Production and Consumption in Knowledge Market: Solving the Old Puzzles with New Techniques

A Thesis
submitted to the Faculty of the
Graduate School of Arts and Sciences
of Georgetown University
in partial fulfillment of the requirements for the
degree of
Doctor of Philosophy
in Economics

By

Dongbo Guo, M.A.

Washington, DC
April 20, 2018

Production and Consumption in Knowledge Market: Solving the Old Puzzles with New Techniques

Dongbo Guo, M.A.

Thesis Advisor: John Rust, Ph.D.

Abstract

The first chapter investigates the drivers of citation counts of academic papers. I match yearly citation data, full texts, and yearly author data of 4,482 papers in the top 5 economics journals, and use textual analysis to construct high dimensional vectors of features of papers and authors. The 10-year citation distribution is highly right-skewed, and the upper tail of the distribution is well approximated by a power law. In addition, higher 10-year citation counts are associated with higher popular topic coverage, numbers of authors, and total citations of authors' co-authors, while associated with lower "Micro" intensity, paper complexity, and numbers of authors' top field publications. I use several state-of-the-art machine learning methods and develop a hybrid method that combines variable construction of dictionary-based textual analysis, variable selection of regression shrinkage, and model fitting of Gradient Boosted Trees to predict papers' 10-year citations with the information available as of the year of publication. My proposed hybrid method gives the smallest Mean Squared Error for 10-year citation out-of-sample prediction test while using a relatively small number of variables compared to other machine learning methods. It correctly predicts 72.7% of the papers that are in the upper half of the citation distribution and correctly predicts 76.7% of the papers that are in the lower half of the citation distribution.

The second chapter analyzes editorial decision making in the academic publishing process. I analyze data on keywords, abstract, referee recommendations, historical records of authors, and records of editorial decision making of 13,517 manuscripts submitted to four academic journals, linked with data on paper citation counts. I use textual analysis to analyze keywords and abstracts of each paper to construct high dimensional measures of research topics and fields. Then, I estimate the effects of features of papers, authors, and referee recommendations on editorial decision making, duration from submission to decision, and paper citations. Empirical results suggest that papers with higher referee recommendation scores, higher scientific contribution scores, lower standard deviation of referee recommendation scores, higher share of positive referee recommendations, higher coverage of popular research topics, and written by authors with longer and more solid submission history (higher number of submissions and lower rejection rate) are more likely to be published. Papers with lower coverage of popular research topics and written by authors with shorter and weaker submission history are more likely to be desk rejected. For non-desk-rejected papers, the ones with higher referee recommendation scores and lower standard deviation of the scores have shorter durations of the first round of review. The results for paper citations suggest that accepted papers on average get higher citations than rejected ones, and higher paper citation counts are associated with higher coverage of popular research topics, referee recommendation scores, and scientific contribution scores. In the prediction part, I use machine learning methods (regression shrinkage methods, Random Forest, and Gradient Boosted Trees) to predict paper citations with the information available at the time of submission. The model that uses Random Forest method, measures of publication information, measures of research fields and topics, and high dimensional measures of the appearance of popular topic words gives the best out-of-sample prediction performance. Using the preferred prediction model, I test the possibility of combining artificial intelligence (AI) and human experts in the academic publishing process. The experiment shows that the average number of cumulative citations of the published papers is more than 24% higher than all submissions. This result suggests that papers published by the human intelligence based academic publishing process turn to have higher average citations than rejected ones, even though editors may not use paper's expected citations as one of the criteria when they decide which paper to publish. As an exercise, I use the citation prediction model to decide which papers to publish based on maximizing citations. For a comparable acceptance rate as the human-based editorial process, the papers published by the algorithm have 2% higher citation counts. In addition, the average number of cumulative citations of the papers selected by the artificial intelligence from the publishable paper is 22% higher than all publishable papers. Admittedly, there are other factors that affect editors' decision on which paper to publish. However, the artificial intelligence based prediction model may help editor to identify the papers that are more likely to be highly cited from publishable papers.

INDEX WORDS: Paper citation distribution, Textual analysis, Machine learning, Scientific impact prediction, Editorial decision, Academic publishing process, Knowledge market

Acknowledgments

I am indebted to John Rust for his invaluable guidance and support. I also thank my dissertation committee members Francis Vella and Dan Cao for their constructive comments.

I thank Axel Anderson, Eric Auerbach, Oeystein Daljord, Garance Genicot, Keisuke Hirano, Whitney Newey, Bernard Salanie, Kuansan Wang, and participants in various seminars for helpful comments and discussions.

I thank research assistants at Georgetown University: Li Sun, Menghan Li, Sihan Zheng, Yulin Ding, Yizhang Zhu, Mengyang Su, Jun Chu, Jiaqi Li, Tianyi Fan, Sangyu Wang, Hong Yang, Ruhan Wang, for their help in data collection and textual analysis.

I thank my friends at Georgetown University, the World Bank Group, and the International Monetary Fund for their help and support throughout my graduate school years.

I acknowledge the data and computing facility support from Microsoft Research. I also thank the academic journals that donate their editorial decision data to this research. This research has been reviewed and approved by the Georgetown University Institutional Review Board. The IRB number is 2017-0078.

Table of Contents

Chapter

1 What Drives Paper Citations? 1
- 1.1 Introduction 1
- 1.2 Data Collection and Textual Analysis 10
- 1.3 Estimation and Prediction Strategy 31
- 1.4 Estimation Results and Discussion 40
- 1.5 Prediction Results and Discussion 49
- 1.6 Conclusion 60

2 What Drives Editor Decisions? 62
- 2.1 Introduction 62
- 2.2 Model 68
- 2.3 Data Collection and Textual Analysis 75
- 2.4 Estimation and Prediction Strategy 78
- 2.5 Results and Discussion 81
- 2.6 Conclusion 115

Appendix

A Data Collection Details 118
- A.1 Academic Data 118
- A.2 Paper Full Text 120
- A.3 Journal Databases 120
- A.4 List of Top Field Journals 122

B Textual Analysis Algorithms 123
- B.1 Algorithms for Measuring Paper Information 123
- B.2 Algorithms for Measuring Author Information 126

C Additional Figures and Tables 129

Bibliography 144

List of Figures

1.1 Average 10-year citations of papers in the top 5 economics journals . 15

1.2 Empirical cumulative distribution function of papers with $C_{i,10} \leq 1000$ 15

1.3 Complementary cumulative distribution function of paper 10-year citations 17

1.4 Kernel density estimation applied to papers with $C_{i,10} < 300$ 18

1.5 Citation paths of papers in the top 5 economics journals 1990-2011 . 19

1.6 Cloud of popular topic words in the top 5 economics journals 1990-2011 24

1.7 Popular topic coverage and complexity of papers in the top 5 economics journals 26

1.8 Evolution of author profile at the year of publication 31

1.9 Predicted distribution v.s. actual distribution 55

1.10 Correct rate of decile prediction 56

2.1 Editor's decisions in the academic publishing process 69

2.2 Rejection rate by journal 86

2.3 Number of submissions by journal 86

2.4 Duration of the first round of review 93

2.5 Duration of the first round of review: Non-desk-rejected papers v.s. desk-rejected papers 94

2.6 Citation distribution of accepted papers as of the end of November 2017 97

2.7 Citation distribution of papers as of the end of November 2017: accepted v.s. rejected 98

C.1 Average citations of papers in the top 5 economics journals at the end of 2016 129

