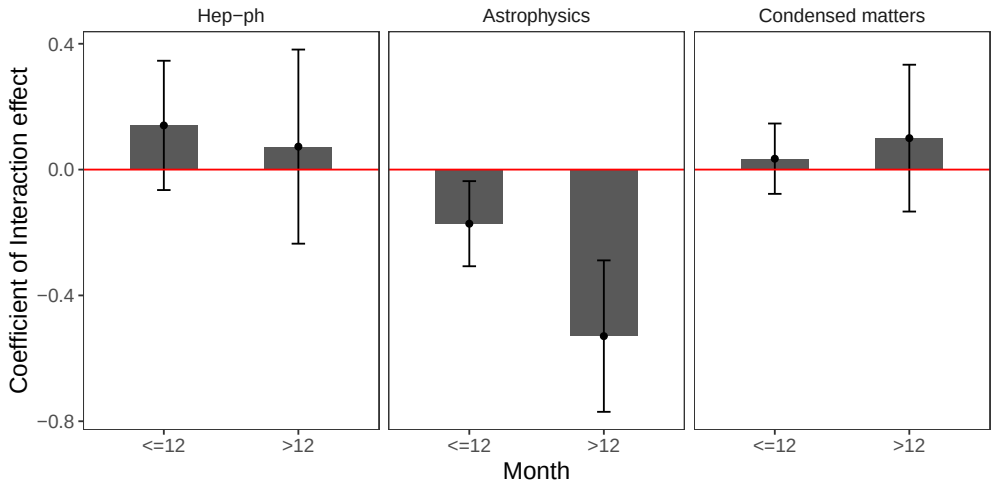


Figure 4.7 and 4.8 summarizes the results of count and hurdle part of Model 3, which includes the interaction effect of journal influence and the time trend in addition to Model 2. Figure 4.7 shows a mixed finding without a clear trend across disciplines. In Hep-ph, the interaction effect is positively statistically significant for the data set staying in arXiv for less than 12 months, but it is not so for those staying longer than 12 months. In astrophysics, the effect is negatively statistically significant, and in condensed matters, there is no sign of meaningful statistical change in the effect. Positive coefficient of interaction effects indicate that the impact of journal influence has been growing over time.

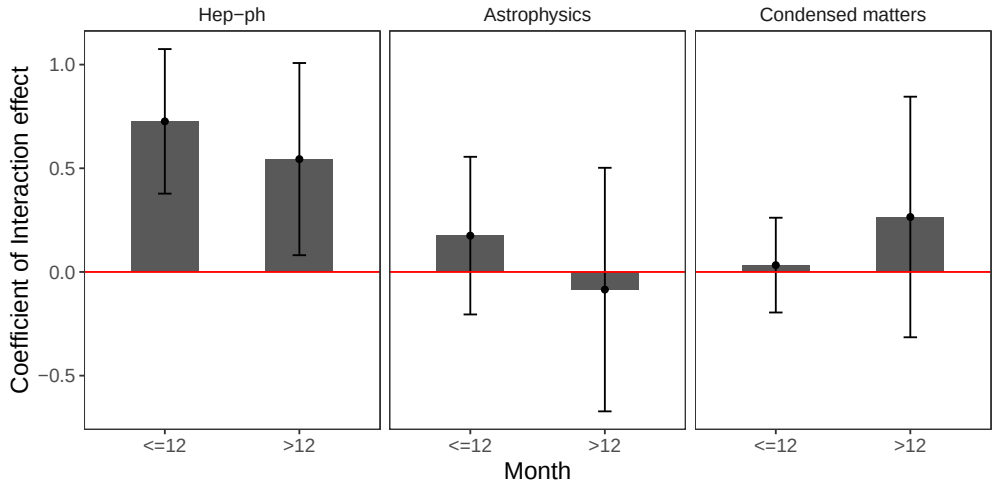
Figure 4.7: Coefficients of the interaction effect of journal influence and the total citation count made in arXiv in a given year from the count part of hurdle regression analysis by three disciplines. Y-axis shows the coefficient of the interaction effect from Model 3. X-axis shows the type of data set. The line at the top of the bar is the 95% confidence interval with robust standard error.



In Figure 4.8, the coefficients turn out to be positive and statistically significant for Hep-ph, however, they are not so for astrophysics and condensed matters. Overall, I cannot find

consistent decreasing or increasing change in the effect of journal influence.

Figure 4.8: Coefficients of the interaction effect of journal influence and the total citation count made in arXiv in a given year from the zero part of hurdle regression analysis by three disciplines. Y-axis shows the coefficient of the interaction effect from Model 3. X-axis shows the type of data set. The line at the top of the bar is the 95% confidence interval with robust standard error.



4.5 Discussion and conclusion

In this research, I have examined the changing meaning of journal publication as the credential system by answering two research questions. The first question asked whether there have been any changes in the composition of the quality of papers pursuing journal publication, and the second one examined the impact of journal status on its included papers' citation count and whether it has changed over time. Given that it is impossible to separate the effect of the quality of an individual paper from journal status because published papers are always read and cited by readers with its journal name, I have approached this problem by using the citation data linking arXiv and Microsoft Academic Database. As readers of

preprint versions of papers in arXiv do not know which journals they will be published in in the future, the citation data accumulated to preprint versions before journal publication can be used as an indicator of pure paper quality independent from journal status.

From the survival analysis, I found that while the cumulative citation count increases the probability of journal publication, the effect has reduced as more people use arXiv in all three disciplines. Notably, the decrease in the marginal effect stands out in astrophysics where the use of arXiv has increased rapidly. This finding implies that as the use of arXiv becomes more popularized among researchers, papers whose contributions have been acknowledged by being uploaded in arXiv only less actively seek a chance to be published in journals.

The effect of journal status on individual paper's citation count turns out to be over-emphasized without controlling individual paper's quality. After controlling paper's quality, the impact of journal status tends to be significantly lower than before; this adjustment has a more substantial effect when the proxy measure of paper's quality is supposed to be more accurate. However, it is only applicable when journal status predicts which papers are particularly cited more. In estimating the probability of ever cited or not, adding paper's quality as a control variable does not necessarily reduce the impact of journal status. There is a mixed finding of whether the effect of journal status has been cut or not over time across disciplines.

All in all, it can be concluded that although researchers have been less bound to the process of journal publication when their works are cited and read enough in other channels like arXiv, the role of a journal as a status marker of its included paper's quality remains strong in rapidly changing academic search environments. This finding is consistent with the conclusion of Larivière et al. [86] stating that there is no evidence yet to declare the demise of peer-reviewed journal articles at this point. While I did not find evidence of temporal changes in journal status, I found that previously measured journal effect might have been overestimated.

There are several possible explanations regarding why the journal's role as a status marker is still resistant in the new search environments. First, journals might actively seek a way to capture the paper's quality by recruiting renowned papers. As the evaluation of journals mostly relies on the metric, JIF, the editorial board actively looks for ways to increase it as much as possible [99]. One possibly safe and successful strategy for editors would be to invite papers that received a significant amount of attention in the conference or arXiv, which are more likely to bring a higher number of citations to a journal after being published.

Second, academic institutions might still actively rely on the status of journal in evaluating the quality of one's research, particularly for researchers in their early career. During the process of hiring new faculties or tenure review, the committee often relies its screening on which the journals the applicants have publications instead of the genuine quality of individual research papers. This is happening everywhere in spite of serious concerns about this practice raised from academic communities [62, 136]. Young scholars pursuing a successful academic career might inherently comply with the criterion and cooperate in regenerating it when they engage with previous literature in the new search environments as well.

Third, another possible institutional factor that influences on the resilient impact of journal is the economic interests of big academic publishers. By emphasizing journal impact factor up front and using this index as an important selling point, publishers can maintain the subscription from universities. The attempt to retain the economic interests by promoting its academic recognition might help continue why scholars consider journal status as an important signal of individual paper's quality.

Fourth, the disciplines studied in this chapter might not be especially influenced by changing search technology. These disciplines aim to analyze particular subjects, and thus disciplines with least interdisciplinary nature [148]. Researchers trained in these disciplines might retain expertise in specialized sub-fields and learn the appropriate set of journal names they need to cite for their work, which reduces the necessity to rely on new search technology.

Thus, there is a possibility that the findings might be driven by a bias of the selected disciplines towards unapproachable field specialty.

Thus, my analysis has a limit in that it can be only be generalized for disciplines that are less interdisciplinary and more specialized. The consequence of changing search environments on scholars' behavior in evaluating journal status might be more noticeable in disciplines that can be accessed easily by scholars from various backgrounds. However, because of the characteristics of arXiv that targets scholars in those specific scientific fields, the comparable longitudinal data for more interdisciplinary fields such as social sciences could not be attained. Perhaps the confidence of the organizers who first launched arXiv originated from the assumption that the scholars using this tool would have enough expertise to discern the scientific quality of papers and cite properly even without peer review processes and signals of journal status hinting to the paper's quality.

Another limitation of this research is its exclusion of one crucial possible confounding variable—a measure of the author's current career stage. Authors in early on in career might feel more pressure for journal publications, particularly in a high-status journal, because the hiring or tenure review committee still evaluates one's performance and future prospects based on which journals they have published in.

Finally, papers in arXiv might have been revised, possibly dramatically, during the review process of a journal. If a paper has been developed in terms of its scientific quality by going through the review, my proxy variable of paper's quality might not uniformly measure the quality of the same paper. For example, the rate of critical citations might be higher in earlier version of paper, which does not necessarily indicate the 'good' quality of paper. Thus, the way of measuring paper's quality used in this chapter can only stand when I assume that the main finding or contribution of papers has not changed during the review.

For future research, it will be helpful to study the performance of research articles published in a journal, but not in arXiv, then compare them to papers both uploaded in arXiv

and published in journals. If the amount of attention measured by received citation count has been similar for the two groups, it will indicate that the role of arXiv in calling more attention than traditional publication outlets might not be significant enough to be noted. In this way, I will be able to evaluate whether arXiv has been a tool in encouraging scholarly communication or not.

Chapter 5

A RE-EXAMINATION THE RELATIONSHIP BETWEEN "THE DEATH OF DISTANCE" HYPOTHESIS, AND INFORMATION AND COMMUNICATION TECHNOLOGY

5.1 Introduction

Does the advancement of information and communication technology (ICT) decrease the importance of the geographical location of occupations? It has been more than 40 years since futurists in the 1960s and 1970s predicted that technological development would make it possible for us to choose where we want to work regardless of where we should [145, 15, 101]. Currently, their predictions have enough technological foundation to be realized that face-to-face interaction can be replaced. Not infrequently, people participate in conference calls via Skype, do business via emails, and communicate via real time messaging applications to collaborate. Castells defines this trend as the rise of the network society where the key social structures and procedures are processed through electronic information networks [30] instead of geographical places. In his definition of the network society, static and physical interactions are replaced by dynamic and virtual interactions.

Previous research supports the futurists' claim that technology fundamentally transforms the face-to-face communication of workers—a sign of being one step closer to "death of distance" [29]. Although the physical proximity of workers has been an important component of co-worker communication [6, 128, 75] and fostering innovation [158], ICT demonstrates the possibility of replacing offline interactions with online ones and creating knowledge different from the ones produced offline. The virtual communication tool, already developed enough to replace simple interaction at work [43], has become the foundation in creating virtual

teams [147]. Also, occupations with flexible schedules are increasingly being conducted at home [5, 74] or at multiple locations [16]. The internet community—based on experts who neither necessarily know each other nor stay at the same location and not only limited to the transfer of simple knowledge—has emerged as a new hub of knowledge creation [12]. All in all, the prospect of what ICT can do for occupations has expanded and been broadly explored.

The ample evidence indicating the declining importance of physical distance and the rise of ICT questions the meaning of cities as the basis of local, economic, and social activities. While cities exist as an agglomeration of people and resources in the local area, ICT development and its role in replacing offline interaction might relate to the decrease in the significance of city. Not only workers who work distant from their designated places, but consumers are now able to shop with a few clicks at home (U.S. Census Bureau 2014), possibly undermining the role of the city as a center of economic consumption. However, we have not found the evidence to support this hypothesis yet; cities still remain meaningful centers for living and maintain their own unique roles. Moretti [107] shows that the location of jobs is still an important factor concerning the wage level. According to his research, the wage level of non-creative jobs is positively influenced by the creative jobs surrounding it. Sassen's theory about global hubs [130] also illustrates the concentrating financial role of cities. Since the fortunes of cities are unevenly distributed across the U.S., where people live has more meaning than ever before [92].

This chapter is focused on filling a gap in the existing literature between emphasizing the declining significance of geography and the concentration of cities. Despite the ample evidence supporting the "death of distance" driven by the development of ICT, why has the importance of geography in economic activities and wealth disparity been highlighted more than ever before? Does ICT equally contribute in distributing economic opportunities across cities or vice versa? Would bringing a higher level of ICT eventually prevent the concentra-

tion of cities? In this chapter, I have tried to answer these research questions by empirically examining the required ICT skills of occupations and their geographical distribution in the U.S. between 2006 and 2016. As occupation is defined by a set of skills needed to accomplish given tasks, it is reasonable to assume that workers in the same occupation are required to understand and use similar ICT skills, which is a better unit of analysis than measuring the overall ICT level of an industry (e.g., [68]) where workers with different levels of ICT are mixed together. Also, the communication of workers is not tightly confined within an organization, but it can be expanded to the fluid conversation between organizations—an essential part in knowledge creation [124]. Using a firm as an analysis unit might ignore the between-firm interactions. Besides, using occupation as a unit of analysis instead of industry has another advantage in explaining the concentration of cities, because occupation systematically represents the socio-economic status of people [45, 61, 138]. Thus, since people in the same occupation arguably have a homogeneous level of technology as well as similar socio-economic status, it is the appropriate unit of analysis to link ICT skills, geographical location, and its implication on the regional concentration of wealth.

Based on this idea, I have examined whether the location of occupations has become less important or independent of its use of ICT in 2006 and 2016. By comparing the two time periods where ICT was remarkably advanced, I have investigated whether a higher level of ICT can affect the importance of geography.

In the analysis, I have defined the importance of geography using two concepts: geographical dispersion of occupations and geographical interdependence of occupations. If occupations are clustered in limited geographical areas, it indicates that they require specific conditions provided by physical places. However, if they are dispersed across regions, geographical location matters less for occupations. The geographical interdependence of occupations, on the other hand, measures a different aspect of the importance of physical places. When the geographical distribution of an occupation is independent from the distribution

of other occupations, the occupation has little need to be in the same location as another, which can be a sign of lesser geographical significance. Based on these two measures, I have investigated the relationship between the ICT level of occupations and the importance of geography using a quantitative analysis, and proceed to analyze the findings.

5.2 Literature review

5.2.1 Importance of interaction at physical place

Physical proximity is considered to have a significant role in the workplace by facilitating communication between co-workers as they are more likely to interact in closer distances. Allen and Fustfeld [6] show that co-workers who are separated by a more than 25-meter walking distance have a lower probability of interaction. More recently, Sailer and McCulloh [128] developed this idea by focusing on the organization of physical space. They find that not only the Euclidean distance between co-workers but also the spatial distance measured by detailed configuration analysis explain more about the frequency of co-worker interaction, once again confirming the significance of physical distance as the basis of creating and maintaining a relationship.

The importance of physical place is particularly emphasized in interactions that nurture innovative ideas. Co-location of business entities is considered one of the most critical factors in collaborative knowledge creation; the success stories of industrial districts or new industrial spaces support this argument [109]. Similarly, Wineman et al. [158] and other follow-up studies [157, 75] also find the mechanism of how workplaces become the basis of creating innovative knowledge by linking physical place, social network, and innovation procedure. The workplace layout separates and unites space, guiding the formation of social ties among workers, facilitating the foundation of the innovation process. The learning process is not only limited to social ties within companies, but also emerges from the collaboration between various institutions that share similar research interests, ranging from universities to private

firms [124]. Whether more proximity and frequent interaction of workers always increase productivity is still an open question [22], but the evidence so far suggests that physical place does contribute to the creation of collaborative knowledge.

Further, some scholars argue that face-to-face interaction at work has its unique contribution that is not replaceable by ICT. For example, sharing activities such as eating or drinking together helps workers to sustain healthy relationships and boost work productivity [112]. Frequent face-to-face meetings turn out to moderate the relationship between team-empowerment and team-performance [78]. Also, offline interaction has its advantage in transferring complex and tacit knowledge rather than codified and standardized transactions [108]. Geographically dispersed virtual teams fail to maintain mutual knowledge because they cannot overcome communicating nuanced information that tells the significance of work [36]. Besides, the feeling of trust in business is something that cannot be transferable to an online relationship [89].

All in all, the physical proximity of workers has a positive impact on facilitating communication. Particularly, offline interactions at work have a unique advantage—in boosting teamwork and enabling nuanced communication of workers—that cannot be replaced by ICT. However, recent developments in ICT opens up a possibility that might threaten the existing role of offline interactions.

5.2.2 Technological change

ICT has rapidly developed since the late 20th century. The first technological development that enabled distant workplaces was telecommunication, which led to the creation of virtual teams. Townsend et al. [147] define virtual teams as "groups of geographically and/or organizationally dispersed coworkers that are assembled using a combination of telecommunications and information technologies to accomplish an organizational task." This new form of the team took the spotlight because it was considered as the practical and flexible survival

solution among lean and downsized organizations. Also, virtual teams could hire the right people without the limitations of geographical boundaries. Despite the free and malleable feature of virtual teams,however, there were limitations as well, mostly concerning the lack of ability in sharing complex ideas via telecommunication [53, 121, 36, 140].

However, the advancement in ICT for virtual communication now provides new services and benefits that facilitate complicated online work relationships [90]. Virtual communication tools now include video calls, and more recently, chatting applications where more than two people can share files, pictures, and videos. The most pronounced feature of the new technology is the quality of user-interactivity in contrast to the early one-way bulletins or emails [111]. Virtual communication gains popularity because it can save time and costs of travel [42] and even reduce greenhouse gas emission from transportation [3]. Though it cannot replace interactions required to complete highly complex tasks or develop new networks, it seems to have enough technological foundation to maintain existing ties and achieve workplace communication with less ambiguity [43].

The Internet also provides an interactive space for knowledge communities who do not necessarily exist in the same physical place or know each other. The success of the distanced group of collaborators in producing new knowledge based on virtual online space undermines the previous argument emphasizing the unique characteristics of physical proximity in collaboration. Amin and Roberts [8] find that presently, knowledge creation includes many organizations in different locations, which reduces the importance of geography in learning. Similarly, Bathelt and Turi [12] argue that computer-mediated communication is not something that exists to mimic face-to-face interaction but possesses its own distinctive role primarily embedded in the economy [11].

Following technological changes, more evidence supports the declining significance of geographical proximity in work interaction. For example, geographical proximity is not required in the entire process of knowledge transfer, but only in certain phases [146], which

suggests a further expansion of the meaning of proximity to organizational, institutional, and cognitive proximity [100]. This expansion of meaning implies that the boundary of people's collaboration can be formed from, geographically, anywhere. The increased inter-regional collaboration of academic scholars also demonstrates the decreasing importance of geographical location [104, 141, 65]. On an individual level, more workers have jobs that are characterized by highly flexible schedules as well as locations [5, 74].

5.2.3 Geographic consequences

The advancement in ICT begins to bring in a reshaping of geographical formation. The use of communication technology is likely to reduce the overall travel distance [90] and the necessity to live in the city center, thereby relating to a reformatting of the urban structure [77, 106]. Also, as ICT facilitates temporal and spatial fragmentation of work activities, it is expected that it will cause cities to be divided from one center to a multitude of medium-sized cities [38].

In contrast to this division of cities, evidence has shown the concentration of resources and people in a few U.S. cities. Until the early 1980s, the regional inequality had been gradually decreasing, but the trend was unexpectedly reversed after 1980. Only a few elite cities located on the coasts, such as Washington DC, San Francisco, and New York, have been rapidly developed [92]; cities such as New York, London, and Tokyo are not only globalization centers but also financial and business hubs [130]. The inequality between cities is not limited to income, but also leads to polarization in education level, political participation [107], and intergenerational mobility [32]. Also, where people live is an essential component that affects wage level [107] as well as the culture [52].