List of Tables

1.1 Pareto distribution estimation 16
1.2 Pareto distribution estimation applied to papers with $C_{i,10} >= 300$ . . 18
1.3 Paper research field intensities: "Math, Quant Methods" as the benchmark 22
1.4 Paper topic information and presentation style information 26
1.5 Variables measuring author information 28
1.6 Author information at the year of publication 30
1.7 Effects on 10-year citation counts part I 43
1.8 Effect of word appearance on 10-year citation counts 46
1.9 Out-of-sample Mean Squared Error 52
1.10 The top 50 predictors selected by Elastic Net Model (9) 54
1.11 Kolmogorov-Smirnov test for predicted distribution with and without journal ID 55
1.12 Prediction performance test: 2 categories 58
1.13 Prediction performance test: 3 categories 59
2.1 Editorial decision by round 83
2.2 Decision making in the first round of review 84
2.3 Rejection rate by round 85
2.4 Categories of referee recommendation score 88
2.5 Categories of scientific contribution score 88
2.6 Distribution of referee recommendation and scientific contribution scores 88
2.7 Referee recommendation score by round 89
2.8 Scientific contribution score by round 90
2.9 Referee recommendation and scientific evaluation: accepted v.s. rejected 91
2.10 Duration of the academic publishing process 95
2.11 Paper citations by journal 97
2.12 Paper and author information 100
2.13 Decision making in the desk review round 102
2.14 Decision making in the first round of peer review 104
2.15 Decision making in the first round of peer review 105
2.16 Paper and author effect on the duration of the first round (non-desk-rejected) 107
2.17 Referee effect on the duration of the first round (non-desk-rejected) . 108
2.18 Assessing the effect of editorial decision making on paper citations . . 109
2.19 Assessing the effect of referee evaluations on paper citations 110
2.20 Assessing paper and author effect on paper citations 110
2.21 Out-of-sample prediction: Papers submitted to the four journals 2006-2010 112

2.22 Identifying highly cited papers: artificial intelligence v.s. human . . . 115
A.1 List of top field journals 122
C.1 Cumulative citations: Top 5 economics journals 130
C.2 The top 5 highly cited papers in the top 5 economics journals 1990-2011131
C.3 Paper research field intensities 132
C.4 Research field effect on paper citation paths 133
C.5 Research topic effect on paper citation paths 134
C.6 Author effect on paper citation paths 135
C.7 Effects on paper citation paths by journal part I 136
C.8 Effects on paper citation paths by author institution ranking part I . 138
C.9 Effects on paper 10-year citation deciles by journal part I 140
C.10 Vectors of variables in prediction models 142
C.11 Machine learning methods in prediction models 143

Chapter 1

What Drives Paper Citations?

1.1 Introduction

Academic publications can be considered as products in the "knowledge market". Researchers in the knowledge market are consumers of knowledge when they read papers, and become producers of knowledge when they publish papers. After a paper is published, it is publicly viewed by researchers, and its citations accumulate when researchers cite it in subsequently published papers. Corresponding to the total sales of products in other markets, a paper's cumulative citation counts can be viewed as a measure of the total sales of the paper in the knowledge market. However, compared with consumption in other markets, consumption in the knowledge market has not been as well studied, even though an increasing share of economic growth comes from the technological improvements and breakthroughs that are often reported in academic research.

Price [56] and Redner [57] find that the numbers of citations of papers in various academic journals are not evenly distributed, and paper citation distribution follows a power law distribution.[1] In addition, as will be shown in this chapter, the 10-year citation distribution of papers in the top 5 economics journals is also highly right-skewed, and the upper tail of the distribution is well approximated by a power law. However, this chapter studies a deeper question: What are the drivers of paper citations?

[1] The power law or Pareto distribution describes the situation where the number of samples having values greater than x decreases as a power function of x. One example of power law distribution is the size distribution of cities [28]. Gabaix [29] provides a survey on power laws in economics and finance, and Clauset et al. [17] give a survey on Power laws in empirical data in broader areas of research.

Previous studies have found that paper's ultimate impact is associated with its early citation history [67], and citations received per year change over the "life-cycle" of academic papers [2, 34, 43]. In addition, empirical evidence suggests that citation counts received by academic papers differ by research field [3, 13, 33, 51], the novelty and conventionality of paper topics [65], profile of authors [14, 61], and position in a journal [19]. Previous studies only use low dimensional features of papers and authors to explain the variation of paper citations. However, the vast majority of the variation is not explained. In addition, some potentially important determinants of paper citations (e.g., topic words of papers and authors' collaboration networks) are not included in previous studies, which may lead to omitted variable bias. To address these issues, this chapter studies whether higher dimensional features of papers and authors can help expand our understanding of the drivers of paper citations.

In addition to the studies on paper citations, Card and DellaVigna [13] and Angrist et al. [3] find the fields and style of economics papers have become more empirical, and Ellison [21] proposes a model of paper quality that explains the increasing length of academic papers. Since paper citations reflect the response of academicians to published papers, investigating the drivers of citations may help understand the evolution of economic research.

Understanding the drivers of paper citations is not only useful for understanding the consumption in knowledge market but also potentially useful in various decision-making processes in academia, especially when available resources are constrained. For instance, it could be used to help editors to pre-screen publishable papers from large numbers of submissions, to assist review committees to allocate scarce resources among competing projects in research funding applications, and to help universities and colleges to make tenure and promotion decisions. Even though the citation counts of a subset of samples (e.g., some of the rejected papers) cannot be observed, a model that can explain the variation of citation counts of the published papers and predict out-of-sample may still be used to improve the decision-making processes in academia. Previous studies have developed citation indices to compare the productivity of researchers [23, 40, 54]. Empirical studies have found that paper/author citations are associated with decision making in academia, including referee recommendations and editor decisions [14], National Science Foundation(NSF) review scores [49], promotions at universities [38, 39, 64], and elections of fellows of the econometric society [35, 36]. Hamermesh [34] provides a survey on the use of citations in economics. The potential usefulness of paper citations raises the following questions: Is it possible to predict citations of individual papers with the information available as of the year of publication? Can we use these predictions to help improve academic decision making?

This chapter investigates the drivers of paper citations and uses machine learning methods to predict paper citations with the information available as of the year of publication. I match yearly citation data, full texts, and yearly author data of 4,482 papers in the top 5 economics journals – The American Economic Review (AER), Econometrica (ECMA), Journal of Political Economy (JPE), The Quarterly Journal of Economics (QJE), and The Review of Economic Studies (RES) during 1990-2011.[2] The following facts can be seen from the data: 1. The 10-year citation distribution is highly right-skewed, and the upper tail of the distribution is well approximated by a power law. 2. The slopes of paper citation paths are quite different with each other. 3. Seminal papers differ widely in research fields, topic words, and author information. Given the noticeable differences in the citation counts of papers and the diversity in

[2] The papers in other journals were not used in this study because of the legal risk of excessively downloading paper full texts from digital journal libraries, and the computational burden of analyzing a large amount of unstructured data. The data on rejected papers of the top 5 economics journals were not available for this study.

the features of highly cited papers, higher dimensional measurement of features of papers and authors may be useful for explaining the variation of paper citations.

To measure features of papers and authors, I use dictionary-based textual analysis to parse unstructured paper data and author data. For each paper, I parse its full text to construct high dimensional vectors of variables that measure its research fields, topic words, presentation style, and journal information. For authors of each paper, I parse their publication lists to construct high dimensional vectors of variables measuring their publication records, cumulative citations, and collaboration networks. These variables are arguably the measures of paper and author information that drives researchers' citing decisions. The variables constructed in this chapter are consistent with previous studies on the drivers of paper citations. However, with the help of dictionary-based textual analysis, higher dimensional measures of these features are constructed. The dictionary-based textual analysis used in this chapter is consistent with other economic studies that use this technique, including the measurement of investor sentiment [63], media slant [30], tone in financial text [52], and economic policy uncertainty [6]. Gentzkow et al. [31] provide a survey on economic research using text as data.

The results of the textual analysis show the following facts: 1. Papers in ECMA and RES on average have higher "Mathematical and Quantitative Methods" and "Microeconomics" intensity, and cover more topics in "Mathematical and Quantitative Methods" and "Microeconomics" than the other three journals. 2. Papers in QJE have the highest average coverage of popular topics, while papers in ECMA have the highest average paper complexity.[3] 3. Average author citations of papers in AER and QJE are relatively higher than the papers in the other three journals as of the year

[3] Paper complexity is measured by the average length of sentences. Papers with many long sentences, theorems, and equations usually have higher paper complexity. The intuition is that a paper with a higher average length of sentences might be harder to read.

of publication. 4. Authors of papers in QJE have the strongest average collaboration network as of the year of publication.[4] 5. A bigger proportion of authors in ECMA and RES are from institutions in the lower quantiles of rankings of economic research. The results of the textual analysis confirm the differences in the objectives of these journals and raise the following question: What causes the "QJE effect", namely, the higher citations of papers published in this journal?

To investigate the QJE effect, paper effect and author effect on paper's long-term scientific impact, I estimate the coefficients of the measures of paper and author information as of the year of publication on the 10-year citations. The results show that the QJE effect exists for paper 10-year citations, meaning that papers in QJE get higher citations even after controlling various paper and author information, though the QJE effect decreases after more control variables are added. One potential explanation for the QJE effect might be that QJE performed better in advertising publications. It is also possible that editors of QJE preferred papers that would be highly cited, while editors of the other journals did not have strong preferences for highly cited papers. Another cause of the QJE effect could be the differences in the pools of submitted manuscripts. However, the QJE effect becomes much less important in prediction models with many variables. For the prediction model with the smallest out-of-sample Mean Squared Error, adding journal ID variable only marginally improves the prediction performance.

The estimation results confirm the importance of paper research field in determining paper citations, and papers with higher 10-year citation counts are associated with higher "Macro, Monetary Econ" intensity and lower "Micro" intensity. The results also show that higher 10-year citation counts are associated with

[4] The criteria used to measure the strength of an author's collaboration network are the total citations of her/his co-authors and the total number of top publications of her/his co-authors.

higher popular topic coverage and lower paper complexity. In addition, papers with higher 10-year citation counts are associated with the appearance of some topic words and word pairs (e.g., "gdp", "correl", "bank", "school", ("capit","share"), ("product","develop","growth")).[5]

Within the variables measuring author information, numbers of authors, numbers of authors' co-authored publications, and total citations and numbers of the top 5 publications of authors' co-authors are positively correlated with the 10-year citation counts, while numbers of authors' top field publications and numbers of top field publications of authors' co-authors have negative coefficients. However, the coefficients of author experience and numbers of authors' publications are not statistically significant in any of these regressions.

To investigate the drivers of paper citation paths, I use a quadratic function to estimate the effects of variables measuring paper information and time-varying author information on papers' cumulative citations. The results show that a steeper slope of paper citation path is associated with higher "Math, Quant Methods", "Econ Development, Growth", "Econ Systems" intensity, popular topic coverage, number of pages, number of authors, and author citations. The analysis of paper citation paths reveals substantial heterogeneity among journals. Notably, papers in QJE have extremely large positive coefficients of "Econ Development, Growth" and "Econ Systems" intensity, papers in AER and QJE are negatively affected the most by higher paper complexity, and papers in JPE benefit the most from bigger teams of highly cited authors. In addition, the heterogeneity among journals also exists in the adjusted R-squared. The regression for JPE gives the largest adjusted R-squared of 0.54, while

[5] Both dictionaries and paper texts were standardized using Porter stemming algorithm [55] to increase the "hits" of topic word detection. For instance, "correl" can not only detect the appearance of "correlate", but also the appearance of "correlation", "correlating", and any other words that contain "correl".

the regression for RES gives the smallest adjusted R-squared of 0.16. However, the coefficients seem to have no clear trend across author groups.

The estimation results could help deepen our understanding of the drivers of paper citations, while the low adjusted R-squared (less than 0.6 in all of these regressions) shows that simple regression models might be inadequate to model the variation of paper citation counts. To better model the variation of the citation counts and predict out-of-sample, I turn to use some state-of-the-art machine learning methods that can use more covariates and higher-order interactions. The machine learning methods face two challenges: 1. Constructing a map of the features of papers and authors to paper citations. 2. Assigning a weight to each variable.

Recent studies investigate the use of machine learning (including Ordinary Least Squares) in predicting human decision and improving the performance of decision making, including predicting at-risk youth [15], hiring and promoting workers [42], and improving judge decisions [44]. Einav and Levin [20], Varian [66], and Mullainathan and Spiess [53] provide surveys on the use of big data and machine learning in economic research. Compared to the other studies which use machine learning methods to predict human decision, predicting paper citations is challenging due to the difficulty of measuring and assigning appropriate weights to a large number of features of papers and authors.

I compare a variety of state-of-the-art machine learning methods, including regression shrinkage models (Lasso, Post-Lasso, Ridge, and Elastic Net) in Zou and Hastie [69] and Belloni et al. [9], Neural Network [1, 11], Random Forest [12], and Gradient Boosted Trees [25, 26] in terms of their ability to predict papers' 10-year citation deciles using the information available as of the year of publication. Based on my evaluation of the advantages and disadvantages of the state-of-the-art machine learning methods, I develop a hybrid method that combines variable construction of dictionary-based textual analysis, variable selection of regression shrinkage, and model fitting of Gradient Boosted Trees for prediction with textual data.[6]

The Mean Squared Error (MSE) of various prediction methods, except for Ordinary Least Squares (OLS), generally decrease after adding more predictors constructed by textual analysis. The Shrinkage-Gradient Boosted Trees Hybrid method proposed in this chapter gives the smallest MSE in 10-year citation out-of-sample prediction test, while only using a relatively small number of predictors compared to other machine learning methods. This property of the hybrid method significantly reduces the cost of data collection and computation for using it to predict 10-year citation counts of a new paper. However, it seems hard for the prediction models to predict the citation counts of seminal papers. Even the prediction model with the smallest Mean Squared Error (MSE) cannot predict the citation counts of papers in the highest and lowest deciles well. One potential explanation could be that some important features of papers in the highest and lowest deciles are not well captured by the variables used in the analysis.

In a test of applying these machine learning methods to the academic publishing process, the hybrid method predicts papers that are in the upper half of the citation distribution correctly 72.7% of the time and predicts the papers that are in the lower half of the citation distribution correctly 76.7% of the time. In addition, within the papers being predicted by the hybrid method to be "highly cited" (the top 30% of the distribution), 65.0% of them turn out to be "highly cited", and only 4.7% of them turn out to be "lowly cited" (the bottom 30% of the distribution). Within the papers being predicted to be "lowly cited", 66.7% of them turn out to be "lowly cited", and only 2.6% turn out to be "highly cited" after 10 years of publication.