5.2.4 Research questions and conceptualization of the significance of location at work

The main goal of this study is to provide new evidence that disentangles the discussion about the relationship between ICT and the significance of location at work, and its possible impact on geographical inequality. Specifically, I have aimed to examine three related research questions.

First, are occupations with a higher level of ICT associated with a low significance of location? If ICT contributes to reduce the necessity of geographical proximity in communication at work, occupations with a higher level of ICT will be less bound with locations. In contrast, if there is an element of face-to-face communication at work that cannot be replaced by online relationship, the level of ICT will not be associated with geographical significance .

Second, has the progress of ICT positively or negatively influenced the significance of location at work over time? Some might argue that although ICT has not been developed enough to substitute offline communication, it is a matter of time before it eventually reduces the significance of local interaction. If there is a possibility that ICT can lead to the "death of distance", the association between the ICT level of occupations and their geographical importance will be strengthened in 2016 rather than 2006.

In this chapter, I have focused on two aspects of the significance of location. First, the most intuitive way of studying the significance of location at work is to see how many occupations are geographically dispersed or concentrated. If the percentage of workers in one occupation out of the total workforce in each city is stable across metropolitan areas, it indicates that occupations can exist in any regional conditions. If the location of work relates to specific conditions or resources such as natural environments, infrastructure, or labor pool that fulfills the education requirement of work that can be only provided by a limited number of metropolitan areas, the occupation will concentrate on a few metropolitan areas.

The second way of measuring the significance of location is to look at how many occupations are geographically related to others. Even for an occupation with low geographical dispersion, if it always needs other occupations in close distance, it is geographically bounded to the existence of other occupations. In this case, an occupation can be dispersed across metropolitan areas not because a location is less critical, but the geographical condition necessary for the occupation might have expanded to various metropolitan areas.

There are several possibilities regarding how occupations are geographically dependent on each other. First, occupations might be functionally in need of other occupations to accomplish the shared task. For example, doctors and nurses or judges and law clerks work together in proximity as they need the work of other occupations to finish their tasks. If ICT is able to replace functional communication, occupations will be less geographically dependent on each other. Secondly, occupations might happen to be geographically dependent when they have a similar cultural taste that emerges when similar people gather in the same area. The key literature to support this idea is Florida's creative class [52], which argues that the creative class, a newly emerged class consisting of the highly intelligent and innovative from the population, who share similar lifestyles, pursue diversity and individuality, flock together in those cities equipped with technology, other talented people, and tolerance. In this case, creative class occupations are likely to be geographically related for cultural reasons. The development of ICT might strengthen cultural interdependence as it is more likely to substitute dry work communication than cultural interaction. Finally, occupations might be interdependent due to specific natural environments. For example, sailors and meteorologists tend to be in the same city for neither functional nor cultural reasons but natural conditions for work. Since it is not able to transmit the information or experience collected from natural environments, ICT will not influence this type of interdependence.

Regardless of the type of geographical interdependence, since proximity is required between people from various occupations, systematic interdependence is the evidence that ge-

ography still retains its influence. While analyzing the dispersion of occupations tells us the importance of geography regarding a city's original conditions apart from human interaction, analyzing the interdependence of occupations focuses on the necessity of physical distance in communicating with others. Thus, tracing the changes in the relationship between ICT and geographical interdependence will reveal if ICT can replace offline interaction and lower geographical significance.

5.3 Data and method

5.3.1 Data

I have used the Occupational Employment Statistics (OES) of 2006 and 2016 created by the Bureau of Labor Statistics to know the distribution of occupations by geographic areas. I have used occupations This data includes information about employment and wage estimates for around 800 occupations published every year based on the combination of six semi-annual surveys on 1.2 million establishments in non-farm industries of the United States. The OES survey covers both full-time and part-time wage and salary workers but excludes the self-employed and unpaid family workers. The data provides the number of occupations according to various geographic boundaries such as nation, state, and metropolitan or non-metropolitan areas. The geographic boundary of metropolitan area was used because it is the boundary where people can meet for a reasonable amount of time while staying in the same economic community. The OES survey uses Standard Occupational Classification (SOC) system. As the 2006 OES survey uses the 2000 SOC system and the 2016 OES uses the 2010 SOC,the SOC code has been standardized to 2000 based on the crosswalks provided by the Bureau of Labor Statistics (BLS). For instances when there are more than two categories from the 2000 SOC matched to one category from the 2010 SOC, the category with the closest job title was chosen. I have only included the top 70% occupations in terms of its size for both time periods because some categories include only a small number of workers who are distributed

about 400 metropolitan statistical areas.

Additionally, I have collected the required data to measure the ICT level of occupations from the Occupational Information Network (O*NET) program supported by the U.S. Department of Labor/Employment and Training Administration. The O*NET database includes various occupation-specific information such as required tasks, technology skills, knowledge, work values as well as wages and employment trends since 1998. All the occupation-related descriptions in the database were collected by surveys of occupation holders and job analysts since 2002. From the O*NET database archive, I have used the data version 10.0 published in 2006 and the version 21.0 published and updated as of 2016. Since the 10.0 version, O*NET database has been providing occupation information based on SOC systems. Similar to the OES occupation classification, two versions of O*NET data use different SOC systems. By applying the same crosswalk used for the OES data, the SOC code of version 21.0 was standardized with the one in version 10.0.

5.3.2 Response variables

Geographical dispersion of occupations

To measure the amount of geographical dispersion, I have used the pre-normalized clustering index developed by Benson [17], which generalizes Duncan's dissimilarity index by comparing two groups to more than two. The generalized index, since I have compared the amount of concentration in more than 400 metropolitan areas, is suitable for my research. The clustering index of occupation i is computed by using the following formula:

$$C_i^* = \frac{1}{2} \sum_{i=1}^{I} \left| \frac{n_{im}}{n_i} - \frac{n_m - n_{im}}{n - n_i} \right|$$

where m indicates a metropolitan area and n indicates the number of workers. According to Benson [17], the intuitive explanation of the clustering index is "the share of workers within an occupation that must relocate for the share of workers to be balanced in every

metropolitan area". In other words, jobs with higher clustering index are more clustered, and vice versa. One of the most clustered occupations in the 2006 data is service unit operators (oil, gas, and mining) and it needs special environments that provide natural resources.

Geographical interdependence of occupations on other occupations

To measure the geographical dependence of other occupations, a method created and used to explain the development path of countries [63] and cities [110, 134] was used. Muneepeerakul et al. and Shutters et al. compute the interdependence among occupations and use it as a way of understanding transformation in urban economies. The same method of computing interdependence was used but interpreted differently here as the geographical interdependence of an occupation. The measure of Muneepeerakul et al. and Shutters et al. argue that the interdependence of two occupations increases when the prominence of occupation in one area continually coincides with the prominence of another occupation. Muneepeerakul et al. define that an occupation is predominant in a city when the percentage of workers in this occupation is higher than the national average percentage. They call it LQ_i^m, for occupation i in Metropolitan Statistical Area (MSA) m, and it is defined as follows:

$$LQ_i^m = \frac{(n_i^m / \sum_i n_i^m)}{(\sum_m n_i^m / (\sum_m \sum_i n_i^m))}$$

In this formula, n_i^m represents the number of workers with occupation i located in MSA m. Thus, when $LQ_i^m > 1$, it means that an occupation i is over-represented in an MSA m.

Then, the interdependence of two occupations i and j is defined as:

$$\zeta_{ij} = \frac{P[LQ_i^m > 1, LQ_j^m > 1]}{P[LQ_i^{m'} > 1]P[LQ_j^{m''} > 1]} - 1$$

where $m^{'}$ and $m^{''}$ represent two randomly selected MSAs. When ζ_{ij} is higher than 0, it means that i and j are more likely to be over-represented in the same city than them being a random probability. This computation method is not influenced by the size of

city or occupation and also includes an expression of negative coincidence. To simplify the interpretation, the interdependence of occupations is dichotomized as follows: when ζ_{ij} is larger than 1, the occupation dyad is defined as interdependent.

With these links, the degree centrality of the node as well as the number of links it possesses has been computed. In my data, the degree centrality indicates the number of occupations geographically interdependent to a given node. When the degree centrality is high, it means an occupation has many geographically interdependent occupations. However, when it is low, it means that an occupation is distributed randomly and unrelated to the location of other occupations.

5.3.3 Explanatory variables

As key explanatory variables, two items from the O*NET questionnaire have been employed to measure an occupation's ICT level. The first item measures the level of working with computers from low (e.g., entering employee information into a database) to high (e.g., the deployment of a new computer system). While it does not measure how often employers use e-mail or other virtual communication tools, it indirectly gives information about the level of computer skills that encompasses communication tools. The second item computes the frequency of using e-mail at work from never to every day. Both variables are measured as continuous because they provide the average of survey respondents in the same occupation. While the use of e-mail is a very basic form of ICT in comparison to other advanced techniques such as video conferencing or media sharing tools, it is the only item that directly relates to the job holder's ICT use. These two items are used together in the analysis to complement the limit of their measurement and check the robustness of results.

In addition to key variables, the size, creativity of tasks, and the broad category of occupations were controlled. Creativity of tasks was obtained from the O*NET data, and a continuous variable was computed in the same way as other O*NET variables were. From

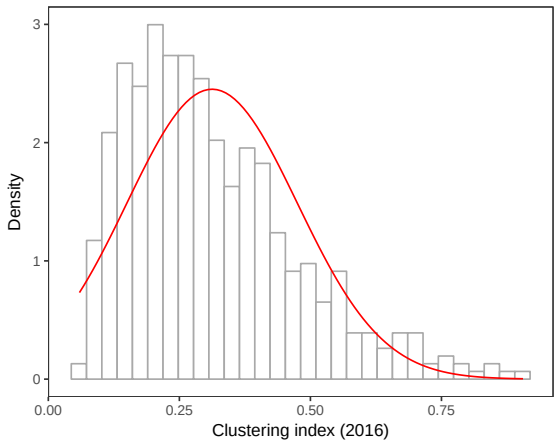
the questionnaire, creativity of tasks was measured from low (e.g., changing the space of the printed report) to high (e.g., creating new computer software). Creativity was controlled because it represents the level of intellectual activity that might still influence the reason that online communications cannot replace offline interactions(e.g., [108]). Finally, I have controlled the broad categories of occupations because the amount of ICT use or its impacts might differ with the type of work. The broad categories were classified based on the SOC system, namely: management and professional, service, sales and office, farming or fishing, construction trades, and production occupations.

5.3.4 Method

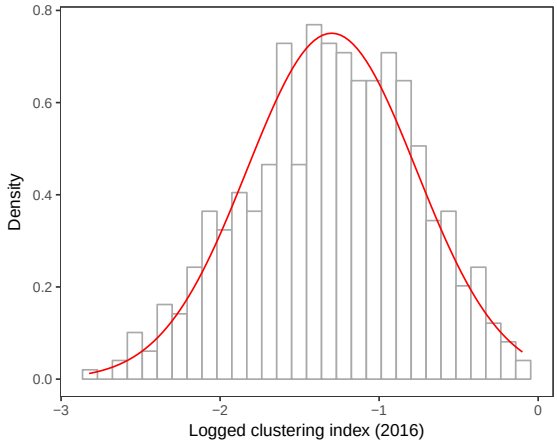
The analysis consists of three parts. In the first part, the relationship between ICT and the geographical dispersion of work measured by the clustering index was analyzed using a linear regression analysis with two models per period. The first model includes, in addition to all control variables, the level of working with computers as a measure of ICT. In the second model, the frequency of e-mail use replaces the level of working with computers. Since the dependent variable—the clustering index of occupations—showed a positively skewed distribution (Figure 5.1a) for both the time periods, a natural log to the clustering index was employed. Figure 5.1b shows that the distribution is close to the normal distribution after being logged.

Figure 5.1: The distribution of clustering index in 2016

(a) Original data



(b) Logged data

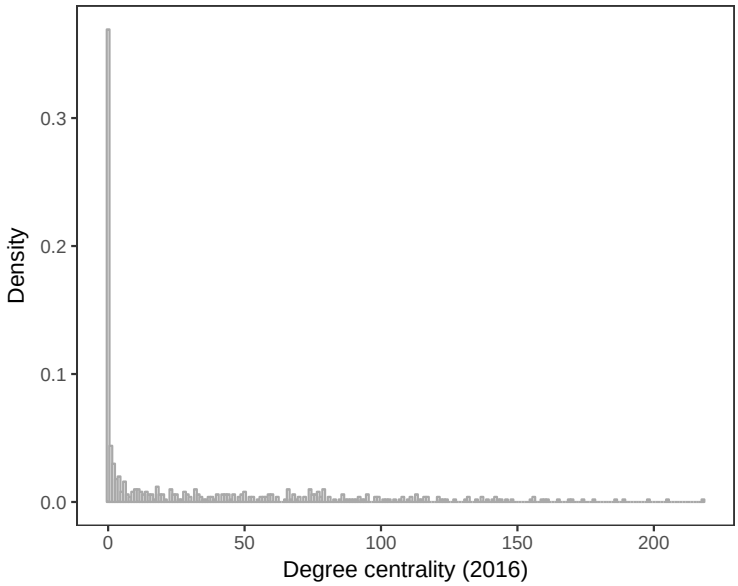


The second part of the analysis examines how ICT is associated with the geographical interdependence of work. Similar to the analysis of clustering index, I have tested two models with different measures of ICT. However, since the distribution of degree centrality was not appropriate to apply the linear regression analysis, I have used a different approach. As Figure 5.2 shows, the distribution of degree centrality in 2016 is highly skewed to the right with a lot of occupations with zero degree centrality.[1] In addition, the degree centrality only includes 0 and positive integers. In order to adjust excessive zeros and positive integers in the distribution, the zero-inflated regression model with Poisson distribution has been used. This method consists of two components that allow generating zeros, structural zeros, and ones created as part of Poisson distribution. The coefficients of variables are separately

[1]Zero degree centrality indicates that the occupations have no geographically interdependent occupations. In other words, these occupations are randomly distributed geographically and not interdependent with any other occupations.

estimated for each component. All control variables in the first part of analysis are included in addition to the clustering index.

Figure 5.2: The distribution of degree centrality in 2016



As the third step, I have compared the level of income between occupation dyads that are and are not geographically interdependent to understand the characteristics of interdependent occupations and their possible effects on wealth distribution between cities.

5.3.5 Descriptive statistics

Table 5.1 summarizes the list of variables and their descriptive statistics. For positively skewed distributions, its median instead of mean is presented as a summary statistic. While I have aimed to analyze the same set of occupations in both time periods, 2006 ended up having fewer number of observations due to gaps in the O*NET data, particularly for construction and production occupations. The median of the clustering index and degree centrality has

not changed much between 2006 and 2016. In contrast, the level of working with computers as well as the frequency of e-mail use have increased from 2006 to 2016, indicating that more occupations are engaged with higher levels of ICT. The level 3 of working with computers indicates the level of knowledge required to use a word processor. In 2006, the average use of e-mail at work is found to be slightly above "once a month or more but not every week", but it became closer to "once a week or more but not every day". Among control variables, the median occupation size, following the overall increase in labor forces, is higher in 2016 . Creativity of tasks has slightly increased as well. About a half of the total number of occupations are accounted by the management and professional categories while the other four share the other half.

Table 5.1: Descriptive statistics of used variables

Variable			Year	
			2006	2016
Response variable	Clustering index (median) (possible range 0-1)		.29	.27
	Degree centrality (median) (possible range 0-max occupation count)		6	6
Explanatory variables	Level of working with computers (mean) (possible range 0-7)		2.78	3.00
	Frequency of Email use (mean) (possible range 0-100)		54.03	64.24
Control variables	Occupation size (median)		72,100	82,210
	Creativity of tasks (possible range 0-7)		3.53	3.68
	Broad category	Management and professional	52%	47%
		Service	11%	11%
		Sales and office	13%	13%
		Construction	10%	12%
		Production	14%	18%
Total occupations (N)			441	501

5.4 Results

5.4.1 Geographical dispersion of work

In the first analysis, the relationship between ICT and geographical dispersion of work was observed by analyzing factors that influence the clustering index of occupations in 2006 and

2016. Models 1 and 2 in Table 5.2 summarize the coefficients from the linear regression models explaining the clustering index in 2006. Both measures of ICT—the level of working with computers and the frequency of e-mail use—are found not to be associated with the clustering index after other variables are controlled. In the 2016 results, the level of working with computers is not statistically significant, while the frequent use of email decreases the clustering index (= more dispersion of work). However, the size of the coefficient is relatively small. For an occupation with a median clustering index (0.27), a 10 unit increase in the frequency of e-mail use decreases its clustering index to 0.268, showing no substantive difference from 0.27.

In all the four models, control variables show similar effects on clustering index. Occupations with large labor forces are more likely to be dispersed, and creativity of tasks does not have a statistically significant effect except in Model 4. In Model 4, creative jobs are more likely to be clustered. In broad categories, production occupations are the most geographically clustered ones while sales and office occupations are the most dispersed.

The main finding from the analysis in Table 5.2 is that there is no clear evidence to support the claim that ICT use is related to the geographical dispersion of occupations. While some of the previous literature imply that a higher level ICT use will free people from where they should work, my findings suggest that these two factors are not necessarily linked to each other. In addition, despite technological advancements between 2006 and 2016, the relationship between ICT usage and the dispersion of occupations has neither strengthened nor weakened, partially forecasting the some possibility that the development of ICT will result in occupations unbound by locations.

Table 5.2: Coefficients from Linear Regression Models of Clustering Index: 2006 and 2016

		Year			
		2006		2016	
		Model 1	Model 2	Model 3	Model 4
Level of working with computers		0.002		-0.003	
		(.006)		(0.006)	
Frequency of Email use			0.000121		-.00060**
			(0.000214)		(0.00022)
Occupation size (logged)		-.089***	-.089***	-.089***	-.089***
		(0.004)	(0.004)	(.004)	(.004)
Creativity of tasks		0.008	0.007	0.010	.015*
		(.006)	(.006)	(.006)	(.006)
Broad category of occupations (ref=management and professional)	Service	.062***	.064***	.058**	.041*
		(.019)	(.019)	(.018)	(.018)
	Sales and office	0.014	0.014	0.018	0.021
		(.017)	(.017)	(.017)	(.016)
	Construction	.053**	.055**	.042*	0.02
		(.019)	(.019)	(.017)	(.016)
	Production	.074***	.077***	.097***	.073***
		(.017)	(.018)	(.016)	(.018)
Intercept		1.264***	1.262***	1.261***	1.280***
		(.054)	(.055)	(.052)	(.052)
Adjusted R-square		0.574	0.574	0.59	0.596

* <.05, ** <.01, *** <.001

5.4.2 Geographical interdependence of work

In the next stage of the analysis, the association between ICT and the geographical interdependence of work was examined by measuring the degree centrality. 5.3 displays the results from the 2006 data. Each model consists of two parts: the count part represents the Poisson distribution and the zero-inflation part shows the process of generating structural zeros. The count part of Model 1 suggests that the occupations with high level of working with computers are more likely to be geographically interdependent. Another measure of ICT, the frequency of e-mail use, also turned out to be positively related to degree centrality and statistically significant in the count part of Model 2. A similar implication is found in the zero-inflation part of models. Higher levels of computer skill as well as frequent use of e-mail increase the possibility of having a degree centrality that is non-zero.[2]

Regarding the control variables in the count part of both Model 1 and 2, the larger occupation size is more likely to decrease the degree centrality. Also, more creative occupations are likely to be geographically interdependent with other occupations, matching with the findings from previous literature which indicate that face-to-face interaction is not yet completely replaceable by online communication, particularly for complicated and nuanced materials [36, 108, 89]. In the zero-inflation part of both models, control variables do not influence the degree centrality except the clustering index.

[2]According to the statistical tool that I have used in this analysis—"pscl" package in R—the coefficients of the zero-inflation part estimate the probability of having structural zeros as the event in contrast to having non-zero observations. Thus, when a coefficient is positive, it means that the increase in a variable is positively associated with having a zero degree centrality.