[6] Gentzkow et al. [31] summarize the general steps to analyze text as data. The hybrid method I propose is consistent with their general steps while focusing on improving the accuracy of prediction with textual data.

Based on the prediction results, the hybrid method proposed in this chapter may be helpful in identifying articles that will turn out to be lowly cited to enable editors to reject a significant fraction of inappropriate submissions, thereby allowing the editors to focus their scarce time evaluating the more promising subset of submissions to their journals. In addition, its performance in identifying highly cited papers may be helpful in preventing rejection of submissions that will turn out to be highly cited.

One concern about the hybrid prediction model might be that it discriminates some types of authors or papers by assigning enormous negative weights to few features of authors or papers. However, since the preferred hybrid prediction model predicts paper citations using hundreds of features of authors and papers, and there is not any feature with dominant weight, it is not likely to severely discriminate against some specific types of authors or papers. On the contrary, using the hybrid prediction model to help editors in editorial decision making may attenuate possible discrimination in the academic publishing process because it captures and "objectively" assigns weights to hundreds of features of papers and authors that human referees may ignore. On the other hand, there are potential dangers to using algorithms that recommend certain papers for rejection and others for acceptance based on specific keywords or other features of the paper. If authors discover the criteria used by these algorithms, it could encourage strategic behavior or "gaming" in the way scientific papers are written in a way that may be mostly cosmetic rather than encouraging authors to invest more in improving the true quality of their submissions.

This chapter contributes to existing literature in the following aspects. First, this chapter contributes to the research on the scientific impact of academic papers by investigating the factors that explain the variation of paper citations, as well as the factors that predict paper citations. The findings in this chapter may not only deepen our understanding of the drivers of paper citations, but also contribute to the work on designing data-driven models to predict scientifically impactful work. Admittedly, the number of citations is not a perfect measure of a paper's scientific impact. However, it is among the few quantitative indicators of a paper's impact. Second, the estimation and prediction strategy developed in this chapter has potential to be used to investigate the drivers of decision making and predict decision making in other markets, where information in textual data are likely to be important drivers of decision making. Third, this chapter makes extensive use of unstructured data collection techniques, textual analysis and machine learning methods that may shed new insights into the application of these techniques in other economic studies that use large-scale high dimensional data.

In Section 1.2, data collection, textual analysis, and descriptive statistics are presented. Section 1.3 discusses estimation and prediction strategy. Section 1.4 presents estimation results. Section 1.5 presents prediction results. Section 1.6 concludes.

1.2 Data Collection and Textual Analysis

1.2.1 Data collection

I collected and matched yearly citation data, full texts, and yearly author data of 4,482 papers in the top 5 economics journals. The main focus was on analyzing the data of papers in the top 5 economics journals for three reasons. First, the top 5 economics journals are general interest economics journals, and the paper dataset arguably can represent the major research fields and topics of economic research. Second, the digital journal libraries prohibit massive downloading, and downloading a large number of full texts of papers published in a broader range of journals may violate their terms and conditions. Third, the computational burden of collecting and analyzing a larger amount of unstructured paper data and author data was beyond the capacity of the computing facility available at the time of data collection and textual analysis. The data collection was conducted between January and August 2017. The details of data collection are documented in Appendix A.

Academic data

I collected academic data including paper citation lists, paper information (including the journal of publication, publication date, title, abstract, and author name list), and author publication lists from Microsoft Academic (MA) database. An overview of Microsoft Academic database is provided by Sinha et al. [60]. I used MA database as the main source of the academic data for this chapter because of the abundance of its academic data and the efficiency of using MA Application Programming Interface (API) to query and collect data from it. Due to the noisy nature of large-scale academic data, the raw academic data collected from the MA API might have measurement error in paper citations and author publications. To reduce measurement error in the academic data, I used preprocessing algorithms to check whether the retrieved author names matched with the names on the paper, subtract duplicated citations, and removed mistakenly listed publications. The preprocessing algorithm for matching Microsoft Academic data and paper data is described in Appendix A.1.1, and the preprocessing algorithm for increasing the accuracy of author information is described in Appendix B.2.

Google Scholar (GS) is another online database of paper citation data. However, the built-in barriers requested by the publishers to foil automatic queries and the other constraints in data retrieval increases the difficulty of collecting a large amount of paper and author information from its database. A comparison of the MA database and the GS database is presented in Appendix A.1.3.

Paper full texts

The full texts of papers in the top 5 economics journals during 1990-2011 were collected from digital journal libraries (ScienceDirect and JSTOR).[7]

For the top 5 economics journals, I excluded papers in The American Economic Review: Papers and Proceedings, papers published as comments and replies, papers that have less than 11 pages[8], papers that could not be recognized by the optical character recognition algorithm, and papers that did not return a correct paper ID from MA database when I attempted to look up its academic data. After these exclusions, my dataset contained 4,482 papers, with 1,299 papers in The American Economic Review, 941 papers in Econometrica, 754 papers in Journal of Political Economy, 767 papers in The Quarterly Journal of Economics, and 721 papers published in The Review of Economic Studies.

Other data

In addition to the data sources above, I also collected the following data in public domain: keywords in JEL classification codes on the AEA website[9], a list of selected adjective words, a list of advanced words, a list of "top field journals"[10], location

[7] I only collected the full texts of 9,241 papers (including comments, replies, and short papers) published in 1990-2011 because some of the papers published before the 1990s were stored as scanned documents that were significantly harder for the optical character recognition algorithm to parse, and the papers published after 2011 only had less than five years to accumulate citations at the time when I collected paper citation data.

[8] Many of the papers that have less than 11 pages are short notes, and their structures are quite different from full-length papers. These papers are not in the scope of this study. The numbers of pages of papers in The American Economic Review before 2008 were doubled in page counting because papers in The American Economic Review before 2008 are printed in two columns per page.

[9] https://www.aeaweb.org/jel/guide/jel.php

[10] The "top field papers" were the papers published in 73 of the 75 "good" or above journals ranked by [59]. The two excluded journals were IMF Staff Papers and The Journal of Law, Economics, and Organization. They were not included because papers published in

information of academic institutions from OpenStreetMap[11], and economic research score from The Tilburg University Economics Schools Research Ranking[12]. These data were used in the paper and author information measurement.

1.2.2 Paper data analysis

For each paper, I counted its yearly citations, constructed categorical variables based on its journal information, and measured its research fields, topic words, and presentation style[13] using paper text in the first 10 sentences, the first 100 sentences, the first 200 sentences, and the full text.[14]

Paper citations

Paper citation paths were constructed based on the data from MA database. I searched for each paper's paper ID in the MA database using the paper title. Then, I compared the author names in the matching entry in the MA database with the actual author names to check whether the returned paper was correct or not. If the returned paper had matched title and authors, I used the publication year of each paper in paper's citation list to count the number of citations after t years of publication, and constructed $C_{i,t}$: the cumulative citations of paper i after t years of publication.

these two journals could not be retrieved from Microsoft Academic database at the time when I collected academic data. The list of the journals is shown in table A.1.

[11] https://www.openstreetmap.org

[12] https://econtop.uvt.nl/

[13] The measures of paper's presentation style include the "descriptiveness" of paper's writing style, the "richness" of paper's vocabulary, the "complexity" of paper's sentences, and the number of pages. A detailed explanation will be provided in Section 1.2.2.

[14] For the papers with abstract, the first 10 sentences are approximately the abstract. For the papers without an abstract, the first 10 sentences are approximately the first paragraph. The first 100 sentences are approximately the introduction section of most papers. The first 200 sentences are approximately the first half of the text of most papers.