Table 5.3: Coefficients from Zero-Inflated Poisson Regression Models of Degree Centrality: 2006

		2006			
		Model 1		Model 2	
		Count	Zero-inflation	Count	Zero-inflation
Level of working with computers		.193***	-.508**		
		(.009)	(0.161)		
Frequency of Email use				.0062***	-.0168**
				(0.0004)	(.0057)
Occupation size (logged)		-.281***	0.255	-.243***	0.276
		(0.011)	(0.174)	(0.011)	(0.171)
Creativity of tasks		.067***	-.135	.081***	-.162
		(0.009)	(0.152)	(0.010)	(0.152)
Clustering index		1.887***	-14.465***	1.938***	-14.083***
		(0.060)	(2.180)	(0.061)	(2.135)
Broad category of occupations (ref= management and professional)	Service	0.045	0.082	.063*	0.157
		(.031)	(.551)	(.032)	(.540)
	Sales and office	-.156***	-.527	-.134***	-.637
		(.036)	(.452)	(.036)	(.450)
	Construction	-.489***	1.010	-.400***	0.969
		(.037)	(.534)	(.040)	(.537)
	Production	-.344***	0.022	-.269***	-.116
		(.030)	(.482)	(.032)	(.499)
Intercept		5.555***	2.048	5.255***	1.309
		(.134)	(2.432)	(.136)	(2.353)
AIC		11226.76		11420.70	

* <.05, ** <.01, *** <.001

The summary of coefficients from Model 3 and 4 is illustrated in Table 5.4. The result of the 2016 data is similar to that of 2006. Both measures of ICT turned out to increase the degree centrality in the count and zero-inflation parts of the model. While it was not possible to directly compare the size of coefficients between 2006 and 2016, I could not find any signs of the declining effect of ICT on geographical interdependence during this period. Instead, creativity of tasks turned out to be statistically significant in the zero-inflation part of models. According to the results, highly creative occupations are not likely to have zero degree centrality, indicating that they are more likely to be geographically interdependent with at least one other occupation.

Table 5.4: Coefficients from Zero-Inflated Poisson Regression Models of Degree Centrality: 2016

		2016			
		Model 3		Model 4	
		Count	Zero-inflation	Count	Zero-inflation
Level of working with computers		.083***	-.474**		
		(.010)	(.180)		
Frequency of Email use				.0058***	-.0171*
				(.0004)	(.0068)
Occupation size (logged)		-.103***	0.193	-.094***	0.138
		(.011)	(.169)	(.011)	(.167)
Creativity of tasks		.074***	-.476**	.062***	-.451*
		(.010)	(.176)	(.010)	(.180)
Clustering index		2.975***	-17.801***	3.044***	-18.408***
		(.063)	(2.311)	(.063)	(2.331)
Broad category of occupations (ref= management and professional)	Service	-.132***	-0.442	-0.027	-0.526
		(.035)	(.537)	(.036)	(.555)
	Sales and office	-.090**	-.882*	-.080*	-0.863
		(.035)	(.439)	(.035)	(.441)
	Construction	-.575***	0.057	-.386***	-0.108
		(.034)	(.477)	(.038)	(.510)
	Production	-.580***	0.178	-.368***	-0.212
		(.029)	(.491)	(.034)	(.548)
Intercept		3.488***	5.103*	3.197***	5.595*
		(.132)	(2.492)	(.134)	(2.489)
AIC		11399.4		11280.47	

* <.05, ** <.01, *** <.001

According to the analysis results in Table 5.3 and Table 5.4, there is no evidence to support that ICT decreases the geographical interdependence of occupations in both time periods. In contrast, two factors are positively associated, implying the opposite of the "death of distance" hypothesis. When occupations use higher level of computer skills or interact via e-mail more frequently, they are more likely to be geographically consistent with other occupations.

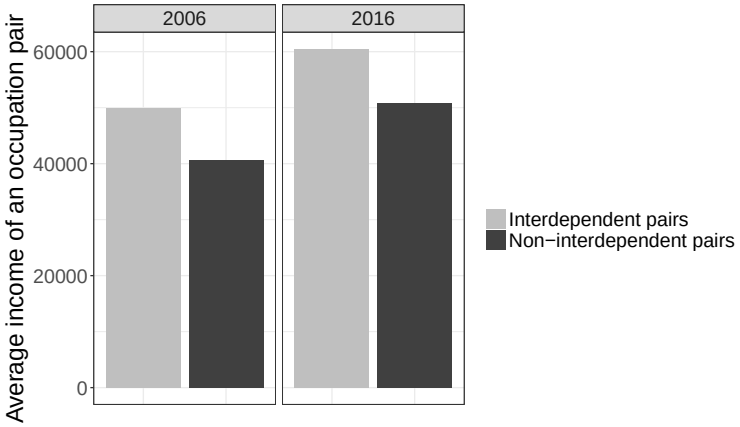
5.4.3 The income difference between geographically interdependent and non-interdependent occupation pairs

So far, I have investigated the relationship between the ICT level of an occupation and the significance of location at work based on two measures—geographical dispersion and interdependence—and found a positive association between ICT level and geographical interdependence. To study whether the level of ICT contributes to the increase in wealth concentration in few cities, I have looked at the average income difference between interdependent and non-interdependent occupation pairs in Figure 5.3. The average income is used as a way of measuring the socio-economic status of occupation pairs. While it is shown that occupations with higher ICT levels are more likely to co-locate in the same region, if geographically interdependent occupation pairs have higher income levels than non-interdependent pairs, it suggests that ICT helps gather highly paid jobs to the same regions instead of distributing them by obsoleting offline interaction.

According to Figure 5.3, the average annual income of a linked occupation pair is about 10,000 dollars higher than non-interdependent pairs in both time periods. The difference between two groups in two time periods are statistically significant according to t-test with .05 alpha level. A higher average income of interdependent occupation pairs indicates that high earners are more likely to coincide in the same metropolitan areas. Additionally, the similar level of income difference between the two groups in 2006 and 2016 shows that

interdependent occupation pairs still consisted of high earning occupations during the ten years. This result provides the evidence that ICT has contributed to an increase in the concentration of wealth only in a few cities.

Figure 5.3: The average income of geographically interdependent and non-interdependent groups in 2006 and 2016



5.5 Discussion and conclusion

In this chapter, I have studied how the development of ICT is related to "the death of distance" and whether this relationship has changed between 2006 and 2016. Further, I have briefly analyzed the impact of ICT on the concentration of wealth. By using the data from the Occupational Employment Statistics and O*NET database, I have examined the effect of ICT on the importance of geography for occupations by considering geographical dispersion and interdependence to other occupations. From the results, I could not find the evidence that ICT either increases or decreases geographical dispersion except the small amount of positive effect on dispersion by the frequency of e-mail use in 2016. However, ICT is positively associated with the geographical interdependence of work, and this effect has not been reduced in 2016.

The lack of association between ICT and the clustering index of occupations indicates that ICT use does not necessarily disperse or concentrate occupations. This finding contrasts with the finding of the previous literature that the use of communication technology decreases the overall travel distance and the preference to live in the city center [90, 77, 106]. In contrast, my finding implies that these effects might have been limited to the distance that can be still reached in a reasonable amount of time when necessary. Additionally, the minor change in this tendency between 2006 and 2016 shows that higher ICT has not affected the dispersion of occupations across metropolitan areas during the ten years covered in the analysis.

In contrast to the relationship between ICT and geographical dispersion, the use of ICT is positively associated with geographical interdependence in both 2006 and 2016, and the findings are robust with either measure of ICT in occupations. Regardless of the cause of interdependence, my findings show that ICT does not diminish the geographical consistency of occupations. Besides, the similar size of ICT coefficients in two time periods suggests that the advancement in ICT does not necessarily decrease its geographical reliance.

Combining the two findings, since the trend was found to barely change in the last ten years. I could not find any evidence for the claim that ICT has brought the "death of distance" and is likely to bring it in near future. Instead, occupations with higher ICT skills are more likely to have a relationship with other occupations' location. It might be due to the characteristics of the industry where most of these occupations with high ICT level belong. As the example of Silicon Valley illustrates, information and knowledge-intensive industries consist of occupations with high ICT level, and these industries were developed based on informal technical communities within minimial physical distances [131]. This finding is also consistent with previous literature emphasizing that physical proximity facilitates communication between workers for industrial innovation [158, 157, 75, 124]. The collective intelligent energy by being situated in the same region remains as something that is not yet possible to be replaced by online communication.

It was found that geographically interdependent occupation pairs have a higher average income level than non-interdependent pairs, implying that high earning jobs are more likely to stay together in the same city, which partially explains why the wealth has been concentrated in a few cities. While there is only indirect evidence, it was found that ICT has not contributed to decrease wealth concentration; the use of ICT does not reduce geographical interdependence but instead increases it, and these interdependent occupations have in general higher incomes. Therefore, this finding suggests that ICT is not helpful in reducing wealth inequality between cities.

This research has a limitation as it focused only on a relatively limited window of time: between 2006 and 2016. While comparing the two time periods can give a hint whether the development of ICT achieved for ten years can bring the "death of distance", it is still a narrow window considering the dramatic changes in ICT in the last several decades. I could not analyze the data that goes further back due to its unavailability at O*NET, but if a comparable data source is found, it would be helpful to analyze the effect of ICT with a long-term perspective.

Chapter 6

CONCLUSION

The dissertation examined the impact of technology on work practices by focusing on two cases: citation behavior of researchers and offline interactions at work. Through four empirical chapters, I have illustrated that the social consequences of adopting new technology at work are complicated and nuanced because technology is neither an isolated object nor an absolute solution, but rather a tool planted in the existing social structure.

The four chapters aimed to trace the consequences of technological changes with specific cases and viewpoints. Various databases and quantitative methodologies were used to demonstrate the empirical findings. Chapter 2 documented the temporal trend of the citation distribution's inequality measures—such as the Gini coefficient and the percentage of every cited papers for each year between 1996 and 2014—after adjusting all the conditions of the publication environment to be similar to that of 1996. In the results, I could not find the evidence to support that the distribution has become more dispersed over time, which contrasts with the findings of the existing literature. In Chapter 3, I studied whether technology differently influences scholars depending on their level of expertise. According to the results, scholars without expertise in the field are more likely to be influenced by the algorithm of new search engines. In Chapter 4, I investigated a journal's role as the credential system of its included articles. By following an innovative approach to measure the quality of a paper, I succeeded in separating the effect of a journal's status, which is arguably independent of an article's quality. The results showed that while the quality of the article has become less critical in pursuing journal publication, the effect of a journal on the citations received by an article has not decreased for published ones. However, the impact of a journal

on the citations of articles has been over-estimated without controlling individual paper's quality as the quality was found to confound the journal's effects. In Chapter 5, I examined whether the advancement in information and communication technology (ICT) influenced "the death of distance" by analyzing how essential geography is for occupations. According to my findings, there is no evidence to support that the technological development relates to the decreasing importance of geography. Instead, occupations with higher technical skills are more interdependent on other occupations' location, which contributes to the increasing economic inequality between cities.

Combining the results from the empirical chapters, I summarize the major implications of the dissertation. First of all, the overall findings of the dissertation in evaluating the effects of academic search as well as information and communication technology indicate that there is little evidence to support the powerful and immediate impacts of these technologies on society. It is true that our daily working practices have been dramatically changed; we heavily rely on the internet to search information instead of using a library and to communicate with co-workers as well as friends rather than meeting them offline. While these technologies seem to transform a large part of our micro behaviors, my research shows that the actual effects of recent technological development has not been as significant as it is being currently discussed. When the effect is systematically examined with structured data and proper methodology that controls possible confounding factors, the impact of technology is found to be gradual rather than disruptive. The findings of Chapter 2 represent this point well. After separating elements that might interrupt measuring changes of scholar's citation behavior, the temporal trend of inequality stays stable instead of decreasing dynamically.

Secondly, my research shows that the existing social structure is resilient in the course of technological changes. Instead of creating an evident disjuncture by transforming all social components from the beginning, the recent technological changes are found to be somewhat continuous. For example, in Chapter 4, I found that the effect of a journal on

its included individual paper's received citation count has not decreased, even though the new tools, including preprint services and academic search engines, had helped reduce the reliance on journal status. Similarly, in Chapter 5, the development of new communication tools could not stop good occupations being concentrated in a limited number of cities. Regardless of what the newly introduced technologies are supposed to do, the real changes, examined based on a macro-perspective with the systematic methodology, illustrate that many social processes continue to exert a powerful influence and hinders technology from having the only power to transform current work practices. Even when new technology changes people's behavior, the degree of its influence is amplified or negligible depending on existing social factors. In Chapter 3, my findings demonstrate that the behavior of scholars who are susceptible to social influence—outsiders of the field—is more likely to be influenced by the new search technology, while the scholars with expertise—who used to rely less on the evaluation of others—has barely any change in their behaviour.

However, my observations do not necessarily mean that technology does not have enough power to bring out transformative social changes. Instead, what it shows is that before jumping to conclusions and stating the consequences of new technology—mostly judged by the proliferation of their use—we need to approach this problem scientifically with macro perspectives.

In the rest of the concluding chapter, I summarize the limitations of the overall dissertation by stating two major points: methodologies to show the relationship between technology and its consequences, and the possible data bias. Then, I expand the meaning of my empirical findings and draw broader implications from them. In addition, I recommend a few policy suggestions for academic training and interpreting the influence of technology. Finally, I illustrate the results of a brainstorming session inspired by this dissertation that concerns future research—the most exciting part for me to think and write.

6.1 Limitations

This dissertation has a limit in that it was not able to provide any direct evidence for the link between technology and its consequences. Because researchers do not specify how they locate previous literature in article's references, the bibliographic data does not indicate if they use search engines for their research. Similarly, workers (or companies) do not explain the relocation of jobs with the amount of communication technology they use. Instead, the use of technology is more smoothly embedded in our daily lives, and thus hard to separate from existing conditions. Due to these features of technology and the limitations in the data, my findings depend on temporal trends or analytic strategies designed to infer the implied changes driven by technology, from which it might be weak to argue that technology is the primary source of observed transformation.

Also, I need to note the possible bias in databases used in this research. Regarding the Web of Science database used in Chapter 2 and 3, while it is true that it is one of the most extensive bibliographic databases continuously covering articles from various disciplines since the emergence of modern science, it is also true that Web of Science has a limited coverage in non-English journals that are still participating in the academic discussion. These journals are not only omitted from the analysis but also not included in computing citation counts and Journal Influence Factor, which might contribute to the imperfect representation of what really happens in search engines. The database from search engines, such as Microsoft Academic Graph used in Chapter 4, might reduce this bias because it broadly scrapes web pages regardless of language. However, since the online database never stops updating as web pages change their contents, the perfect time to collect the best representational data does not exist. Thus, even if I realize that the database has become consistently better, I need to stop updating the database and use current uncompleted data. Besides, the database of search engines might suffer from the quality control issues as the research on Google Scholar illustrates [58].

6.2 Broader implications and recommendations

In this section, I add a few comments about the implications suggested by the dissertation for specific contexts as well as broader implications in studying the impact of technology in general. Mainly, I focus on recommendations in practices of searching literature and evaluating research results in new technological environments.

One of the most used and most important tools for scientists is the academic search engine, but little attention has been paid to the potential biases and downstream effects of these tools, including their impact on discovery and the trajectories of scientific careers. Search engines improve access to the existing literature, but there is also a risk that built-in, algorithmic biases, as well as human behavioral biases might impact the kind of scientific information that is actually found, read, cited, and communicated to the public. Since it is essential for scientists to be familiar with the established knowledge to make scientific progress, the existence of invisible echo chambers could hinder the productivity of scientific research. To put it bluntly, if these new technologies are concentrating scientists' attention on only a subset of high-quality research, subsequent discovery and innovation could be dampened.

This research could also help the general public determine the significance of scientific studies by explaining why the most cited scientific articles do not always contain the most important or the best scientific research. As the analysis of Chapter 3 shows, in the era of new search engines, a paper's cumulative number of prior citations is an increasingly significant predictor of whether subsequent authors will cite a piece of research, particularly for interdisciplinary scholars. This finding suggests that as search algorithms promote highly cited papers to top of their search results, they are producing a Matthew effect, in which popular papers become more popular. Possibly, the impact of strategic self-citations might contribute to putting their own papers in the cycle as well, which might subsequently influence academic promotion and funding decisions.

The finding of this dissertation again warns academic committees to rely less on the status of a journal in evaluating a scholar's scientific achievement along with previous manifestos [62, 85]. Although a journal's status remains as an influential factor to predict the following received citations, I show that this impact is overestimated than it should be in Chapter 4. Also, the finding suggests the possibility that more research articles with high quality might choose preprint services instead of the platform of a journal. These findings imply the diversification of the presentation of research articles, demanding focus on the individual research paper's quality than relying on its journal's status for evaluation.

The results of this dissertation should also influence the way in which scientists are trained to engage with academic information technology, especially, when they seek out literature in unfamiliar fields. Both librarians and new scholars must learn how to be savvy in their use of these technologies, which are now deeply embedded in the scientific research process. While it is widely recognized that search engine technology has been rapidly developing, how they transform scholars' practice and information dissemination strategies has not been adequately identified.

This research also calls attention to the significance of enhancing online infrastructure for all scholars. The growing importance of technology in doing research means that the continued gaps between researchers with and without access to electronized articles are of critical importance. The research results could convince school administration and librarians that securing access to digital archives is just as important as storing copies of journals and books in the library.

Finally, as we have seen in the case of search engines as well as ICT, it is necessary to be aware that technology does not always bring consequences that it is expected to bring. Thus, in discussing the role of technology, it is essential to critically observe what really happens in a society as new technology is introduced instead of assuming it based on what the technology initially intends to achieve.

6.3 Future work

Based on the dissertation, I illustrate a few research ideas that will either improve the limitations or develop the presented research topics. First of all, as a complementary analysis to what I mentioned as the limitation of this study, it would be helpful to use or collect data that can more directly show the impact of search engines on researcher's citation behavior. For this purpose, I suggest a qualitative approach such as individual or focus group interviews of scholars that survey their usual protocol of how to browse unfamiliar fields and sort out which papers to put their time and energy into. Also, it might be helpful to conduct an experiment to test whether changes in search algorithms influence users' decision in choosing articles to download and read. The last source of data that can strengthen the argument of the dissertation is JSTOR's usage data. So far, the dissertation has focused on the relationship between search engines and the citation behavior of researchers. The usage data will be able to fill in the gap of this relationship by linking the use of search engines and the way articles are accessed before making a citation.

The dissertation implies the possibility that there might be differences in citing behaviors among those who were trained before, during, and after the process of digitization and the emergence of search engines. Most researchers learn how to engage with literature throughout their training in graduate school, and are not intensively trained again for the rest of their academic career. However, the development of technology in academic research has changed more rapidly than the length of an average academic career, and thereby may require senior researchers to acquire new skills. Also, established and novice scholars may have different levels of expertise as well as different social capital for collecting bibliographic information, which might influence the frequency or intensity of their reliance of search engines. Thus, it will be essential to examine the generational differences since if senior and junior scholars use technology differently, an active academic discussion between scholars might be a challenge to accomplish.

In analyzing the bibliographic data, the dissertation mainly focused on the citation structure to investigate the process of knowledge creation along with the development of search engines. In addition to the structural approach, tracing the actual contents of scholarly articles will be an exciting complement to the current dissertation. For example, I am interested in studying whether the knowledge domain has converged on a few popular topics as the scholarly interaction has been made easier and faster than before by analyzing the abstracts of articles. Another research idea is to examine the gap between the general public and the academic audience's interest in research articles by using data from various forms of non-academic online writings and scholarly citations. This research will reveal whether an increased accessibility of the general public to academic work—as people will quickly locate articles from various online sources—reduces the gap between the understanding of academic articles of a scholar and a non-scholar.

Finally, I suggest using the concept of occupation interdependence for further investigating the two-body problem. The two-body problem is a commonly used term in academia that indicates a couple's difficulty in having two decent jobs close enough. The geographical interdependence of occupations is a measure of how much an occupation pair is likely to exist in the same location. By extending the application of the term to highly educated professionals, the concept of occupation interdependence can help explain whether changes in the location can exacerbate or alleviate the two-body problem.

BIBLIOGRAPHY

[1] Giovanni Abramo, Ciriaco Andrea D'Angelo, and Flavia Di Costa. Citations versus journal impact factor as proxy of quality: could the latter ever be preferable? *Scientometrics*, 84(3):821–833, 2010.