Figure 1.1 shows the average 10-year citations of papers by publication year in the top 5 economics journals. The average 10-year citations of papers in all of these journals increase over time, and the curve of QJE locates relatively higher than the other four journals. Figure C.1 shows the average cumulative citations of papers in the top 5 economics journals at the end of 2016. The trend in Figure C.1 is consistent with the citation trend of a broader range of papers in the top 5 economics journals presented by [13] and [51]. As shown in Figure C.1, the papers published in recent years generally have fewer cumulative citations partially due to the shorter publication years.

Figure 1.2 shows the empirical cumulative distribution function of the 10-year citations of papers with $C_{i,10} \leq 1000$.[15] It can be seen that almost all of the papers published in the top 5 economics journals have less than 1,000 citations, and about 90% of them have less than 300 citations after 10 years of publication. The empirical cumulative distribution function of QJE is located to the right of the curves of the other journals, showing that the distribution of the 10-year citations of papers published in QJE stochastically dominate the citations of papers in the other top economics journals. Table C.1 presents the citation statistics of the top 5 economics journals.[16] On average, the papers in QJE have the highest average citations, while the standard deviations of 10-year citations in these journals are all quite large.

[15] The papers with more than 1,000 citations are truncated because the cumulative densities at 1,000 are very close to 1.

[16] The original format of Table C.1 and some other tables in this chapter were created using the R package: stargazer [41], and then realigned by the author.

Figure 1.1: Average 10-year citations of papers in the top 5 economics journals

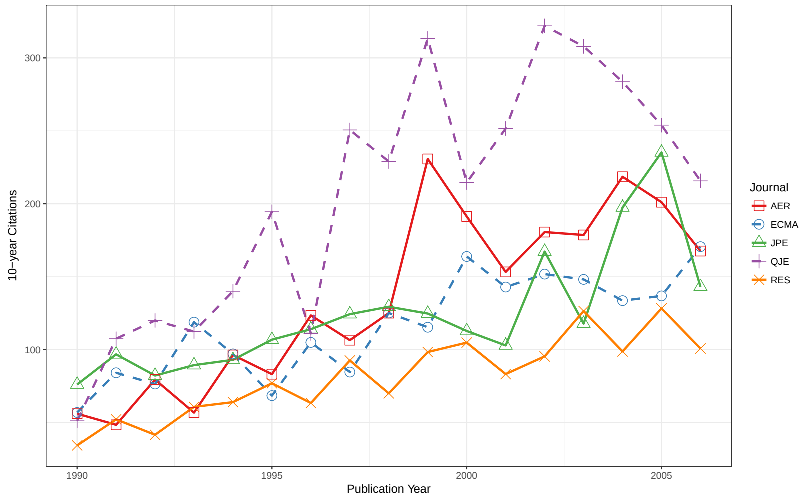
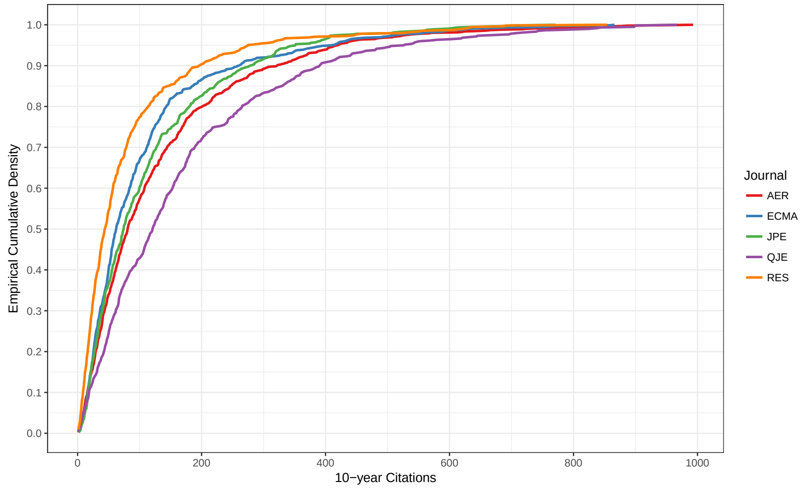


Figure 1.2: Empirical cumulative distribution function of papers with $C_{i,10} \leq 1000$



In academic journals, it is often observed that a few seminal papers account for a disproportionate amount of the published papers' total citations. Thus, to show the distribution of seminal papers, I use the Pareto (power law) distribution. The survival function of Pareto distribution is shown in Equation 1.1.

$$\overline{F}(x) = Pr(X > x) = \begin{cases} (\frac{x_m}{x})^{\alpha} & \text{for} \quad x \geq x_m \\ 1 & \text{for} \quad x < x_m \end{cases} \tag{1.1}$$

where α is the shape parameter, which describes the degree of concentration of the distribution, and x_m is the scale parameter.

Table 1.1 presents the result of Pareto distribution estimation. The hypothesis testing suggests accepting that the data is generated from a power law distribution for all journals except for ECMA. Figure 1.3 shows the complementary cumulative distribution function (CCDF) of paper 10-year citations with logarithmic horizontal and vertical axes.[17] It can be seen that the CCDF of paper citations in all journals, except for ECMA, is well approximated by a straight line, indicating a power law distribution.

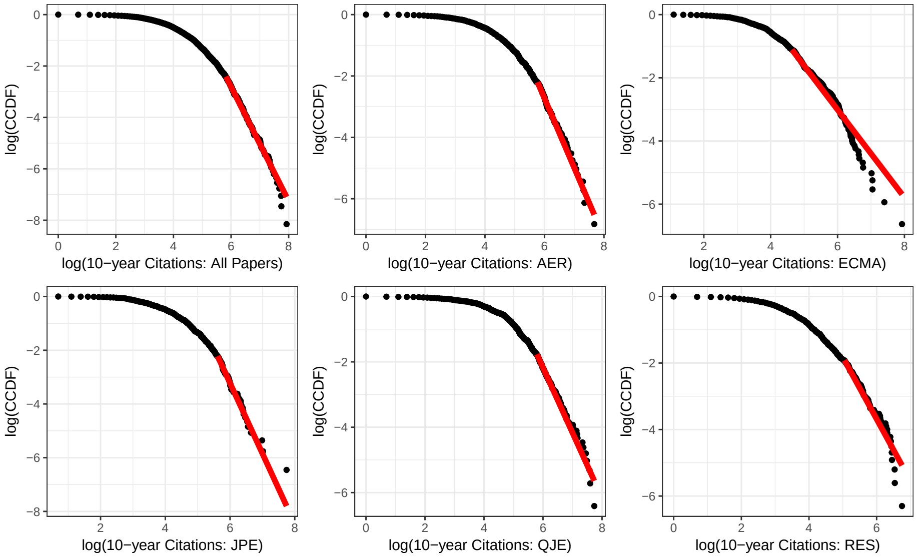
Table 1.1: Pareto distribution estimation

	All Top 5	**AER**	**ECMA**	**JPE**	**QJE**	**RES**
α	3.23	3.25	2.40	3.54	2.98	2.75
	(0.17)	(0.38)	(0.12)	(0.48)	(0.29)	(0.39)
x_m	352.43	309.86	105.97	252.21	324.93	135.83
	(56.39)	(77.64)	(27.18)	(55.17)	(75.99)	(56.19)
Observations	3,472	923	759	636	609	545
P-value	0.46	0.82	0.01	0.35	0.97	0.23

Note: The standard errors are estimated via 1,000 times of bootstrapping by Numerical Maximum Likelihood Estimation algorithm in Gillespie [32]. The P-value in each column is generated via 1,000 times of bootstrapping by the algorithm in Clauset et al. [17], and it quantifies the plausibility of the null hypothesis that the data is generated from a power law distribution.