[2] Anurag Acharya, Alex Verstak, Helder Suzuki, Sean Henderson, Mikhail Iakhiaev, Cliff Chiung Yu Lin, and Namit Shetty. Rise of the rest: The growing impact of non-elite journals. *arXiv preprint arXiv:1410.2217*, 2014.

[3] Anne Aguilera. Business travel and mobile workers. *Transportation Research Part A: Policy and Practice*, 42(8):1109–1116, 2008.

[4] Arthur S Alderson. Explaining deindustrialization: globalization, failure, or success? *American Sociological Review*, pages 701–721, 1999.

[5] Tooran Alizadeh. Teleworkers' characteristics in live/work communities: Lessons from the united states and australia. *Journal of Urban Technology*, 19(3):63–84, 2012.

[6] Thomas John Allen and Alan R Fustfeld. Research laboratory architecture and the structuring of communications. *R&D Management*, 5(2):153–164, 1975.

[7] Benjamin M. Althouse, Jevin D. West, Carl T. Bergstrom, and Theodore Bergstrom. Differences in impact factor across fields and over time. *Journal of the American Society for Information Science and Technology*, 60(1):27–34, 2009.

[8] Ash Amin and Joanne Roberts. Knowing in action: Beyond communities of practice. *Research policy*, 37(2):353–369, 2008.

[9] Éric Archambault and Vincent Larivière. History of the journal impact factor: Contingencies and consequences. *Scientometrics*, 79(3):635–649, 2009.

[10] Albert-László Barabási, Chaoming Song, and Dashun Wang. Handful of papers dominates citation. *Nature*, 491(7422):40–41, 2012.

[11] Harald Bathelt and Sebastian Henn. The geographies of knowledge transfers over distance: toward a typology. *Environment and Planning A*, 46(6):1403–1424, 2014.

[12] Harald Bathelt and Philip Turi. Local, global and virtual buzz: The importance of face-to-face contact in economic interaction and possibilities to go beyond. *Geoforum*, 42(5):520–529, 2011.

[13] Joeran Beel, Stefan Langer, Marcel Genzmehr, Bela Gipp, Corinna Breitinger, and Andreas Nürnberger. Research paper recommender system evaluation: a quantitative literature survey. In *Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation*, pages 15–22. ACM, 2013.

[14] Jöran Beel and Bela Gipp. Google scholar's ranking algorithm: an introductory overview. In *Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI'09)*, volume 1, pages 230–241. Rio de Janeiro (Brazil), 2009.

[15] Daniel Bell. The coming of the post-industrial society. In *The Educational Forum*, volume 40, pages 574–579. Taylor & Francis, 1976.

[16] Eran Ben-Elia, Bayarma Alexander, Christa Hubers, and Dick Ettema. Activity fragmentation, ict and travel: An exploratory path analysis of spatiotemporal interrelationships. *Transportation Research Part A: Policy and Practice*, 68:56–74, 2014.

[17] Alan Benson. Rethinking the two-body problem: The segregation of women into geographically dispersed occupations. *Demography*, 51(5):1619–1639, 2014.

[18] David B Bills. Credentials, signals, and screens: Explaining the relationship between schooling and job assignment. *Review of educational research*, 73(4):441–449, 2003.

[19] Lutz Bornmann and Loet Leydesdorff. Skewness of citation impact data and covariates of citation distributions: A large-scale empirical analysis based on web of science data. *Journal of Informetrics*, 11(1):164–175, 2017.

[20] Lutz Bornmann and Rüdiger Mutz. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, 66(11):2215–2222, 2015.

[21] Ángel Borrego and Lluís Anglada. Faculty information behaviour in the electronic environment: attitudes towards searching, publishing and libraries. *New Library World*, 117(3/4):173–185, 2016.

[22] Ron Boschma. Proximity and innovation: a critical assessment. *Regional studies*, 39(1):61–74, 2005.

[23] Hamid Bouabid. Revisiting citation aging: a model for citation distribution and life-cycle prediction. *Scientometrics*, 88(1):199, 2011.

[24] Meredith Broussard. *Artificial Unintelligence: How Computers Misunderstand the World*. MIT Press, 2018.

[25] Cecelia Brown. The e-volution of preprints in the scholarly communication of physicists and astronomers. *Journal of the Association for Information Science and Technology*, 52(3):187–200, 2001.

[26] Cecelia Brown. The role of electronic preprints in chemical communication: Analysis of citation, usage, and acceptance in the journal literature. *Journal of the Association for Information Science and Technology*, 54(5):362–371, 2003.

[27] Erik Brynjolfsson, Yu Hu, and Duncan Simester. Goodbye pareto principle, hello long tail: The effect of search costs on the concentration of product sales. *Management Science*, 57(8):1373–1386, 2011.

[28] Erik Brynjolfsson, Yu Jeffrey Hu, and Michael D Smith. From niches to riches: Anatomy of the long tail. *Sloan Management Review*, 47(4):67–71, 2006.

[29] Frances Cairncross. *The death of distance: How the communications revolution will change our lives*. {Harvard Business School Press}, 1997.

[30] Manuel Castells. *The rise of the network society*. Oxford: Blackwell, 1996.

[31] Ivan D Chase, Craig Tovey, Debra Spangler-Martin, and Michael Manfredonia. Individual differences versus social dynamics in the formation of animal dominance hierarchies. *Proceedings of the National Academy of Sciences*, 99(8):5744–5749, 2002.

[32] Raj Chetty, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez. Where is the land of opportunity? the geography of intergenerational mobility in the united states. *The Quarterly Journal of Economics*, 129(4):1553–1623, 2014.

[33] Harry M. Collins, Luis Reyes-Galindo, and Paul Ginsparg. A note concerning primary source knowledge. *Journal of the Association for Information Science and Technology*, 68(5):1105–1110, 2017.

[34] Randall Collins. *The credential society: An historical sociology of education and stratification*. Academic Pr, 1979.

[35] Tanya Cothran. Google scholar acceptance and use among graduate students: A quantitative study. *Library & Information Science Research*, 33(4):293–301, 2011.

[36] Catherine Durnell Cramton. The mutual knowledge problem and its consequences for dispersed collaboration. *Organization science*, 12(3):346–371, 2001.

[37] Kate Crawford. Artificial intelligence's white guy problem. *The New York Times*, 2016.

[38] Daniel Czamanski and Dani Broitman. Information and communication technology and the spatial evolution of mature cities. *Socio-Economic Planning Sciences*, 58:30–38, 2017.

[39] Philip Davis and Michael Fromerth. Does the arxiv lead to higher citations and reduced publisher downloads for mathematics articles? *Scientometrics*, 71(2):203–215, 2007.

[40] Prabuddha De, Yu Hu, and Mohammad S Rahman. Technology usage and online sales: An empirical study. *Management Science*, 56(11):1930–1945, 2010.

[41] D. J. de Solla Price. Networks of scientific papers. *Science*, 149:510–515, 1965.

[42] Jon Martin Denstadli. Impacts of videoconferencing on business travel: the norwegian experience. *Journal of Air Transport Management*, 10(6):371–376, 2004.

[43] Jon Martin Denstadli, Tom Erik Julsrud, and Randi Johanne Hjorthol. Videoconferencing as a mode of communication: A comparative study of the use of videoconferencing and face-to-face meetings. *Journal of Business and Technical Communication*, 26(1):65–91, 2012.

[44] Fereshteh Didegah and Mike Thelwall. Determinants of research citation impact in nanoscience and nanotechnology. *Journal of the Association for Information Science and Technology*, 64(5):1055–1064, 2013.

[45] Otis Dudley Duncan. A socioeconomic index for all occupations. *Class: Critical Concepts*, 1:388–426, 1961.

[46] Anita Elberse. Should you invest in the long tail? *Harvard business review*, 86(7/8):88, 2008.

[47] James A. Evans. Electronic publication and the narrowing of science and scholarship. *Science*, 321(5887):395–399, 2008.

[48] James A Evans. Response. *Science*, 323:37–38, 2009.

[49] James A Evans and Jacob Reimer. Open access and global participation in science. *Science*, 323(5917):1025–1025, 2009.

[50] Gunther Eysenbach. Citation advantage of open access articles. *PLoS biology*, 4(5):e157, 2006.

[51] Daniel Fleder and Kartik Hosanagar. Blockbuster culture's next rise or fall: The impact of recommender systems on sales diversity. *Management science*, 55(5):697–712, 2009.

[52] Richard Florida. The rise of the creative class. *New York: Basic books*, 2002.

[53] Jess Gaspar and Edward L Glaeser. Information technology and the future of cities. *Journal of urban economics*, 43(1):136–156, 1998.

[54] Anne Gentil-Beccot, Salvatore Mele, and Travis C Brooks. Citing and reading behaviours in high-energy physics. *Scientometrics*, 84(2):345–355, 2010.

[55] Jim Giles. Science in the web age: Start your engines. *Nature*, 438(7068):554–555, 2005.

[56] Yves Gingras, Vincent Larivière, and Éric Archambault. Literature citations in the internet era. *Science*, 323:36, 2009.

[57] Roger V Gould. The origins of status hierarchies: A formal theory and empirical test. *American journal of sociology*, 107(5):1143–1178, 2002.

[58] Gali Halevi, Henk Moed, and Judit Bar-Ilan. Suitability of google scholar as a source of scientific information and as a source of data for scientific evaluation—review of the literature. *Journal of Informetrics*, 11(3):823–834, 2017.

[59] David P Hamilton. Publishing by–and for?–the numbers. *Science*, 250(4986):1331–1332, 1990.

[60] Lowell L Hargens. Using the literature: Reference networks, reference contexts, and the social structure of scholarship. *American sociological review*, pages 846–865, 2000.

[61] Robert M Hauser and John Robert Warren. Socioeconomic indexes for occupations: A review, update, and critique. *Sociological methodology*, 27(1):177–298, 1997.

[62] Diana Hicks, Paul Wouters, Ludo Waltman, Sarah De Rijcke, and Ismael Rafols. The leiden manifesto for research metrics. *Nature*, 520(7548):429, 2015.

[63] César A Hidalgo, Bailey Klinger, A-L Barabási, and Ricardo Hausmann. The product space conditions the development of nations. *Science*, 317(5837):482–487, 2007.

[64] Joseph M Hodge, Gerald Hödl, Martina Kopf, Alberto Accomazzi, Carolyn S Grant, Donna Thompson, Elizabeth Bohlen, Stephen S Murray, Paul Ginsparg, and Simeon Warner. E-prints and journal articles in astronomy: a productive co-existence. *Learned publishing*, 20(1):16–22, 2007.

[65] Jarno Hoekman, Koen Frenken, and Robert JW Tijssen. Research collaboration at a distance: Changing spatial patterns of scientific collaboration within europe. *Research Policy*, 39(5):662–673, 2010.

[66] Tad Hogg and Kristina Lerman. Disentangling the effects of social signals. *arXiv preprint arXiv:1410.6744*, 2014.

[67] Tad Hogg and Kristina Lerman. Disentangling the effects of social signals. *Human computation*, 2(2):189–208, 2015.

[68] Junjie Hong and Shihe Fu. Information and communication technologies and the geographical concentration of manufacturing industries: Evidence from china. *Urban Studies*, 48(11):2339–2354, 2011.

[69] Beibei Hu, Xianlei Dong, Chenwei Zhang, Timothy D Bowman, Ying Ding, Staša Milojević, Chaoqun Ni, Erjia Yan, and Vincent Larivière. A lead-lag analysis of the topic evolution patterns for preprints and publications. *Journal of the Association for Information Science and Technology*, 66(12):2643–2656, 2015.

[70] Mu-Hsuan Huang, Han-Wen Chang, and Dar-Zen Chen. The trend of concentration in scientific research and technological innovation: A reduction of the predominant role of the us in world research & technology. *Journal of Informetrics*, 6(4):457–468, 2012.

[71] Sven E Hug, Michael Ochsner, and Hans-Dieter Daniel. A framework to explore and develop criteria for assessing research quality in the humanities. *International Journal for Education Law and Policy*, 10:55, 2014.

[72] David Ingold and Spencer Soper. Amazon doesn't consider the race of its customers. should it. *Bloomberg, April*, 2016.

[73] Jerry A Jacobs and Scott Frickel. Interdisciplinarity: A critical assessment. *Annual review of Sociology*, 35:43–65, 2009.

[74] Tammy Johns and Lynda Gratton. The third wave of virtual work. *Harvard Business Review*, 91(1):66–73, 2013.

[75] Felichism Kabo, Yongha Hwang, Margaret Levenstein, and Jason Owen-Smith. Shared paths to the lab: A sociospatial network analysis of collaboration. *Environment and Behavior*, 47(1):57–84, 2015.

[76] Qing Ke, Emilio Ferrara, Filippo Radicchi, and Alessandro Flammini. Defining and identifying sleeping beauties in science. *Proceedings of the National Academy of Sciences*, 112(24):7426–7431, 2015.

[77] Seung-Nam Kim, Patricia L Mokhtarian, and Kun-Hyuck Ahn. The seoul of alonso: New perspectives on telecommuting and residential location from south korea. *Urban Geography*, 33(8):1163–1191, 2012.

[78] Bradley L Kirkman, Benson Rosen, Paul E Tesluk, and Cristina B Gibson. The impact of team empowerment on virtual team performance: The moderating role of face-to-face interaction. *Academy of Management Journal*, 47(2):175–192, 2004.

[79] Christopher Kollmeyer and Florian Pichler. Is deindustrialization causing high unemployment in affluent countries? evidence from 16 oecd countries, 1970-2003. *Social forces*, 91(3):785–812, 2013.

[80] Balázs Kovács and Chengwei Liu. Audience structure and status multiplicity. *Social Networks*, 44:36–49, 2016.

[81] Adam DI Kramer, Jamie E Guillory, and Jeffrey T Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24):8788–8790, 2014.

[82] Vincent Larivière and Yves Gingras. The impact factor's matthew effect: A natural experiment in bibliometrics. *Journal of the Association for Information Science and Technology*, 61(2):424–427, 2010.

[83] Vincent Larivière and Yves Gingras. On the relationship between interdisciplinarity and scientific impact. *Journal of the Association for Information Science and Technology*, 61(1):126–131, 2010.

[84] Vincent Larivière, Yves Gingras, and Éric Archambault. The decline in the concentration of citations, 1900–2007. *Journal of the Association for Information Science and Technology*, 60(4):858–862, 2009.

[85] Vincent Lariviere, Veronique Kiermer, Catriona J MacCallum, Marcia McNutt, Mark Patterson, Bernd Pulverer, Sowmya Swaminathan, Stuart Taylor, and Stephen Curry. A simple proposal for the publication of journal citation distributions. *Biorxiv*, page 062109, 2016.

[86] Vincent Larivière, Cassidy R Sugimoto, Benoit Macaluso, Staša Milojević, Blaise Cronin, and Mike Thelwall. arxiv e-prints and the journal of record: An analysis of roles and relationships. *Journal of the Association for Information Science and Technology*, 65(6):1157–1169, 2014.

[87] Vincent Larivière, Éric Archambault, and Yves Gingras. Long-term variations in the aging of scientific literature: From exponential growth to steady-state science (1900–2004). *Journal of the American Society for Information Science and Technology*, 59(2):288–296, 2008.

[88] Vincent Larivière, George A. Lozano, and Yves Gingras. Are elite journals declining? *Journal of the Association for Information Science and Technology*, 65(4):649–655, 2014.

[89] Edward E Leamer and Michael Storper. The economic geography of the internet age. In *Location of International Business Activities*, pages 63–93. Springer, 2014.

[90] SH Lee, YT Leem, and JH Han. Impact of ubiquitous computing technologies on changing travel and land use patterns. *International Journal of Environmental Science and Technology*, 11(8):2337–2346, 2014.

[91] Kristina Lerman and Tad Hogg. Leveraging position bias to improve peer recommendation. *PloS one*, 9(6):e98914, 2014.

[92] Phillip Longman. Why the economic fates of america's cities diverged. *The Atlantic*, 2015.

[93] George A Lozano, Vincent Larivière, and Yves Gingras. The weakening relationship between the impact factor and papers' citations in the digital age. *Journal of the Association for Information Science and Technology*, 63(11):2140–2145, 2012.

[94] Freda B Lynn. Diffusing through disciplines: Insiders, outsiders, and socially influenced citation behavior. *Social Forces*, 93(1):355–382, 2014.

[95] Freda B Lynn, Joel M Podolny, and Lin Tao. A sociological (de) construction of the relationship between status and quality. *American Journal of Sociology*, 115(3):755–804, 2009.

[96] Adrian Mackenzie. The performativity of code: Software and cultures of circulation. *Theory, Culture & Society*, 22(1):71–92, 2005.

[97] Donald MacKenzie and Judy Wajcman. *The social shaping of technology: How the Refrigerator got its Hum*. Milton Keynes: Open university press, 1985.

[98] Noah P Mark, Lynn Smith-Lovin, and Cecilia L Ridgeway. Why do nominal characteristics acquire status value? a minimal explanation for status construction. *American Journal of Sociology*, 115(3):832–862, 2009.

[99] Ben R Martin. Editors' jif-boosting stratagems–which are appropriate and which not?, 2016.

[100] Jannika Mattes. Dimensions of proximity and knowledge bases: innovation between spatial and non-spatial factors. *Regional Studies*, 46(8):1085–1099, 2012.

[101] Marshall McLuhan. *The Gutenberg galaxy: The making of typographic man*. University of Toronto Press, 1962.

[102] Robert K Merton et al. The matthew effect in science. *Science*, 159(3810):56–63, 1968.

[103] Shinji Mine. The roles and place of arxiv in scholarly communication. *Library and Information Science*, 61:25–58, 2009.

[104] Guang Ying Mo, Zack Hayat, and Barry Wellman. How far can scholarly networks go? examining the relationships between distance, disciplines, motivations, and clusters. In *Communication and Information Technologies Annual*, pages 107–133. Emerald Group Publishing Limited, 2015.

[105] Henk F Moed. The effect of "open access" on citation impact: An analysis of arxiv's condensed matter section. *Journal of the Association for Information Science and Technology*, 58(13):2047–2054, 2007.

[106] Markus Moos and Andrejs Skaburskis. Workplace restructuring and urban form: The changing national settlement patterns of the canadian workforce. *Journal of Urban Affairs*, 32(1):25–53, 2010.

[107] Enrico Moretti. *The new geography of jobs*. Houghton Mifflin Harcourt, 2012.

[108] Kevin Morgan. The exaggerated death of geography: learning, proximity and territorial innovation systems. *Journal of economic geography*, 4(1):3–21, 2004.

[109] Frank Moulaert and Farid Sekia. Territorial innovation models: a critical survey. *Regional studies*, 37(3):289–302, 2003.

[110] Rachata Muneepeerakul, José Lobo, Shade T Shutters, Andrés Goméz-Liévano, and Murad R Qubbaj. Urban economies and occupation space: can they get "there" from "here"? *PloS one*, 8(9):e73676, 2013.

[111] Philip M Napoli. Revisiting 'mass communication'and the 'work'of the audience in the new media environment. *Media, Culture & Society*, 32(3):505–516, 2010.

[112] Bonnie A Nardi and Steve Whittaker. The place of face-to-face communication in distributed work. *Distributed work*, pages 83–110, 2002.

[113] David Nicholas, Anthony Watkinson, Hamid R Jamali, Eti Herman, Carol Tenopir, Rachel Volentine, Suzie Allard, and Kenneth Levine. Peer review: Still king in the digital age. *Learned Publishing*, 28(1):15–21, 2015.

[114] Xi Niu and Bradley M. Hemminger. A study of factors that affect the information-seeking behavior of academic scientists. *Journal of the American Society for Information Science and Technology*, 63(2):336–353, 2012.

[115] Safiya Umoja Noble. *Algorithms of Oppression: How search engines reinforce racism*. NYU Press, 2018.

[116] Alireza Noruzi. arxiv popularity from a citation analysis point of view. *Webology*, 13(2):1–7, 2016.

[117] Candela Ollé and Ángel Borrego. A qualitative study of the impact of electronic journals on scholarly information behavior. *Library & Information Science Research*, 32(3):221–228, 2010.

[118] Natsuo Onodera and Fuyuki Yoshikane. Factors affecting citation rates of research articles. *Journal of the Association for Information Science and Technology*, 66(4):739–764, 2015.