[17] Complementary cumulative distribution function or survival function is defined to be $P(x) = Pr(X > x) = 1 - F(x)$.

Figure 1.3: Complementary cumulative distribution function of paper 10-year citations



To capture the pattern near the bulk of the distribution and the pattern of the upper tail separately, I partition papers into two groups based on their 10-year citations and set the cutoff point as 300 (the cutoff between the roughly top 10% and the other papers). For the papers with less than 300 citations, I use Gaussian kernel density estimation to fit the distribution. For the papers with at least three hundred citations, I use Pareto distribution estimation.

Figure 1.4 shows the kernel density estimation of 10-year citation distribution of papers with $C_{i,10} < 300$. The citation distribution is highly right-skewed, meaning that while most of the published papers have relatively low citations, a small portion of them are very highly cited, producing the long tail to the right of the distribution.

Table 1.2 presents the results of the Pareto distribution estimation for papers with $C_{i,10} >= 300$. The QJE has the smallest α among the top 5 economics journals, meaning the thickest upper tail.

Figure 1.4: Kernel density estimation applied to papers with $C_{i,10} < 300$

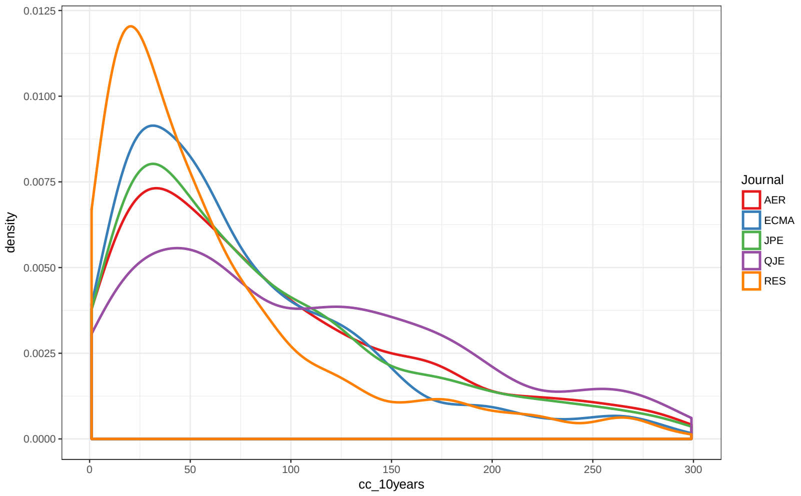


Table 1.2: Pareto distribution estimation applied to papers with $C_{i,10} >= 300$

	All Top 5	**AER**	**ECMA**	**JPE**	**QJE**	**RES**
α	3.05	3.08	2.99	3.83	2.83	3.32
	(0.11)	(0.21)	(0.25)	(0.38)	(0.18)	(0.49)
Observations	368	107	66	58	112	25
Share of papers with $C_{i,10} >= 300$	10.6%	11.6%	8.7%	9.1%	18.4%	4.6%

Note: The standard errors are estimated via 1,000 times of bootstrapping by Numerical Maximum Likelihood Estimation algorithm in Gillespie [32], constraining x_m. x_m is constrained at 300.

The papers in the top 5 economics journals are not only different in their 10-year citations, but also quite different in their citation paths. Figure 1.5 shows the citation paths of papers by 10-year citation quantiles. It can be seen that the slopes of paper citation paths are quite different from each other. The slope of citation path of papers in the top 25% is increasing over time, while the citation path of papers in the bottom 25% is almost a straight line with a lower slope. The discrepancy in paper citation paths raises the following questions: Is the discrepancy in paper citation paths caused by the differences in paper contents and author profiles? What are the typical features of highly cited papers?

Figure 1.5: Citation paths of papers in the top 5 economics journals 1990-2011

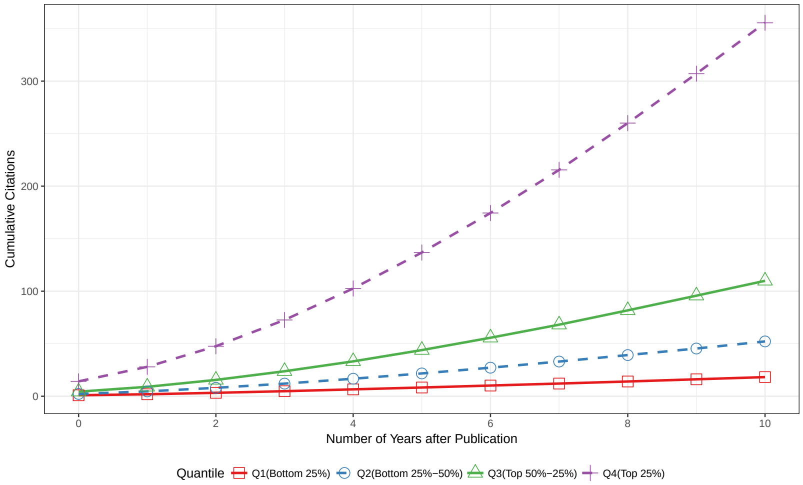


Table C.2 lists some examples of seminal papers in the top 5 economics journals. The rank of papers is based on the citation counts in MA database.[18] Seminal papers cover different topics in different research fields, ranging from applied economics papers (e.g., the highly cited QJE paper by Paolo Mauro) to econometric theory papers (e.g., the highly cited RES paper by Charles Manski). In addition, the profiles of the authors of these papers are also quite different. For instance, the seminal JPE paper by Paul Krugman was published 14 years after he received his Ph.D. degree, while the seminal ECMA paper by Marc Melitz was his first publication. Thus, it seems hard to predict seminal papers using simple measures of the research field and author profile of papers.

In the remaining parts of this section, I describe the process of measuring high dimensional features of papers and authors using dictionary-based textual analysis.

[18] The citation counts in GS database are about double the number of the citation counts in MA database. The difference in citation counts is caused by the difference in their citation counting algorithms. The algorithm used by MA may undercount paper citations, while the algorithm used by GS may overcount. However, the ranks of paper citations in MA and GS are in general consistent.

The dictionary-based textual analysis algorithms rely on the prior information supplied by the econometrician to create dictionaries that are used to parse textual data. If the econometrician's prior information is reliable, the dictionary-based textual analysis will be able to capture potentially important features that might be hard for unsupervised learning methods to capture.[19] In addition, compared to unsupervised learning methods, the variables constructed by dictionary-based textual analysis are usually easier to interpret.

Research fields of papers

I measured the research fields of papers by parsing paper texts using research field dictionaries based on the keywords in JEL classification codes.[20] First, I created research field dummy variables equal one if any keyword listed in a JEL classification code appears in paper text.[21] Second, I calculated the percentage of research fields identified in the text and named it as the coverage of research fields. Third, I created research field intensity variables, which measured the frequency of keywords in each research field classified by JEL codes. Fourth, I manually checked the measurement of research fields of 20 randomly selected papers. The measurements of research fields of these papers were consistent with my understanding of their research fields, though the values of research field intensities were very similar for related fields (e.g., macroeconomics and financial economics). Compared to research field measurement that only assigned one research field to each paper (e.g., Anauati et al. [2] and Angrist et al.

[19] Unsupervised learning, as another branch of machine learning, relies on the algorithm to find hidden patterns in data.

[20] The keywords in field A (General Economics and Teaching), B (History of Economic Thought, Methodology, and Heterodox Approaches), N (Economic History), Y (Miscellaneous Categories), and Z (Other Special Topics) were excluded because of the small number of available keywords in these JEL classification codes.

[21] It is possible and common for papers to have more than one research field.

[3]), the method used in this chapter "objectively" created a continuous measure of the closeness of the paper and each research field.

The vector of variables measuring a paper's research fields ($\boldsymbol{F_i}$) includes the dummy variable for each research field, coverage of research fields, and research field intensity in the first 10 sentences, the first 100 sentences, the first 200 sentences and the full text of each paper.

Table 1.3 presents the relative research field intensities of each journal, and the intensity of "Math, Quant Methods" of each journal is used as the benchmark.[22] For each journal, if the intensity of some research field other than "Math, Quant Methods" is higher than 1, it means the frequency of keywords from that field is higher than the frequency of keywords from "Math, Quant Methods". It can be seen from Table 1.3 that papers in ECMA and RES have relatively higher "Mathematical and Quantitative Methods" and "Microeconomics" intensities than the intensities of the other research fields, while the research fields of papers in the other three journals are more evenly distributed.

[22] The absolute value of research field intensities are presented in Table C.3.

Table 1.3: Paper research field intensities: "Math, Quant Methods" as the benchmark

	Mean				
	AER	**ECMA**	**JPE**	**QJE**	**RES**
Math, Quant Methods	1.00	1.00	1.00	1.00	1.00
Micro	1.60	1.21	1.72	1.54	1.76
Macro, Monetary Econ	0.69	0.40	0.70	0.72	0.62
International Econ	0.60	0.40	0.57	0.61	0.57
Financial Econ	0.42	0.26	0.43	0.43	0.38
Public Econ	0.40	0.21	0.39	0.41	0.35
Health, Education, Welfare	0.36	0.23	0.37	0.41	0.30
Labor Econ	0.56	0.30	0.59	0.65	0.43
Law, Econ	0.09	0.05	0.11	0.11	0.08
Industrial Organization	0.73	0.44	0.72	0.76	0.65
Business Econ, Marketing	0.40	0.23	0.39	0.41	0.35
Econ Development, Growth	0.53	0.30	0.52	0.61	0.46
Econ Systems	0.47	0.26	0.48	0.50	0.41
Agricultural, Environmental Econ	0.16	0.07	0.15	0.15	0.11
Urban Econ	0.18	0.12	0.17	0.20	0.14

Note: The field intensity of "Math, Quant Methods" of each journal is set to be 1 and is used as the benchmark. The field intensities of the other fields are the relative intensities compared with the field intensity of "Math, Quant Methods" of each journal. The measures of paper research fields are constructed by parsing papers' full texts.

Topic words of papers

I used a dictionary based on the keywords in JEL classification codes to parse the texts in each section and created dummy variables for the appearance of highly frequent keywords and keyword pairs, as well as popular topic word and word pair coverage variables measuring the coverage of highly frequent keywords and keyword pairs.[23] The algorithm for measuring topic words of papers is detailed in Appendix B.1.1.

[23] Due to the large number of keywords in JEL classification codes, I only used 405 highly frequent keywords, 566 highly frequent two-keyword pairs, and 594 highly frequent three-keyword pairs in creating dummy variables and measuring coverages.

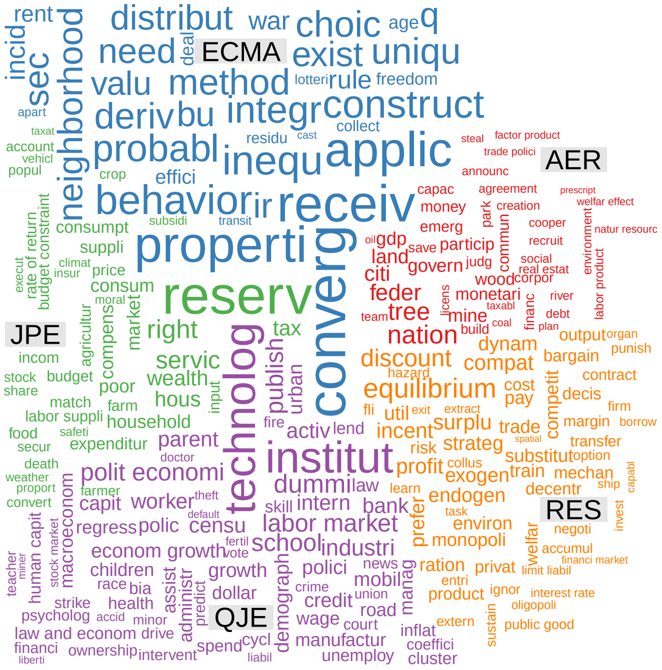
The vector of variables measuring a paper's topic words ($\boldsymbol{W_i}$) includes popular topic coverage, popular two-word pair coverage, popular three-word pair coverage, dummy variables for highly frequent topic words, dummy variables for highly frequent two-word pairs, and dummy variables for highly frequent three-word pairs in the first 10 sentences, the first 100 sentences, the first 200 sentences, and the full text of each paper.

Figure 1.6 presents the cloud of popular topic words in the top 5 economics journals. The size of word measures the level of deviation from the average share of papers having that word in the top 5 economics journals. The color of each word is determined by the journal that has the largest share of papers having that word.

It can be seen that the area of AER is the smallest, meaning that it publishes papers related to a wide range of research fields. ECMA publishes relatively bigger share of papers having topic words related to microeconomics (e.g., "choic"), and mathematical and quantitative methods (e.g., "converg", "probabl", "properti"). JPE publishes relatively bigger share of papers having topic words related to macroeconomics and public economics (e.g., "reserv", "servic", "budget"). QJE publishes relatively bigger share of papers having topic words related to some applied fields including development (e.g., "technolog", "institut"), labor (e.g., "labor market"), and educational economics (e.g., "school"). RES is similar to ECMA in publishing relatively bigger share of papers having topic words related to microeconomics, and mathematical and quantitative methods, but leaning to different topic words (e.g., "equilibrium", "dynam", "discount").[24]

[24] The topic words are standardized using Porter Stemming algorithm. For instance, "converg" may be standardized from: convergence, converge, and converging.

Figure 1.6: Cloud of popular topic words in the top 5 economics journals 1990-2011



Presentation style of papers

I measured the presentation style of papers from several aspects: first, the frequency of selected adjective words[25], which measures the "descriptiveness" of paper's writing style; second, the frequency of selected advanced words[26], which measures the "richness" of paper's vocabulary; third, the average length of sentences measures the "complexity" of paper's sentences; fourth, the number of words and the number of pages, which measures the length of papers. I did not create either dummy variables for adjective words or advanced words, because these words might be too noisy to be credible measures of the academic contribution of a paper. For instance, the

[25] Some examples of the selected adjective words are: innovative, first, extensive, and unique.

[26] These advanced words are selected from the word lists of Graduate Record Examinations(GRE) available in the public domain. Some examples of the selected advanced words are: ameliorate, exacerbate, utilize, and retrieve.

appearance of "innovative" does not necessarily mean a paper is as innovative as it has claimed. However, the frequencies of these words can be used as indicators of the presentation style of a paper. The algorithm for measuring presentation style of papers is presented in Appendix B.1.2.