[119] Wanda J Orlikowski and Susan V Scott. What happens when evaluation goes online? exploring apparatuses of valuation in the travel sector. *Organization Science*, 25(3):868–891, 2013.

[120] Raj K Pan, Alexander M Petersen, Fabio Pammolli, and Santo Fortunato. The memory of science: Inflation, myopia, and the knowledge network. *arXiv preprint arXiv:1607.05606*, 2016.

[121] Alexandros Panayides and Clifford R Kern. Information technology and the future of cities: an alternative analysis. *Urban Studies*, 42(1):163–167, 2005.

[122] Pietro Della Briotta Parolo, Raj Kumar Pan, Rumi Ghosh, Bernardo A Huberman, Kimmo Kaski, and Santo Fortunato. Attention decay in science. *Journal of Informetrics*, 9(4):734–745, 2015.

[123] Joel M Podolny. A status-based model of market competition. *American journal of sociology*, 98(4):829–872, 1993.

[124] Walter W Powell, Kenneth W Koput, and Laurel Smith-Doerr. Interorganizational collaboration and the locus of innovation: Networks of learning in biotechnology. *Administrative science quarterly*, pages 116–145, 1996.

[125] Isuru Ranasinghe, Abbas Shojaee, Behnood Bikdeli, Aakriti Gupta, Ruijun Chen, Joseph S Ross, Frederick Masoudi, John A Spertus, Brahmajee K Nallamothu, and Harlan M Krumholz. Poorly cited articles in peer-reviewed cardiovascular journals from 1997-2007: Analysis of 5-year citation rates. *Circulation*, pages CIRCULATIONAHA–114, 2015.

[126] Cecilia L Ridgeway and James W Balkwell. Group processes and the diffusion of status beliefs. *Social Psychology Quarterly*, pages 14–31, 1997.

[127] Cecilia L Ridgeway, Elizabeth Heger Boyle, Kathy J Kuipers, and Dawn T Robinson. How do status beliefs develop? the role of resources and interactional experience. *American Sociological Review*, pages 331–350, 1998.

[128] Kerstin Sailer and Ian McCulloh. Social networks and spatial configuration—how office layouts drive social interaction. *Social networks*, 34(1):47–58, 2012.

[129] Matthew J Salganik, Peter Sheridan Dodds, and Duncan J Watts. Experimental study of inequality and unpredictability in an artificial cultural market. *science*, 311(5762):854–856, 2006.

[130] Saskia Sassen. *The global city: New york, london, tokyo*. Princeton University Press, 2001.

[131] AnnaLee Saxenian. *Regional advantage*. Harvard University Press, 1996.

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

[133] Alexander Serenko and John Dumay. Citation classics published in knowledge management journals. part ii: studying research trends and discovering the google scholar effect. *Journal of Knowledge Management*, 19(6):1335–1355, 2015.

[134] Shade T Shutters, Rachata Muneepeerakul, and José Lobo. Constrained pathways to a creative urban economy. *Urban Studies*, 53(16):3439–3454, 2016.

[135] Rieneke Slager, Jean-Pascal Gond, and Jeremy Moon. Standardization as institutional work: The regulatory power of a responsible investment standard. *Organization Studies*, 33(5-6):763–790, 2012.

[136] Paula Stephan, Reinhilde Veugelers, Jian Wang, et al. Blinkered by bibliometrics. *Nature*, 544(7651):411–412, 2017.

[137] David I Stern. High-ranked social science journal articles can be identified from early citation information. *PloS one*, 9(11):e112520, 2014.

[138] Gillian Stevens and David L Featherman. A revised socioeconomic index of occupational status. *Social Science Research*, 10(4):364–395, 1981.

[139] Joseph E Stiglitz. The theory of" screening," education, and the distribution of income. *The American economic review*, 65(3):283–300, 1975.

[140] Michael Storper and Anthony J Venables. Buzz: face-to-face contact and the urban economy. *Journal of economic geography*, 4(4):351–370, 2004.

[141] Yutao Sun and Cong Cao. Intra-and inter-regional research collaboration across organizational boundaries: Evolving patterns in china. *Technological Forecasting and Social Change*, 96:215–231, 2015.

[142] Iman Tahamtan, Askar Safipour Afshar, and Khadijeh Ahamdzadeh. Factors affecting number of citations: a comprehensive review of the literature. *Scientometrics*, 107(3):1195–1225, 2016.

[143] Carol Tenopir, Donald W King, Sheri Edwards, and Lei Wu. Electronic journals and changes in scholarly article seeking and reading patterns. *Aslib proceedings: New information perspectives*, 61(1):5–32, 2009.

[144] Mike Thelwall. Microsoft Academic automatic document searches: Accuracy for journal articles and suitability for citation analysis. *Journal of Informetrics*, 12(1):1–9, 2018.

[145] Alvin Toffler and Toffler Alvin. *The third wave*, volume 484. Bantam books New York, 1980.

[146] André Torre. On the role played by temporary geographical proximity in knowledge transmission. *Regional Studies*, 42(6):869–889, 2008.

[147] Anthony M Townsend, Samuel M DeMarie, and Anthony R Hendrickson. Virtual teams: Technology and the workplace of the future. *The Academy of Management Executive*, 12(3):17–29, 1998.

[148] Richard Van Noorden et al. Interdisciplinary research by the numbers. *Nature*, 525(7569):306–307, 2015.

[149] Alex Verstak, Anurag Acharya, Helder Suzuki, Sean Henderson, Mikhail Iakhiaev, Cliff Chiung Yu Lin, and Namit Shetty. On the shoulders of giants: The growing impact of older articles. *arXiv preprint arXiv:1411.0275*, 2014.

[150] Nicolas Vibert, Jean-François Rouet, Christine Ros, Mélanie Ramond, and Bruno Deshoullieres. The use of online electronic information resources in scientific research: The case of neuroscience. *Library & Information Science Research*, 29(4):508–532, 2007.

[151] Matthew L Wallace, Vincent Larivière, and Yves Gingras. Modeling a century of citation distributions. *Journal of Informetrics*, 3(4):296–303, 2009.

[152] James G Webster and Thomas B Ksiazek. The dynamics of audience fragmentation: Public attention in an age of digital media. *Journal of communication*, 62(1):39–56, 2012.

[153] Ian Wesley-Smith, Ralph J. Dandrea, and Jevin D. West. An experimental platform for scholarly article recommendation. In *Proc. of the 2nd Workshop on Bibliometric-enhanced Information Retrieval (BIR2015)*, pages 30–39, 2015.

[154] J.D. West, T.C. Bergstrom, and C.T. Bergstrom. The eigenfactor metrics: A network approach to assessing scholarly journals. *College and Research Libraries*, 71(3):236–244, 2010.

[155] Jevin D West, Michael C Jensen, Ralph J Dandrea, Gregory J Gordon, and Carl T Bergstrom. Author-level eigenfactor metrics: Evaluating the influence of authors, institutions, and countries within the social science research network community. *Journal of the American Society for Information Science and Technology*, 64(4):787–801, 2013.

[156] William Julius Wilson. *The truly disadvantaged: The inner city, the underclass, and public policy*. University of Chicago Press, 1987.

[157] Jean Wineman, Yongha Hwang, Felichism Kabo, Jason Owen-Smith, and Gerald F Davis. Spatial layout, social structure, and innovation in organizations. *Environment and Planning B: Planning and Design*, 41(6):1100–1112, 2014.

[158] Jean D Wineman, Felichism W Kabo, and Gerald F Davis. Spatial and social networks in organizational innovation. *Environment and Behavior*, 41(3):427–442, 2009.

[159] Soo Jeong Yoon, Dae Young Yoon, Hyung Jin Lee, Sora Baek, Kyoung Ja Lim, Young Lan Seo, and Eun Joo Yun. Distribution of citations received by scientific papers published in the imaging literature from 2001 to 2010: Decreasing inequality and polarization. *American Journal of Roentgenology*, 209(2):248–254, 2017.

[160] H. Peyton Young. Innovation diffusion in heterogeneous populations: Contagion, social influence, and social learning. *American Economic Review*, 99(5):1899–1924, December 2009.

[161] Alejandro Zentner, Michael Smith, and Cuneyd Kaya. How video rental patterns change as consumers move online. *Management Science*, 59(11):2622–2634, 2013.

[162] Shu-Dong Zhang. Judge a paper on its own merits, not its journal's. *Nature*, 442(7098):26, 2006.

Appendix A

APPENDIX - CHAPTER 2

A.1 Aggregation of journals to disciplines and disciplines to fields

Across WoS, journals are classified into one or more disciplines. We choose the first two discipline categories, and when either of two categories matches with one of our aimed broad categories, we include the journal in that field. The classification of disciplines to broad categories follows the National Science Foundation's taxonomy of disciplines created by the Integrated Postsecondary Education Data System (IPEDS) survey following Larivière et al., but we categorize them more in detail to each unit as an integrated entity. Out of 14 available categories, we use four broad categories.

Health

Allergy; Andrology; Anesthesiology; Audiology & Speech-Language Pathology; Cardiac & Cardiovascular Systems; Clinical Neurology; Critical Care Medicine; Dentistry, Oral Surgery & Medicine; Dermatology; Emergency Medicine; Endocrinology & Metabolism; Gastroenterology & Hepatology; Geriatrics & Gerontology; Health Care Sciences & Services; Health Policy & Services; Hematology; Infectious Diseases; Integrative & Complementary Medicine; Medical Ethics; Medicine, General & Internal; Medicine, Legal; Medicine, Research & Experimental; Neuroimaging; Nursing; Obstetrics & Gynecology; Oncology; Ophthalmology; Orthopedics; Pathology; Pediatrics; Peripheral Vascular Disease; Primary Health Care; Psychiatry; Public, Environmental & Occupational Health; Radiology, Nuclear Medicine & Medical Imagin; Radiology, Nuclear Medicine & Medical Imaging; Respiratory System; Rheuma-

tology; Transplantation; Tropical Medicine; Urology & Nephrology; Veterinary Sciences

Humanities

Art; Classics; Dance; Ethics; Film, Radio, Television; Folklore; History; Humanities, Multidisciplinary; Literary Reviews; Literary Theory & Criticism; Literature; Literature, African, Australian, Canadian; Literature, American; Literature, British Isles; Literature, German, Dutch, Scandinavian; Literature, Romance; Literature, Slavic; Logic; Medieval & Renaissance Studies; Music; Philosophy; Poetry; Religion; Theater

Mathematics and computer sciences

Computer Science, Artificial Intelligence; Computer Science, Cybernetics; Computer Science, Hardware & Architecture; Computer Science, Information Systems; Computer Science, Interdisciplinary Applications; Computer Science, Software Engineering; Computer Science, Theory & Methods; Information Science & Library Science; Mathematical & Computational Biology; Mathematics; Mathematics, Applied; Mathematics, Interdisciplinary Applications; Statistics & Probability

Social sciences

Agricultural Economics & Policy; Anthropology; Archaeology; Area Studies; Asian Studies; Behavioral Sciences; Criminology & Penology; Cultural Studies; Demography; Economics; Ethnic Studies; Family Studies; Geography; Geography, Physical; Gerontology; History & Philosophy Of Science; History Of Social Sciences; International Relations; Language & Linguistics; Linguistics; Political Science; Public Administration; Social Issues; Social Sciences, Biomedical; Social Sciences, Interdisciplinary; Social Sciences, Mathematical Methods; Social Work; Sociology; Urban Studies; Women's Studies

A.2 Comparison of adjusted and unadjusted data - additional measures and year window

Figure A.1: The temporal trend of HHI between 1996 and 2014 by four broad categories (two-year citation window); Bright dots – adjusted data, vague dots – unadjusted data; Solid line - statistically significant time trend & dotted line – statistically insignificant time trend (a statistical test is done using a robust regression model)

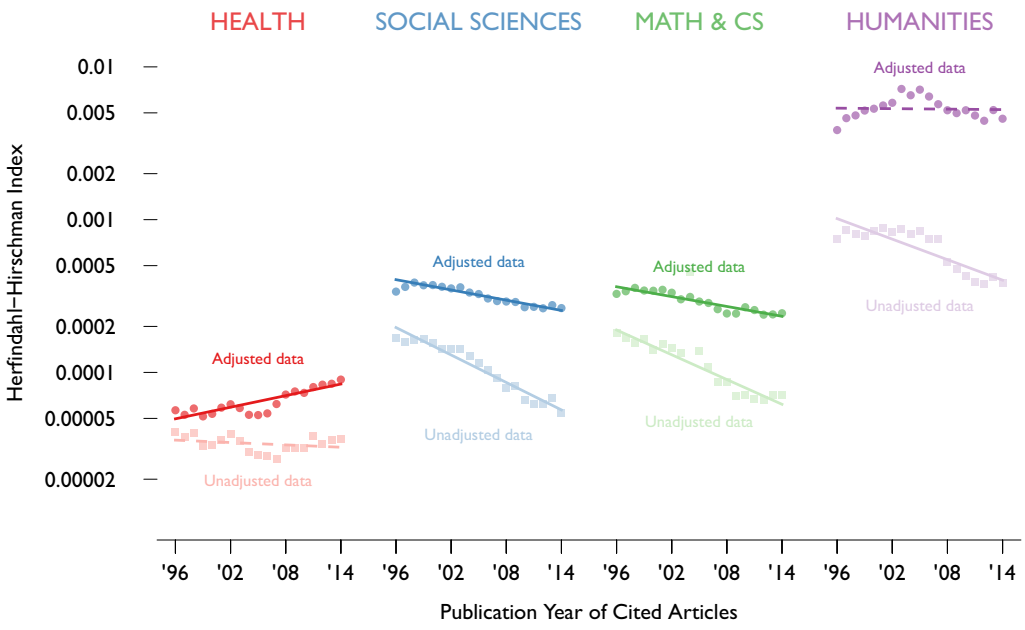


Figure A.2: The temporal trend of percentage of ever cited papers between 1996 and 2014 by four broad categories (six-year citation window); legends are the same as Figure A.1

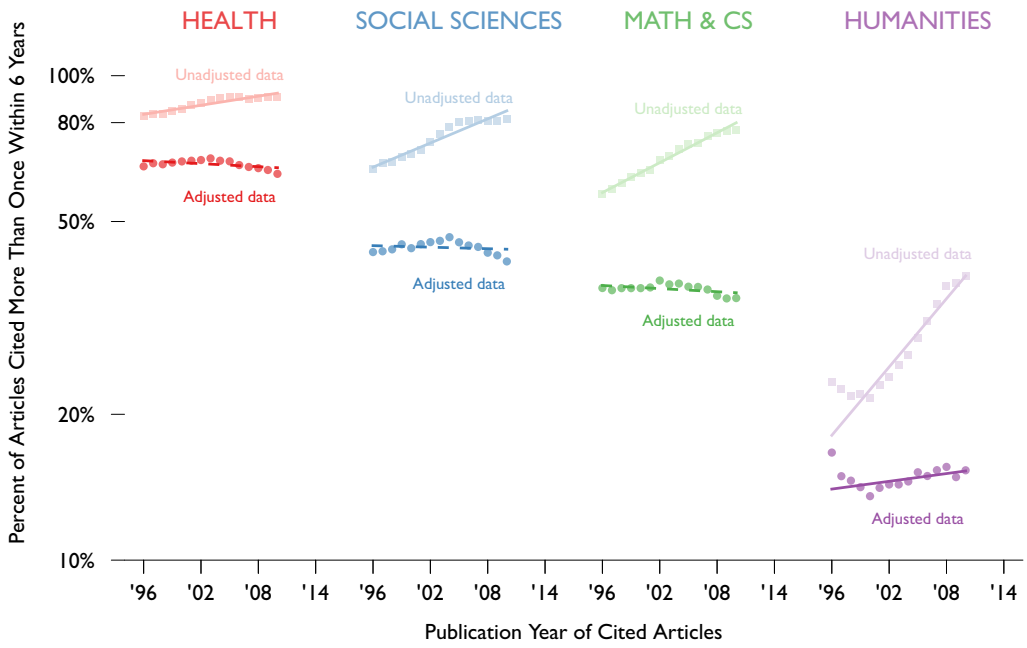


Figure A.3: The temporal trend of percentage of papers that needed to account for 20% and 80% of citations between 1996 and 2014 by four broad categories (six-year citation window); legends are the same as Figure A.1

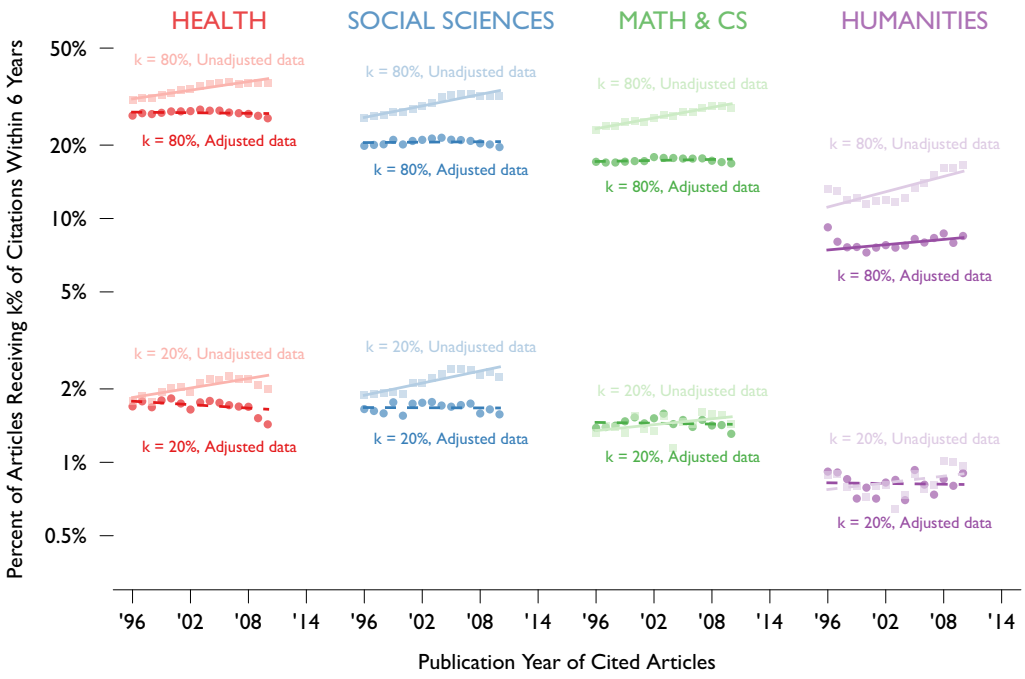


Figure A.4: The temporal trend of Gini coefficient between 1996 and 2014 by four broad categories (six-year citation window); legends are the same as Figure A.1

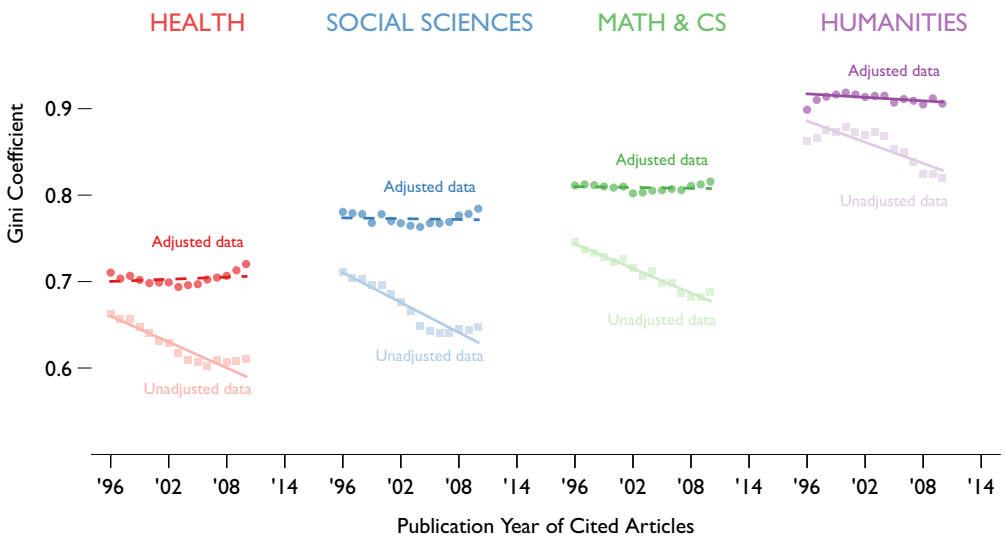
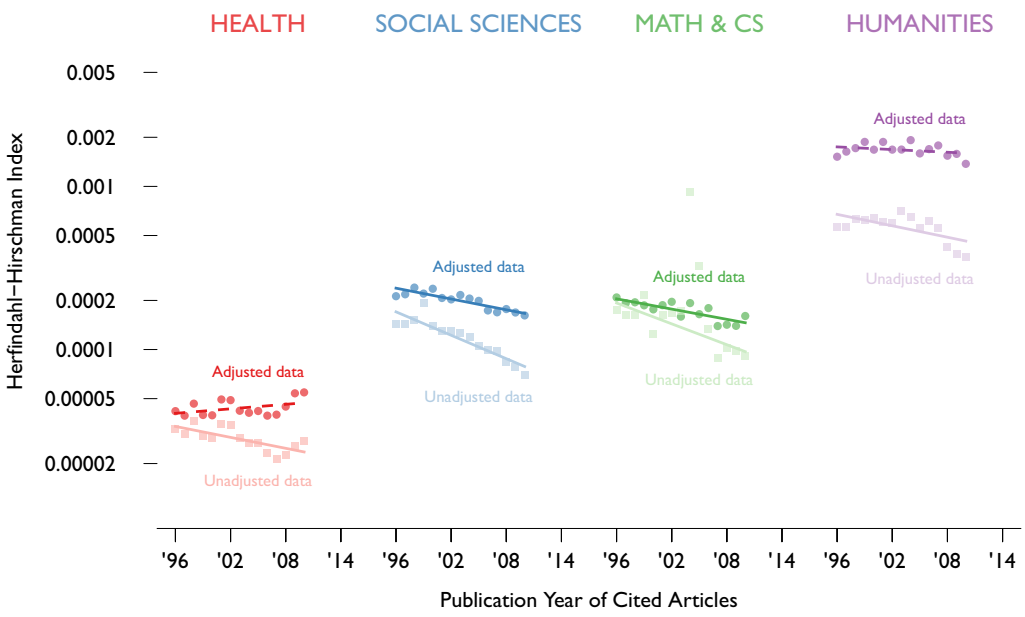


Figure A.5: The temporal trend of HHI between 1996 and 2014 by four broad categories (six-year citation window); legends are the same as Figure A.1



A.3 Comparison of temporal trends with two- and six-year citation windows - additional measures

Figure A.6: The temporal trend of percentage of ever cited papers by four broad categories (1996-2014 for two-year citation window, and 1996-2010 for six-year citation window); Circles – two-year citation window, squares – six-year citation window; Solid line - statistically significant time trend & dotted line – statistically insignificant time trend (a statistical test is done using a robust regression model)

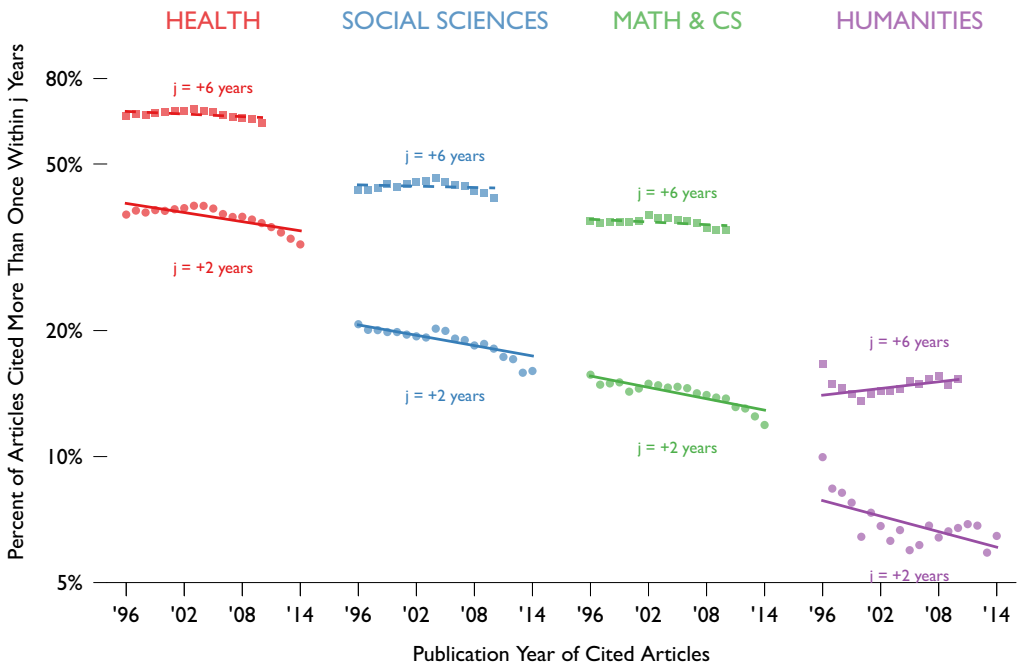
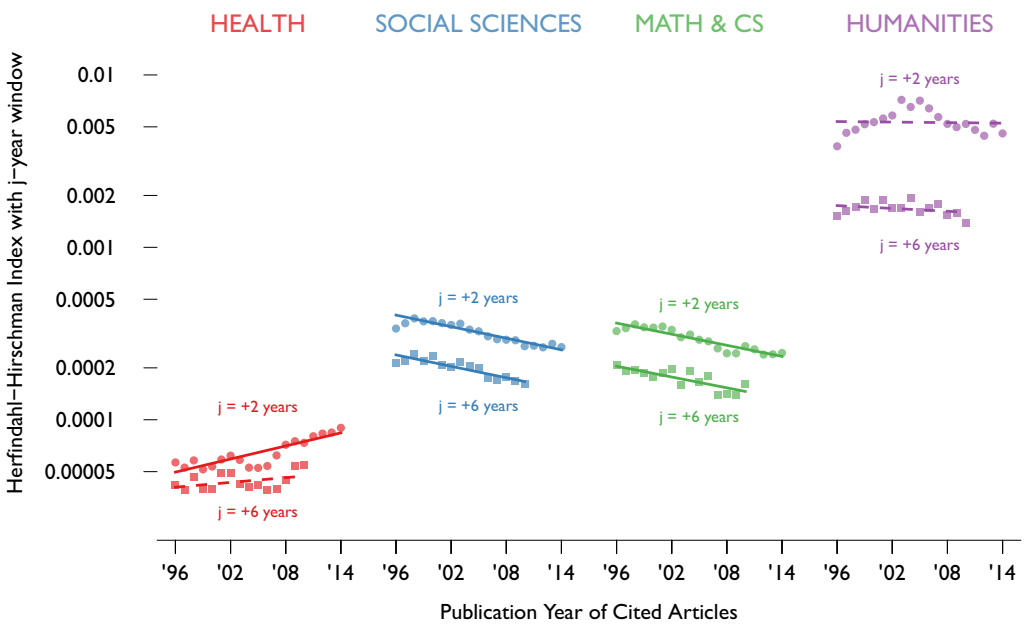


Figure A.7: The temporal trend of HHI by four broad categories (1996-2014 for two-year citation window, and 1996-2010 for six-year citation window); legends are the same as Figure A.6



A.4 Analysis by adjustment component - additional measures

Figure A.8: The temporal trend of Gini coefficient between 1996 and 2014 by four broad categories (two-year citation window); legends are the same as Figure 2.7

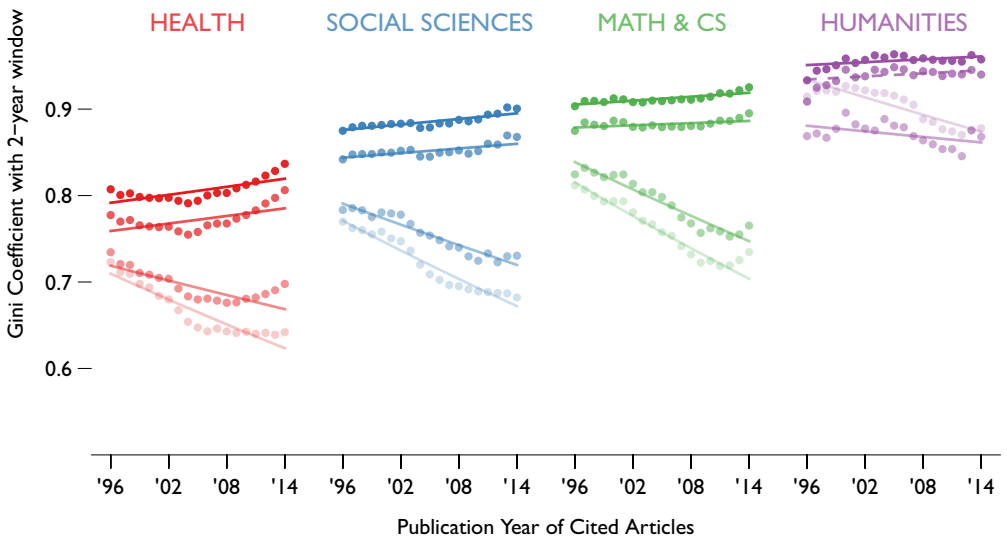


Figure A.9: The temporal trend of percent of ever cited papers by four broad categories between 1996 and 2014 by four broad categories (two-year citation window); legends are the same as Figure 2.7

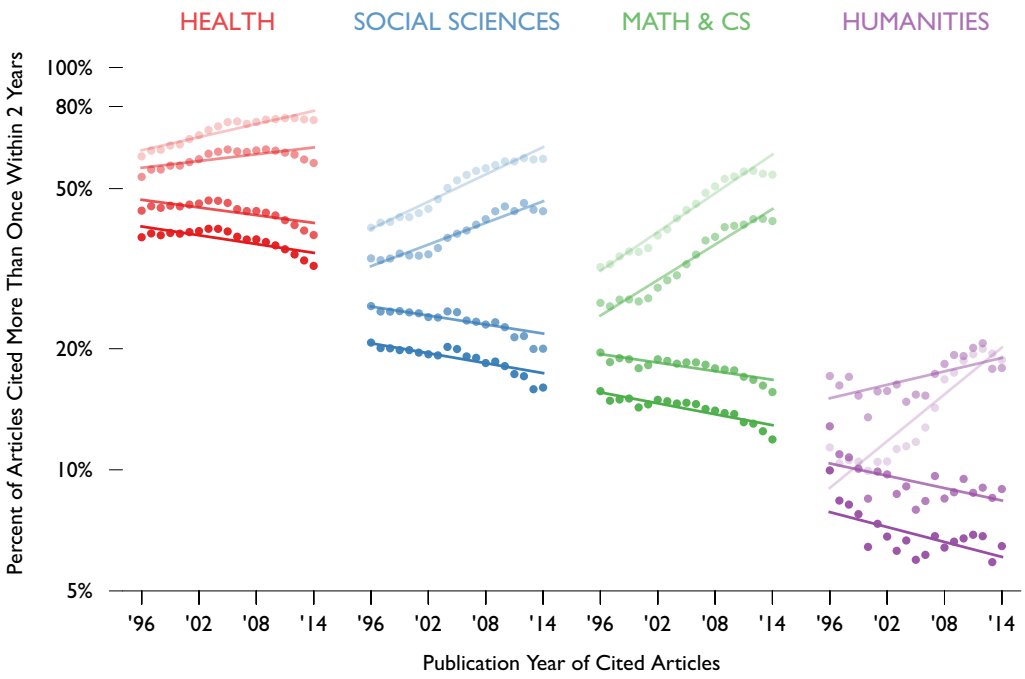
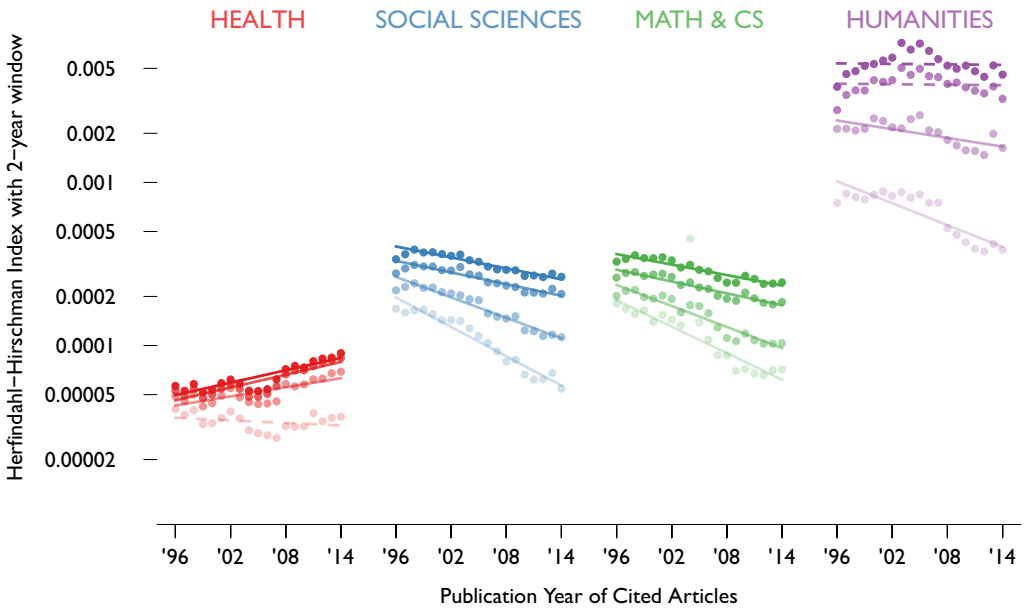


Figure A.10: The temporal trend of HHI by four broad categories between 1996 and 2014 by four broad categories (two-year citation window); legends are the same as Figure 2.7



Appendix B

APPENDIX - CHAPTER 3

B.1 Detailed model results for quantile regression analyses

Table B.1: Models to predict within-disciplinary citations with quantile regression model (tau 0.8) - Sociology, 1999-2007

	1999	2000	2001	2002	2003	2004	2005	2006	2007
Intercept	0.20*	0.27***	0.11	0.21**	0.08	0.13*	0.15*	0.22**	0.38***
	(0.08)	(0.07)	(0.07)	(0.07)	(0.07)	(0.06)	(0.06)	(0.07)	(0.08)
Age	-0.07***	-0.06***	-0.04***	-0.06***	-0.05***	-0.05***	-0.06***	-0.05***	-0.06***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Number of References	0.0001	0.0010	0.0003	0.0024*	0.0013	0.0003	0.0002	0.0013	0.0006
	(0.0008)	(0.0010)	(0.0010)	(0.0010)	(0.0009)	(0.0009)	(0.0010)	(0.0010)	(0.0008)
Page count	0.0047	-0.0021	0.0012	-0.0006	-0.0012	0.0009	0.0025	-0.0004	-0.0032
	(0.0061)	(0.0056)	(0.0065)	(0.0060)	(0.0061)	(0.0059)	(0.0066)	(0.0093)	(0.0069)
Previous cite count	0.33***	0.32***	0.35***	0.32***	0.29***	0.33***	0.30***	0.34***	0.32
	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
JIF	0.39***	0.36***	0.20**	0.27***	0.55***	0.46***	0.41***	0.25**	0.37***
	(0.09)	(0.07)	(0.07)	(0.07)	(0.09)	(0.07)	(0.07)	(0.09)	(0.08)

Standard error in ()

* < .05; ** < .01; *** < .001

Table B.2: Models to predict within-disciplinary citations with quantile regression model (tau 0.8) - Sociology, 2008-2016

.	2008	2009	2010	2011	2012	2013	2014	2015	2016
Intercept	0.12	0.16*	0.03	0.18	0.15	0.26**	0.07	0.11	0.09
	(0.08)	(0.08)	(0.09)	(0.10)	(0.10)	(0.09)	(0.10)	(0.10)	(0.09)
Age	-0.05***	-0.06***	-0.06***	-0.07***	-0.08***	-0.08***	-0.07***	-0.07***	-0.07***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Number of References	0.0011	0.0004	0.0010	0.0005	0.0015	0.0013	0.0005	0.0017	-0.0002
	(0.0011)	(0.0009)	(0.0013)	(0.0015)	(0.0015)	(0.0013)	(0.0014)	(0.0011)	(0.0010)
Page count	0.0012	0.0039	0.0017	0.0068	0.0026	-0.0008	0.0033	0.0026	0.0074
	(0.0068)	(0.0069)	(0.0059)	(0.0066)	(0.0082)	(0.0065)	(0.0098)	(0.0095)	(0.0082)
Previous cite count	0.34***	0.34***	0.34***	0.34***	0.35***	0.38***	0.37***	0.33***	0.37***
	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)
JIF	0.36***	0.31***	0.52***	0.29***	0.31***	0.30***	0.35***	0.31***	0.22***
	(0.09)	(0.08)	(0.08)	(0.08)	(0.06)	(0.07)	(0.06)	(0.06)	(0.06)

Standard error in ()

* < .05; ** < .01; *** < .001

Table B.3: Models to predict inter-disciplinary citations with quantile regression model (tau 0.8) - Sociology, 1999-2007

	1999	2000	2001	2002	2003	2004	2005	2006	2007
Intercept	0.68***	0.64***	0.67***	0.67***	0.67***	0.68***	0.71***	0.63***	0.61***
	(0.08)	(0.08)	(0.07)	(0.09)	(0.09)	(0.08)	(0.08)	(0.07)	(0.07)
Age	-0.14***	-0.14***	-0.15***	-0.15***	-0.16***	-0.16***	-0.16***	-0.15***	-0.15***
	(0.01)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	(0.02)
Number of References	0.0014	0.0018*	0.0023**	0.0021***	0.0020**	0.0015*	0.0020**	0.0026***	0.0036***
	(0.0008)	(0.0007)	(0.0007)	(0.0006)	(0.0007)	(0.0007)	(0.0006)	(0.0008)	(0.0008)
Page count	-0.0010	-0.0014	-0.0090	-0.0121	-0.0116	-0.0082	-0.0125	-0.0063	-0.0120*
	(0.0049)	(0.0058)	(0.0054)	(0.0069)	(0.0076)	(0.0064)	(0.0066)	(0.0055)	(0.0057)
Previous cite count	0.41***	0.42***	0.43***	0.42***	0.43***	0.44***	0.45***	0.46***	0.46***
	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)
JIF	0.22*	0.23**	0.26***	0.29***	0.30***	0.28***	0.27***	0.25***	0.28***
	(0.09)	(0.08)	(0.07)	(0.08)	(0.07)	(0.07)	(0.07)	(0.07)	(0.06)

Standard error in ()

* < .05; ** < .01; *** < .001

Table B.4: Models to predict inter-disciplinary citations with quantile regression model (tau 0.8) - Sociology, 2008-2016

	2008	2009	2010	2011	2012	2013	2014	2015	2016
Intercept	0.56***	0.53***	0.45***	0.42***	0.44***	0.46***	0.47***	0.45***	0.48***
	(0.06)	(0.06)	(0.07)	(0.07)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)
Age	-0.15***	-0.15***	-0.15***	-0.13***	-0.14***	-0.14***	-0.14***	-0.13***	-0.13***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Number of References	0.0035***	0.0031***	0.0026**	0.0030***	0.0023***	0.0022**	0.0025***	0.0020***	0.0022***
	(0.0006)	(0.0008)	(0.0008)	(0.0008)	(0.0005)	(0.0007)	(0.0006)	(0.0003)	(0.0006)
Page count	-0.0080	-0.0003	0.0095*	0.0118***	0.0112***	0.0097***	0.0079*	0.0084*	0.0086*
	(0.0044)	(0.0046)	(0.0048)	(0.0035)	(0.0030)	(0.0024)	(0.0033)	(0.0033)	(0.0036)
Previous cite count	0.47***	0.47***	0.48***	0.48***	0.49***	0.49***	0.49***	0.47***	0.46***
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
JIF	0.27***	0.23***	0.25***	0.21***	0.21***	0.21***	0.19***	0.19***	0.19***
	(0.06)	(0.05)	(0.06)	(0.06)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)

Standard error in ()

* < .05; ** < .01; *** < .001

Table B.5: Models to predict within-disciplinary citations with quantile regression model (tau 0.8) - Political science, 1999-2007

	1999	2000	2001	2002	2003	2004	2005	2006	2007
Intercept	0.23***	0.22**	0.21*	0.09	0.25**	0.29***	0.21*	0.16*	0.31***
	(0.07)	(0.08)	(0.10)	(0.08)	(0.08)	(0.08)	(0.09)	(0.08)	(0.07)
Age	-0.07***	-0.08***	-0.07***	-0.06***	-0.07***	-0.07***	-0.06***	-0.05***	-0.08***
	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Number of References	0.0007	-0.0003	0.0003	0.0006	0.0017*	0.0000	0.0000	0.0007	0.0010
	(0.0006)	(0.0008)	(0.0012)	(0.0008)	(0.0008)	(0.0007)	(0.0006)	(0.0009)	(0.0010)
Page count	0.0017	0.0055	0.0034	0.0050	-0.0012	0.0008	0.0010	-0.0001	-0.0027
	(0.0263)	(0.0241)	(0.0317)	(0.0248)	(0.0259)	(0.0209)	(0.0257)	(0.0257)	(0.0276)
Previous cite count	0.39***	0.40***	0.40***	0.35***	0.39***	0.40***	0.39***	0.37***	0.40***
	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)
JIF	0.41**	0.46***	0.47***	0.53***	0.46***	0.49***	0.57***	0.53***	0.47***
	(0.14)	(0.09)	(0.09)	(0.13)	(0.10)	(0.08)	(0.10)	(0.08)	(0.09)

Standard error in ()

* < .05; ** < .01; *** < .001

Table B.6: Models to predict within-disciplinary citations with quantile regression model (tau 0.8) - Political science, 2008-2016

	2008	2009	2010	2011	2012	2013	2014	2015	2016
Intercept	0.37	0.27	0.31	0.38	0.32	0.29	0.44	0.31	0.39
	(0.09)	(0.09)	(0.08)	(0.09)	(0.08)	(0.09)	(0.11)	(0.08)	(0.09)
Age	-0.07	-0.07	-0.07	-0.08	-0.08	-0.09	-0.10	-0.09	-0.10
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Number of References	0.0007	-0.0004	0.0006	0.0005	0.0007	0.0003	0.0003	-0.0005	0.0003
	(0.0007)	(0.0009)	(0.0007)	(0.0009)	(0.0008)	(0.0008)	(0.0008)	(0.0008)	(0.0007)
Page count	-0.0035	-0.0002	-0.0031	-0.0005	-0.0004	0.0010	-0.0022	0.0012	-0.0005
	(0.0032)	(0.0034)	(0.0023)	(0.0027)	(0.0023)	(0.0023)	(0.0032)	(0.0022)	(0.0031)
Previous cite count	0.42	0.42	0.45	0.43	0.48	0.43	0.49	0.47	0.48
	(0.02)	(0.03)	(0.03)	(0.02)	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)
JIF	0.47	0.55	0.43	0.38	0.29	0.44	0.37	0.40	0.40
	(0.09)	(0.08)	(0.09)	(0.09)	(0.10)	(0.11)	(0.09)	(0.06)	(0.07)

Standard error in ()

* < .05; ** < .01; *** < .001

Table B.7: Models to predict inter-disciplinary citations with quantile regression model (tau 0.8) - Political science, 1999-2007

	1999	2000	2001	2002	2003	2004	2005	2006	2007
Intercept	0.11	0.00	0.31**	0.15	0.18	0.26*	0.32***	0.23***	0.36***
	(0.07)	(0.04)	(0.11)	(0.10)	(0.09)	(0.11)	(0.09)	(0.07)	(0.10)
Age	-0.03*	0.00	-0.05***	-0.03	-0.04**	-0.04*	-0.06***	-0.05***	-0.06***
	(0.01)	(0.01)	(0.01)	(0.02)	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)
Number of References	0.0013	0.0000	0.0022	0.0009	0.0006	0.0009	0.0001	0.0012	0.0014
	(0.0009)	(0.0007)	(0.0011)	(0.0009)	(0.0012)	(0.0010)	(0.0013)	(0.0009)	(0.0009)
Page count	0.0011	0.0000	-0.0017	-0.0034	0.0050	0.0009	0.0030	0.0014	0.0013
	(0.0027)	(0.0017)	(0.0033)	(0.0022)	(0.0029)	(0.0028)	(0.0024)	(0.0027)	(0.0029)
Previous cite count	0.34***	0.36***	0.33***	0.38***	0.36***	0.37***	0.36***	0.40***	0.37***
	(0.03)	(0.02)	(0.03)	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)
JIF	0.18	0.00	0.21	0.16	-0.01	0.00	0.20	0.11	0.11
	(0.13)	(0.06)	(0.11)	(0.11)	(0.11)	(0.10)	(0.14)	(0.08)	(0.11)

Standard error in ()

* < .05; ** < .01; *** < .001

Table B.8: Models to predict inter-disciplinary citations with quantile regression model (tau 0.8) - Political science, 2008-2016

	2008	2009	2010	2011	2012	2013	2014	2015	2016
Intercept	0.47***	0.36***	0.37**	0.34***	0.35***	0.42***	0.43***	0.32***	0.34***
	(0.08)	(0.09)	(0.12)	(0.08)	(0.08)	(0.08)	(0.08)	(0.07)	(0.07)
Age	-0.07***	-0.06***	-0.07***	-0.08***	-0.07***	-0.09***	-0.09***	-0.08***	-0.10***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Number of References	0.0012	0.0009	0.0007	0.0014*	0.0016*	0.0022**	0.0005	0.0024**	0.0008
	(0.0010)	(0.0013)	(0.0013)	(0.0007)	(0.0008)	(0.0008)	(0.0008)	(0.0007)	(0.0007)
Page count	-0.0005	0.0028	0.0029	0.0031	0.0029	-0.0001	0.0022	0.0003	0.0032
	(0.0024)	(0.0031)	(0.0033)	(0.0027)	(0.0020)	(0.0023)	(0.0031)	(0.0029)	(0.0032)
Previous cite count	0.43***	0.46***	0.48***	0.46***	0.48***	0.50***	0.50***	0.49***	0.51***
	(0.02)	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
JIF	0.01	0.03	0.11	0.23*	0.12	0.11	0.12	0.08	0.14
	(0.12)	(0.12)	(0.14)	(0.09)	(0.07)	(0.08)	(0.07)	(0.08)	(0.07)

Standard error in ()

* < .05; ** < .01; *** < .001

Table B.9: Models to predict within-disciplinary citations with quantile regression model (tau 0.8) - Statistics, 1999-2007

	1999	2000	2001	2002	2003	2004	2005	2006	2007
Intercept	0.25*	0.09	0.10	0.06	0.12	0.24**	0.15*	0.18**	0.32***
	(0.10)	(0.08)	(0.08)	(0.08)	(0.07)	(0.08)	(0.07)	(0.07)	(0.07)
Age	-0.06***	-0.04**	-0.04**	-0.03*	-0.04**	-0.06***	-0.04***	-0.05***	-0.07***
	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Number of References	0.0040**	0.0045***	0.0032*	0.0020	0.0042*	0.0016	0.0008	0.0022	0.0007
	(0.0013)	(0.0012)	(0.0016)	(0.0015)	(0.0016)	(0.0015)	(0.0012)	(0.0012)	(0.0015)
Page count	0.0000	0.0037	0.0008	0.0003	0.0002	-0.0008	0.0004	0.0020	0.0030
	(0.0021)	(0.0020)	(0.0023)	(0.0020)	(0.0018)	(0.0022)	(0.0022)	(0.0018)	(0.0024)
Previous cite count	0.37***	0.40***	0.37***	0.38***	0.37***	0.39***	0.37***	0.39***	0.40***
	(0.02)	(0.03)	(0.03)	(0.03)	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)
JIF	0.43***	0.19	0.33*	0.38**	0.22**	0.40***	0.32***	0.30***	0.33***
	(0.11)	(0.10)	(0.14)	(0.13)	(0.08)	(0.08)	(0.09)	(0.08)	(0.08)

Standard error in ()

* < .05; ** < .01; *** < .001

Table B.10: Models to predict within-disciplinary citations with quantile regression model (tau 0.8) - Statistics, 2008-2016

	2008	2009	2010	2011	2012	2013	2014	2015	2016
Intercept	0.35***	0.37***	0.25***	0.26***	0.25***	0.25***	0.26***	0.16**	0.35***
	(0.05)	(0.06)	(0.07)	(0.06)	(0.06)	(0.06)	(0.05)	(0.06)	(0.06)
Age	-0.07***	-0.08***	-0.07***	-0.08***	-0.07***	-0.07***	-0.08***	-0.06***	-0.09***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Number of References	0.0008	0.0025**	0.0026*	0.0025*	0.0032*	-0.0002	0.0017	0.0014	0.0015
	(0.0012)	(0.0008)	(0.0010)	(0.0012)	(0.0013)	(0.0011)	(0.0012)	(0.0011)	(0.0010)
Page count	0.0012	-0.0009	-0.0019	0.0008	-0.0004	0.0028	-0.0013	0.0009	-0.0004
	(0.0013)	(0.0018)	(0.0022)	(0.0019)	(0.0019)	(0.0023)	(0.0020)	(0.0021)	(0.0026)
Previous cite count	0.42***	0.42***	0.42***	0.44***	0.43***	0.42***	0.45***	0.42***	0.43***
	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	(0.02)	(0.02)	(0.03)	(0.03)
JIF	0.31***	0.30***	0.31**	0.29***	0.31***	0.32***	0.35***	0.29***	0.35**
	(0.07)	(0.05)	(0.09)	(0.08)	(0.07)	(0.09)	(0.09)	(0.08)	(0.10)

Standard error in ()

* < .05; ** < .01; *** < .001

Table B.11: Models to predict inter-disciplinary citations with quantile regression model (tau 0.8) - Statistics, 1999-2007

	1999	2000	2001	2002	2003	2004	2005	2006	2007
Intercept	0.21	0.08	0.23*	0.19*	0.33**	0.31**	0.24**	0.35**	0.38***
	(0.12)	(0.07)	(0.11)	(0.09)	(0.11)	(0.10)	(0.09)	(0.12)	(0.09)
Age	-0.05**	-0.04**	-0.05**	-0.05***	-0.07***	-0.07***	-0.07***	-0.07***	-0.08***
	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	(0.02)	(0.01)	(0.02)	(0.01)
Number of References	0.0085***	0.0053***	0.0057**	0.0055***	0.0055***	0.0041**	0.0068***	0.0052***	0.0062***
	(0.0014)	(0.0016)	(0.0020)	(0.0015)	(0.0013)	(0.0015)	(0.0012)	(0.0013)	(0.0009)
Page count	-0.0058	-0.0004	-0.0019	-0.0035	-0.0026	0.0000	-0.0015	-0.0011	-0.0053*
	(0.0034)	(0.0024)	(0.0021)	(0.0024)	(0.0022)	(0.0020)	(0.0029)	(0.0017)	(0.0022)
Previous cite count	0.42***	0.40***	0.45***	0.41***	0.45***	0.47***	0.46***	0.48***	0.47***
	(0.03)	(0.03)	(0.03)	(0.04)	(0.03)	(0.04)	(0.03)	(0.03)	(0.03)
JIF	0.15	0.26*	-0.02	0.23	0.06	0.04	0.14	0.02	0.16*
	(0.19)	(0.12)	(0.11)	(0.13)	(0.12)	(0.11)	(0.09)	(0.10)	(0.07)

Standard error in ()

* < .05; ** < .01; *** < .001

Table B.12: Models to predict inter-disciplinary citations with quantile regression model (tau 0.8) - Statistics, 2008-2016

	2008	2009	2010	2011	2012	2013	2014	2015	2016
Intercept	0.38***	0.29***	0.29*	0.31***	0.24**	0.31***	0.27**	0.24***	0.30***
	(0.09)	(0.08)	(0.12)	(0.08)	(0.09)	(0.07)	(0.08)	(0.06)	(0.07)
Age	-0.08***	-0.08***	-0.08***	-0.08***	-0.07***	-0.07***	-0.07***	-0.08***	-0.09***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Number of References	0.0057***	0.0063***	0.0067***	0.0055***	0.0048***	0.0054***	0.0046***	0.0051***	0.0050***
	(0.0014)	(0.0012)	(0.0012)	(0.0012)	(0.0012)	(0.0008)	(0.0010)	(0.0009)	(0.0009)
Page count	-0.0059***	-0.0013	-0.0060**	-0.0050**	-0.0025	-0.0030*	-0.0040**	-0.0020	-0.0022
	(0.0015)	(0.0021)	(0.0021)	(0.0017)	(0.0020)	(0.0014)	(0.0013)	(0.0018)	(0.0013)
Previous cite count	0.50***	0.49***	0.50***	0.52***	0.53***	0.52***	0.53***	0.50***	0.53***
	(0.03)	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
JIF	0.13	0.11	0.23*	0.06	0.08	0.00	0.12	0.10	0.13
	(0.08)	(0.08)	(0.10)	(0.09)	(0.08)	(0.07)	(0.08)	(0.08)	(0.08)

Standard error in ()

* < .05; ** < .01; *** < .001

Table B.13: Models to predict within-disciplinary citations with quantile regression model (tau 0.8) - Mathematics, 1999-2007

	1999	2000	2001	2002	2003	2004	2005	2006	2007
Intercept	0.21***	0.10*	0.12***	0.23***	0.18***	0.17***	0.17***	0.27***	0.26***
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.03)	(0.04)
Age	-0.04***	-0.03***	-0.03***	-0.05***	-0.04***	-0.04***	-0.04***	-0.06***	-0.05***
	(0.01)	(0.01)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Number of References	0.0036***	0.0012*	0.0040***	0.0016*	0.0014	0.0017*	0.0016*	0.0021**	0.0019**
	(0.0009)	(0.0006)	(0.0007)	(0.0007)	(0.0008)	(0.0008)	(0.0008)	(0.0008)	(0.0007)
Page count	0.0018**	0.0033***	0.0009	0.0017***	0.0029***	0.0018**	0.0022***	0.0014*	0.0008
	(0.0007)	(0.0006)	(0.0006)	(0.0005)	(0.0006)	(0.0006)	(0.0005)	(0.0006)	(0.0007)
Previous cite count	0.37***	0.37***	0.36***	0.36***	0.36***	0.36***	0.36***	0.37***	0.39***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
JIF	0.31***	0.38***	0.34***	0.41***	0.37***	0.41***	0.37***	0.38***	0.37***
	(0.06)	(0.06)	(0.05)	(0.05)	(0.04)	(0.04)	(0.03)	(0.04)	(0.04)

Standard error in ()

* < .05; ** < .01; *** < .001

Table B.14: Models to predict within-disciplinary citations with quantile regression model (tau 0.8) - Mathematics, 2008-2016

	2008	2009	2010	2011	2012	2013	2014	2015	2016
Intercept	0.38***	0.36***	0.24***	0.30***	0.32***	0.31***	0.35***	0.34***	0.28***
	(0.03)	(0.04)	(0.04)	(0.04)	(0.03)	(0.03)	(0.04)	(0.03)	(0.03)
Age	-0.06***	-0.06***	-0.06***	-0.06***	-0.06***	-0.07***	-0.07***	-0.07***	-0.07***
	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)
Number of References	0.0022**	0.0031***	0.0035***	0.0021***	0.0018**	0.0018**	0.0023**	0.0015*	0.0024***
	(0.0007)	(0.0007)	(0.0007)	(0.0006)	(0.0006)	(0.0006)	(0.0007)	(0.0006)	(0.0006)
Page count	0.0005	-0.0004	0.0000	0.0016*	0.0013*	0.0005	0.0007	0.0016*	0.0023***
	(0.0005)	(0.0007)	(0.0008)	(0.0007)	(0.0006)	(0.0007)	(0.0005)	(0.0008)	(0.0007)
Previous cite count	0.39***	0.40***	0.40***	0.42***	0.43***	0.43***	0.43***	0.41***	0.40***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
JIF	0.28***	0.29***	0.36***	0.25***	0.24***	0.24***	0.24***	0.21***	0.25***
	(0.04)	(0.05)	(0.04)	(0.04)	(0.04)	(0.04)	(0.05)	(0.05)	(0.04)

Standard error in ()

* < .05; ** < .01; *** < .001

Table B.15: Models to predict inter-disciplinary citations with quantile regression model (tau 0.8) - Mathematics, 1999-2007

	1999	2000	2001	2002	2003	2004	2005	2006	2007
Intercept	0.04	0.02	0.06	0.05	0.08	0.12	0.07	0.14*	0.25***
	(0.05)	(0.03)	(0.05)	(0.03)	(0.06)	(0.07)	(0.04)	(0.06)	(0.06)
Age	-0.02	-0.01	-0.02	-0.02*	-0.02**	-0.03**	-0.02*	-0.04***	-0.05***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Number of References	0.0036**	-0.0001	0.0007	0.0022*	0.0030**	0.0026**	0.0014	0.0028**	0.0026**
	(0.0014)	(0.0007)	(0.0009)	(0.0010)	(0.0010)	(0.0010)	(0.0008)	(0.0011)	(0.0010)
Page count	0.0015	0.0034**	0.0022*	0.0024*	0.0022	0.0009	0.0001	-0.0001	-0.0007
	(0.0011)	(0.0012)	(0.0011)	(0.0012)	(0.0013)	(0.0012)	(0.0010)	(0.0012)	(0.0017)
Previous cite count	0.33***	0.32***	0.34***	0.33***	0.33***	0.35***	0.34***	0.33***	0.37***
	(0.02)	(0.02)	(0.01)	(0.02)	(0.01)	(0.02)	(0.01)	(0.01)	(0.02)
JIF	0.05	0.07	-0.01	0.03	0.04	-0.03	0.18*	0.25**	0.12
	(0.09)	(0.08)	(0.08)	(0.07)	(0.11)	(0.07)	(0.09)	(0.09)	(0.08)

Standard error in ()

* < .05; ** < .01; *** < .001

Table B.16: Models to predict inter-disciplinary citations with quantile regression model (tau 0.8) - Mathematics, 2008-2016

	2008	2009	2010	2011	2012	2013	2014	2015	2016
Intercept	0.28***	0.25***	0.25***	0.22***	0.18**	0.13**	0.16**	0.11*	0.17**
	(0.07)	(0.06)	(0.06)	(0.06)	(0.06)	(0.05)	(0.06)	(0.05)	(0.06)
Age	-0.06***	-0.06***	-0.06***	-0.06***	-0.06***	-0.06***	-0.06***	-0.05***	-0.06***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Number of References	0.0027**	0.0024**	0.0021*	0.0043***	0.0023***	0.0030***	0.0018	0.0016*	0.0018**
	(0.0009)	(0.0009)	(0.0008)	(0.0010)	(0.0007)	(0.0009)	(0.0010)	(0.0006)	(0.0006)
Page count	-0.0010	-0.0031*	-0.0026	-0.0039**	-0.0013	-0.0014	0.0003	0.0009	0.0002
	(0.0014)	(0.0014)	(0.0014)	(0.0013)	(0.0011)	(0.0009)	(0.0011)	(0.0008)	(0.0011)
Previous cite count	0.39***	0.37***	0.38***	0.38***	0.39***	0.38***	0.37***	0.36***	0.37***
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
JIF	0.14	0.26**	0.26***	0.25***	0.26***	0.28***	0.26***	0.25***	0.24***
	(0.09)	(0.09)	(0.08)	(0.07)	(0.06)	(0.07)	(0.07)	(0.06)	(0.07)

Standard error in ()

* < .05; ** < .01; *** < .001

Table B.17: Models to predict within-disciplinary citations with quantile regression model (tau 0.8) - Microbiology, 1999-2007

	1999	2000	2001	2002	2003	2004	2005	2006	2007
Intercept	0.62***	0.63***	0.58***	0.58***	0.65***	0.63***	0.55***	0.53***	0.55***
	(0.06)	(0.07)	(0.07)	(0.06)	(0.08)	(0.08)	(0.11)	(0.07)	(0.07)
Age	-0.15***	-0.15***	-0.15***	-0.15***	-0.16***	-0.16***	-0.15***	-0.14***	-0.15***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Number of References	0.0016*	0.0007	0.0012*	0.0019**	0.0004	0.0016*	0.0015*	0.0025***	0.0026***
	(0.0008)	(0.0006)	(0.0006)	(0.0007)	(0.0007)	(0.0007)	(0.0007)	(0.0006)	(0.0004)
Page count	0.0052	0.0078	0.0076*	0.0085	0.0122**	0.0077	0.0080	0.0052	0.0041
	(0.0035)	(0.0041)	(0.0034)	(0.0051)	(0.0043)	(0.0055)	(0.0057)	(0.0039)	(0.0043)
Previous cite count	0.41***	0.41***	0.42***	0.42***	0.44***	0.44***	0.45***	0.42***	0.44***
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.02)
JIF	0.12*	0.15*	0.16***	0.15**	0.12	0.11*	0.17*	0.17**	0.15**
	(0.05)	(0.06)	(0.05)	(0.05)	(0.07)	(0.06)	(0.08)	(0.05)	(0.05)

Standard error in ()

* < .05; ** < .01; *** < .001

Table B.18: Models to predict within-disciplinary citations with quantile regression model (tau 0.8) - Microbiology, 2008-2016

	2008	2009	2010	2011	2012	2013	2014	2015	2016
Intercept	0.43***	0.43***	0.40***	0.36***	0.31***	0.25***	0.30***	0.27***	0.35***
	(0.08)	(0.06)	(0.07)	(0.06)	(0.06)	(0.08)	(0.07)	(0.06)	(0.06)
Age	-0.14***	-0.14***	-0.13***	-0.14***	-0.13***	-0.12***	-0.12***	-0.11***	-0.12***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Number of References	0.0022**	0.0018***	0.0010	0.0008	0.0016***	0.0016**	0.0004	0.0009	0.0013*
	(0.0007)	(0.0005)	(0.0005)	(0.0006)	(0.0004)	(0.0006)	(0.0005)	(0.0006)	(0.0006)
Page count	0.0055	0.0028	0.0010	0.0052	-0.0001	0.0018	0.0048	0.0052	-0.0001
	(0.0042)	(0.0031)	(0.0030)	(0.0034)	(0.0034)	(0.0035)	(0.0043)	(0.0033)	(0.0039)
Previous cite count	0.42***	0.41***	0.41***	0.42***	0.41***	0.38***	0.39***	0.38***	0.37***
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
JIF	0.21***	0.23***	0.27***	0.28***	0.28***	0.28***	0.24**	0.19**	0.22**
	(0.05)	(0.04)	(0.05)	(0.04)	(0.06)	(0.06)	(0.07)	(0.07)	(0.07)

Standard error in ()

* < .05; ** < .01; *** < .001

Table B.19: Models to predict inter-disciplinary citations with quantile regression model (tau 0.8) - Microbiology, 1999-2007

	1999	2000	2001	2002	2003	2004	2005	2006	2007
Intercept	0.67***	0.68***	0.62***	0.68***	0.70***	0.73***	0.70***	0.76***	0.73***
	(0.11)	(0.07)	(0.06)	(0.08)	(0.08)	(0.08)	(0.08)	(0.11)	(0.11)
Age	-0.12***	-0.12***	-0.13***	-0.13***	-0.13***	-0.13***	-0.13***	-0.13***	-0.12***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Number of References	0.0000	0.0002	-0.0005	0.0010	0.0004	0.0008	0.0007	0.0008	0.0010
	(0.0008)	(0.0010)	(0.0006)	(0.0009)	(0.0008)	(0.0009)	(0.0008)	(0.0008)	(0.0006)
Page count	0.0008	0.0061	0.0114***	0.0034	0.0056	0.0100*	0.0089*	0.0057	0.0035
	(0.0043)	(0.0037)	(0.0032)	(0.0034)	(0.0042)	(0.0045)	(0.0044)	(0.0052)	(0.0054)
Previous cite count	0.41***	0.39***	0.41***	0.41***	0.42***	0.42***	0.42***	0.44***	0.45***
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.02)
JIF	0.10	0.08	0.12*	0.07	0.05	0.00	0.03	0.00	0.02
	(0.07)	(0.05)	(0.05)	(0.05)	(0.06)	(0.06)	(0.07)	(0.08)	(0.09)

Standard error in ()

* < .05; ** < .01; *** < .001

Table B.20: Models to predict inter-disciplinary citations with quantile regression model (tau 0.8) - Microbiology, 2008-2016

	2008	2009	2010	2011	2012	2013	2014	2015	2016
Intercept	0.74***	0.80***	0.80***	0.81***	0.64***	0.65***	0.57***	0.50***	0.51***
	(0.10)	(0.09)	(0.08)	(0.07)	(0.07)	(0.08)	(0.08)	(0.09)	(0.08)
Age	-0.12***	-0.13***	-0.13***	-0.13***	-0.13***	-0.13***	-0.13***	-0.12***	-0.13***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)	(0.01)	(0.00)	(0.00)
Number of References	0.0013	0.0005	0.0006	0.0002	0.0017*	0.0012*	0.0018**	0.0018**	0.0019***
	(0.0008)	(0.0006)	(0.0007)	(0.0004)	(0.0007)	(0.0005)	(0.0006)	(0.0006)	(0.0006)
Page count	-0.0008	0.0000	-0.0015	-0.0006	-0.0018	0.0001	-0.0017	0.0000	0.0016
	(0.0048)	(0.0042)	(0.0036)	(0.0037)	(0.0037)	(0.0039)	(0.0053)	(0.0058)	(0.0059)
Previous cite count	0.45***	0.46***	0.48***	0.49***	0.50***	0.51***	0.49***	0.47***	0.47***
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
JIF	0.03	0.00	0.01	0.01	0.10*	0.11*	0.14**	0.18*	0.18***
	(0.08)	(0.07)	(0.06)	(0.05)	(0.05)	(0.05)	(0.06)	(0.08)	(0.05)

Standard error in ()

* < .05; ** < .01; *** < .001

Table B.21: Models to predict within-disciplinary citations with quantile regression model (tau 0.8) - Cardiology, 1999-2007

	1999	2000	2001	2002	2003	2004	2005	2006	200
Intercept	0.71***	0.69***	0.68***	0.71***	0.76***	0.80***	0.72***	0.75***	0.66*
	(0.08)	(0.07)	(0.07)	(0.07)	(0.07)	(0.06)	(0.06)	(0.07)	(0.0
Age	-0.15***	-0.15***	-0.15***	-0.15***	-0.16***	-0.15***	-0.15***	-0.14***	-0.14
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.0
Number of References	-0.0028***	-0.0041***	-0.0033***	-0.0045***	-0.0045***	-0.0049***	-0.0060***	-0.0062***	-0.004
	(0.0008)	(0.0010)	(0.0009)	(0.0010)	(0.0009)	(0.0009)	(0.0010)	(0.0010)	(0.00
Page count	0.0074	0.0150**	0.0086	0.0167**	0.0155*	0.0072	0.0202**	0.0184*	0.018
	(0.0061)	(0.0056)	(0.0065)	(0.0060)	(0.0061)	(0.0059)	(0.0066)	(0.0093)	(0.00
Previous cite count	0.47***	0.46***	0.45***	0.44***	0.44***	0.44***	0.42***	0.42***	0.43*
	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.0
JIF	0.21*	0.23**	0.25***	0.19**	0.19*	0.18*	0.23**	0.21*	0.21
	(0.09)	(0.07)	(0.07)	(0.07)	(0.09)	(0.07)	(0.07)	(0.09)	(0.0

Standard error in ()

* < .05; ** < .01; *** < .001

Table B.22: Models to predict within-disciplinary citations with quantile regression model (tau 0.8) - Cardiology, 2008-2016

	2008	2009	2010	2011	2012	2013	2014	2015	2016
Intercept	0.63***	0.59***	0.64***	0.59***	0.58***	0.58***	0.58***	0.56***	0.55***
	(0.09)	(0.08)	(0.09)	(0.10)	(0.10)	(0.09)	(0.10)	(0.10)	(0.09)
Age	-0.15***	-0.15***	-0.15***	-0.15***	-0.16***	-0.15***	-0.15***	-0.14***	-0.14***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Number of References	-0.0048***	-0.0049***	-0.0050***	-0.0027	-0.0012	-0.0005	0.0002	0.0006	0.0007
	(0.0011)	(0.0009)	(0.0013)	(0.0015)	(0.0015)	(0.0013)	(0.0014)	(0.0011)	(0.0010)
Page count	0.0235***	0.0249***	0.0202***	0.0145*	0.0112	0.0075	0.0052	0.0020	0.0022
	(0.0068)	(0.0069)	(0.0059)	(0.0066)	(0.0082)	(0.0065)	(0.0098)	(0.0095)	(0.0082)
Previous cite count	0.42***	0.43***	0.44***	0.44***	0.44***	0.45***	0.44***	0.42***	0.42***
	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)
JIF	0.21*	0.24**	0.24**	0.25**	0.23***	0.24***	0.25***	0.21***	0.24***
	(0.09)	(0.08)	(0.08)	(0.08)	(0.06)	(0.07)	(0.06)	(0.06)	(0.06)

Standard error in ()

* < .05; ** < .01; *** < .001

Table B.23: Models to predict inter-disciplinary citations with quantile regression model (tau 0.8) - Cardiology, 1999-2007

	1999	2000	2001	2002	2003	2004	2005	2006	2007
Intercept	0.68***	0.64***	0.67***	0.67***	0.67***	0.68***	0.71***	0.63***	0.61***
	(0.08)	(0.08)	(0.07)	(0.09)	(0.09)	(0.08)	(0.08)	(0.07)	(0.07)
Age	-0.14***	-0.14***	-0.15***	-0.15***	-0.16***	-0.16***	-0.16***	-0.15***	-0.15***
	(0.01)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	(0.02)
Number of References	0.0014	0.0018*	0.0023**	0.0021***	0.0020**	0.0015*	0.0020**	0.0026***	0.0036***
	(0.0008)	(0.0007)	(0.0007)	(0.0006)	(0.0007)	(0.0007)	(0.0006)	(0.0008)	(0.0008)
Page count	-0.0010	-0.0014	-0.0090	-0.0121	-0.0116	-0.0082	-0.0125	-0.0063	-0.0120*
	(0.0049)	(0.0058)	(0.0054)	(0.0069)	(0.0076)	(0.0064)	(0.0066)	(0.0055)	(0.0057)
Previous cite count	0.41***	0.42***	0.43***	0.42***	0.43***	0.44***	0.45***	0.46***	0.46***
	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)
JIF	0.22*	0.23**	0.26***	0.29***	0.30***	0.28***	0.27***	0.25***	0.28***
	(0.09)	(0.08)	(0.07)	(0.08)	(0.07)	(0.07)	(0.07)	(0.07)	(0.06)

Standard error in ()

* < .05; ** < .01; *** < .001

Table B.24: Models to predict inter-disciplinary citations with quantile regression model (tau 0.8) - Cardiology, 2008-2016

	2008	2009	2010	2011	2012	2013	2014	2015	2016
Intercept	0.56***	0.53***	0.45***	0.42***	0.44***	0.46***	0.47***	0.45***	0.48***
	(0.06)	(0.06)	(0.07)	(0.07)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)
Age	-0.15***	-0.15***	-0.15***	-0.13***	-0.14***	-0.14***	-0.14***	-0.13***	-0.13***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Number of References	0.0035***	0.0031***	0.0026**	0.0030***	0.0023***	0.0022**	0.0025***	0.0020***	0.0022***
	(0.0006)	(0.0008)	(0.0008)	(0.0008)	(0.0005)	(0.0007)	(0.0006)	(0.0003)	(0.0006)
Page count	-0.0080	-0.0003	0.0095*	0.0118***	0.0112***	0.0097***	0.0079*	0.0084*	0.0086*
	(0.0044)	(0.0046)	(0.0048)	(0.0035)	(0.0030)	(0.0024)	(0.0033)	(0.0033)	(0.0036)
Previous cite count	0.47***	0.47***	0.48***	0.48***	0.49***	0.49***	0.49***	0.47***	0.46***
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
JIF	0.27***	0.23***	0.25***	0.21***	0.21***	0.21***	0.19***	0.19***	0.19***
	(0.06)	(0.05)	(0.06)	(0.06)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)

Standard error in ()

* < .05; ** < .01; *** < .001

Appendix C

APPENDIX - CHAPTER 4

C.1 Steps to match one arXiv ID to one Microsoft Academic Graph ID

The goal of the data management process is to get one Microsoft ID matched to one arXiv ID. For the records that have more than one matched Microsoft ID, I discard any records that does not have a DOI match, and has a poor title match (fuzzy ratio less than 50). Then, I keep all records that match one arxiv ID to one Microsoft ID. For the remaining, if an arxiv ID has some records that have incoming citations and some that do not, discard the records with zero incoming citations. Again, keep all records that match one arxiv ID to one Microsoft ID. For the remaining, keep the most recent record. If there are multiple most recent records, keep the most recent record with the highest Eigenfactor (generally the more cited record). For the final remaining duplicates (same pub date, same Eigenfactor), I take the lowest of the Microsoft IDs.

C.2 Model summary of the survival analysis

Table C.1: Model summary of the survival analysis

	Hep-ph	Astrophysics	Condensed matters
Cumulative citation (logged)	.89***	2.37***	1.13***
	(.13)	(.19)	(.01)
Total citation count (logged)	.20***	.14***	.19***
	(.01)	(.01)	(-.12)
Cumulative citation (logged) * Total citation count (logged)	-.09***	-.26***	-.12***
	(.02)	(.78)	(.03)
N	249,884	434,167	578,387
Number of events	33,878	73,397	86,639

Standard error in ()

* < .05; ** < .01; *** < .001

C.2.1 Model summary of the Hurdle regression model

Table C.2: Hep-ph: Model summary of the hurdle regression model for the data between 6<Month<=12

	Count			Hurdle		
	Model1	Model2	Model3	Model1	Model2	Model3
Intercept	-4.54***	-.80*	.10	-3.73***	-1.32	2.55*
	(.52)	(.36)	(.77)	(.65)	(.68)	(1.08)
Months after uploaded	.06	-.08	-.07	-.58**	-.74***	-.72***
	(.15)	(.11)	(.11)	(.21)	(.22)	(.22)
Total citation count	.68***	.21***	.09	.72***	.42***	-.14
	(.05)	(.03)	(.10)	(.05)	(.05)	(.15)
Journal influence	1.27***	.83***	-.17	1.83***	1.45***	-3.61**
	(.07)	(.06)	(.75)	(.12)	(.12)	(1.24)
Received citations in arXiv		3.82***	3.82***		10.93***	10.82***
		(.10)	(.10)		(.77)	(.78)
Journal influence* Total citation count			.14			.73***
			(.10)			(.18)
N	6,710					

Standard error in ()

* < .05; ** < .01; *** < .001

Table C.3: Hep-ph: Model summary of the hurdle regression model for the data Month>12

	Count			Hurdle		
	Model1	Model2	Model3	Model1	Model2	Model3
Intercept	-4.88***	.04	.46	-4.48***	-2.67***	-.29
	(.97)	(.81)	(1.20)	(.74)	(.75)	(1.20)
Months after uploaded	.00	-.06	-.06	-.34*	-.30*	-.30*
	(.19)	(.17)	(.17)	(.15)	(.15)	(.16)
Total citation count	.60***	.05	-.01	.74***	.44***	.11
	(.11)	(.08)	(.16)	(.09)	(.09)	(.16)
Journal influence	1.46***	.65***	.13	1.81***	1.47***	-2.31
	(.11)	(.09)	(1.11)	(.17)	(.17)	(1.61)
Received citations in arXiv		4.97***	4.95***		13.41***	13.33***
		(.29)	(.29)		(1.53)	(1.54)
Journal influence* Total citation count			.07			.54*
			(.14)			(.24)
N	2,228					

Standard error in ()

* < .05; ** < .01; *** < .001

Table C.4: Astrophysics: Model summary of the hurdle regression model for the data between 6<Month<=12

	Count			Hurdle		
	Model1	Model2	Model3	Model1	Model2	Model3
Intercept	-2.20***	.10	-1.40*	-5.12***	-3.56***	-2.41
	(.46)	(.31)	(.68)	(1.00)	(1.04)	(1.67)
Months after uploaded	-.00	-.23*	-.23*	.04	-.15	-.17
	(.13)	(.10)	(.10)	(.32)	(.31)	(.31)
Total citation count	.43***	.20***	.39***	.69***	.52***	.37*
	(.03)	(.02)	(.08)	(.07)	(.07)	(.18)
Journal influence	1.41***	1.02***	2.36***	2.82***	2.65***	1.33
	(.09)	(.05)	(.55)	(.15)	(.15)	(1.49)
Received citations in arXiv		3.11***	3.11***		10.89***	10.88***
		(.11)	(.11)		(1.17)	(1.18)
Journal influence* Total citation count			-.17**			.18
			(.07)			(.19)
N	6,539					

Standard error in ()

* < .05; ** < .01; *** < .001

Table C.5: Astrophysics: Model summary of the hurdle regression model for the data Month>12

	Count			Hurdle		
	Model1	Model2	Model3	Model1	Model2	Model3
Intercept	-3.40***	-.14	-4.42***	.35	2.21*	1.72
	(.77)	(.56)	(1.14)	(1.04)	(1.10)	(2.10)
Months after uploaded	-.17	-.19	-.20	-.98***	-.86***	-.86***
	(.15)	(.12)	(.11)	(.18)	(.18)	(.18)
Total citation count	.56***	.18***	.73***	.39***	.06	.13
	(.08)	(.06)	(.14)	(.11)	(.12)	(.25)
Journal influence	1.87***	1.17***	5.28***	2.01***	1.73***	2.37
	(.18)	(.08)	(1.00)	(.21)	(.20)	(2.47)
Received citations in arXiv		3.48***	3.50***		16.56***	16.55***
		(.21)	(.26)		(2.19)	(2.19)
Journal influence* Total citation count			-.53***			-.09
			(.13)			(.32)
N	2,003					

Standard error in ()

* < .05; ** < .01; *** < .001

Table C.6: Condensed matters: Model summary of the hurdle regression model for the data between 6<Month<=12

	Count			Hurdle		
	Model1	Model2	Model3	Model1	Model2	Model3
Intercept	-2.37***	-1.57***	-1.32**	-3.65***	-3.45***	-3.27***
	(.30)	(.25)	(.51)	(.39)	(.39)	(.81)
Months after uploaded	-.31***	-.34***	-.34***	-.45***	-.45***	-.45***
	(.09)	(.07)	(.07)	(.11)	(.11)	(.11)
Total citation count	.50***	.41***	.37***	.71***	.67***	.64***
	(.03)	(.03)	(.07)	(.04)	(.04)	(.11)
Journal influence	1.02***	.85***	.61	1.64***	1.60***	1.38
	(.03)	(.02)	(.40)	(.06)	(.06)	(.83)
Received citations in arXiv		5.37***	5.37***		9.52***	9.52***
		(.16)	(.16)		(.60)	(.60)
Journal influence* Total citation count			.03			.03
			(.06)			(.12)
N	20,140					

Standard error in ()

* < .05; ** < .01; *** < .001

Table C.7: Condensed matters: Model summary of the hurdle regression model for the data Month>12

	Count			Hurdle		
	Model1	Model2	Model3	Model1	Model2	Model3
Intercept	-3.28***	-1.81***	-1.18	-3.86***	-3.48***	-2.14
	(.54)	(.42)	(.86)	(.66)	(.68)	(1.48)
Months after uploaded	.12	.06	.06	-.10	-.11	-.10
	(.10)	(.07)	(.07)	(.12)	(.12)	(.12)
Total citation count	.44***	.31***	.21	.62***	.55***	.35
	(.07)	(.06)	(.12)	(.09)	(.09)	(.22)
Journal influence	1.14***	.67***	-.03	1.57***	1.52***	-.29
	(.08)	(.05)	(.81)	(.12)	(.13)	(1.88)
Received citations in arXiv		6.73***	6.71***		15.42***	15.37***
		(.37)	(.37)		(1.59)	(1.61)
Journal influence* Total citation count			.10			.27
			(.12)			(.28)
N	4,526					

Standard error in ()

* < .05; ** < .01; *** < .001

VITA

Lanu Kim earned the bachelor's degree in Sociology and Economics at Seoul National University in 2007, and the master's degree in Sociology at the University of Washington in 2014.