The vector of variables measuring a paper's presentation style ($\boldsymbol{P_i}$) includes frequency of selected adjective words, frequency of selected advanced words, number of words, and the average length of sentences in the first 10 sentences, the first 100 sentences, the first 200 sentences, and the full text of each paper, and number of pages of each paper.

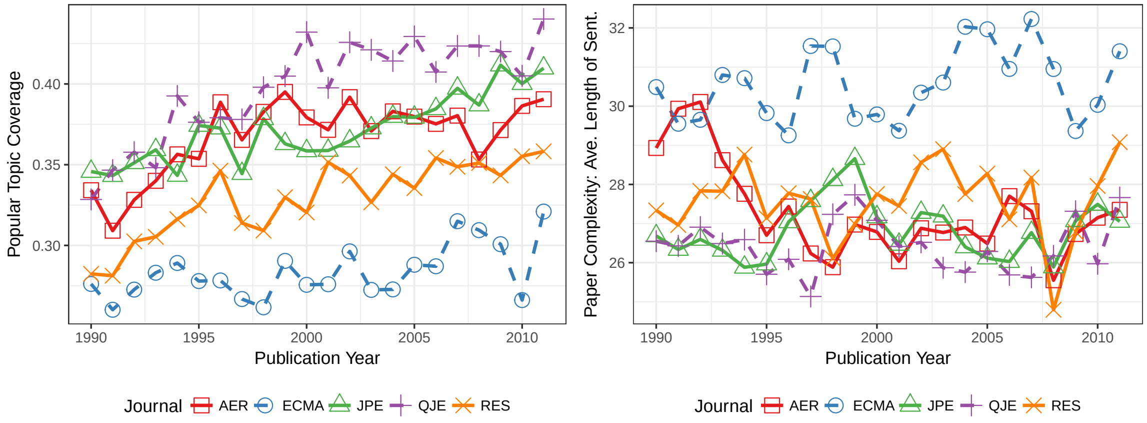
Table 1.4 presents the statistics of some selected variables measuring paper topic information and paper presentation style information, and Figure 1.7 shows time series of coverage of popular topics and paper complexity (measured by the average length of sentences) as of the year of publication. Notably, all of these journals are publishing papers covering more popular topics than they were in the 1990s, and papers in QJE have the highest average popular topic coverage, while papers in ECMA have the lowest. In addition, papers in ECMA have the highest average paper complexity, and papers in QJE have the largest average number of pages.

Table 1.4: Paper topic information and presentation style information

	Mean(Standard Deviation)				
	AER	**ECMA**	**JPE**	**QJE**	**RES**
Popular Topic Coverage (First 10 Sent.)	0.07	0.05	0.06	0.06	0.06
	(0.02)	(0.02)	(0.01)	(0.02)	(0.02)
Popular Topic Coverage (First 100 Sent.)	0.22	0.17	0.23	0.22	0.20
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Popular Topic Coverage (First 200 Sent.)	0.29	0.22	0.29	0.29	0.25
	(0.05)	(0.06)	(0.05)	(0.05)	(0.05)
Popular Topic Coverage (Full Text)	0.37	0.28	0.37	0.39	0.33
	(0.08)	(0.08)	(0.07)	(0.07)	(0.07)
Presentation: Descriptiveness	0.0005	0.0003	0.0005	0.0005	0.0004
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)
Presentation: Vocabulary Richness	0.03	0.03	0.03	0.03	0.03
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Presentation: Complexity	27.24	30.55	26.81	26.42	27.60
	(4.02)	(3.97)	(3.67)	(2.93)	(3.71)
Num. Pages	33.57	29.65	30.32	34.21	24.55
	(12.67)	(11.98)	(10.05)	(10.69)	(7.06)

Note: "Sent." is an abbreviation of "Sentences". The measures of paper presentation style are constructed by parsing papers' full texts.

Figure 1.7: Popular topic coverage and complexity of papers in the top 5 economics journals



Journal information of papers

I retrieved journal of publication information from digital journal libraries and publication year information from MA database. The publication year information of paper in MA database was assigned based on the first date when the paper was accessible publicly, which could be years ahead or behind of the journal date. I manually checked the publication years of 20 randomly selected papers and found almost all of them were assigned correct publication year. The vector of variables measuring a paper's journal information ($\boldsymbol{J_i}$) includes dummy variables for publication years and dummy variables for the top 5 economics journals.

1.2.3 Author data analysis

I measured author information using the data from MA database, as well as data from other sources linked to the author. First, I measured each author's academic experience by measuring the number of years of author citation record. Second, I calculated the number of author's publications (top 5, top field, and total publications) and author citations at the end of each year. Third, I constructed variables measuring the publication records of author's co-authors. Fourth, I used author's institution information in MA database and economic research score linked to that institution to measure economic research score of author's institution and the country of author's institution. The algorithms for measuring author information are documented in Appendix B.2. The vector of variables measuring a paper's author information is denoted as $\boldsymbol{A_{i,t}}$.

Table 1.5 summarizes the variables that were constructed to measure author information (including author's collaboration network information). The number of authors can be used to estimate the effect of the size of author team on paper citations. For the other variables measuring author information, if a paper had more than one author, I used the average value of all of the authors' information.[27] Though a more sophisticated measurement of the average strength of authors is possible, the

[27] For the papers with one or several authors having missing information, I assumed the missing information of these authors were the same as the average value of the available information of the other authors.

simple average is one of the few feasible ways to measure a large amount of author data (millions of records of the papers by authors and their co-authors).

Within the variables measuring author information, the variables measuring author's number of (top and non-top) publications can be used to estimate the effect of author's publication record on the citation counts, author citations and author experience can be used to estimate the effect of author's academic status on the citation counts, the variables measuring author's affiliation score and country score can be used to estimate the effect of the strength of author's affiliation on the citation counts, and the variables measuring the information of author's co-authors can be used to estimate the effect of the strength of author's collaboration network on the citation counts.

Table 1.5: Variables measuring author information

Variable	**Description**
Num. Authors	Number of authors
Au: Cumulative Citations	Author citations
Au: Num. Pub.	Number of author's publications
Au: Num. Top Field	Number of author's top field publications
Au: Num. Top 5	Number of author's top 5 publications
Au: Num. Co-authored Pub.	Number of author's co-authored publications
Au: Experience	Author experience
Au: Affiliation Score	Economic research score of author's institution
Au: Country Score	Economic research score of the country of author's institution
Au: Num. Co-authors	Number of author's co-authors
Au: Num. Top Field Co-authors	Number of author's co-authors within top field publications
Au: Num. Top 5 Co-authors	Number of author's co-authors within the top 5 publications
Co-au: Cumulative Citations	Total citations of author's co-authors
Co-au: Num. Pub.	Total number of publications of author's co-authors
Co-au: Num. Top Field	Total number of top field publications of author's co-authors
Co-au: Num. Top 5	Total number of the top 5 publications of author's co-authors

Table 1.6 presents the statistics of some selected variables measuring author publication information and author's collaboration network information as of the year of publication. As shown in Table 1.6, noticeable differences in author information among the top 5 economics journals exist, and authors of papers in ECMA are most experienced with highest average number of top publications. In addition, papers

in AER and QJE have higher number of authors with higher average cumulative citations as of the year of publication. Besides, authors of papers in QJE have the strongest average collaboration network measured by the total number of top 5 publications and the total cumulative citations of papers written by co-authors of the author, as well as the highest author affiliation score.

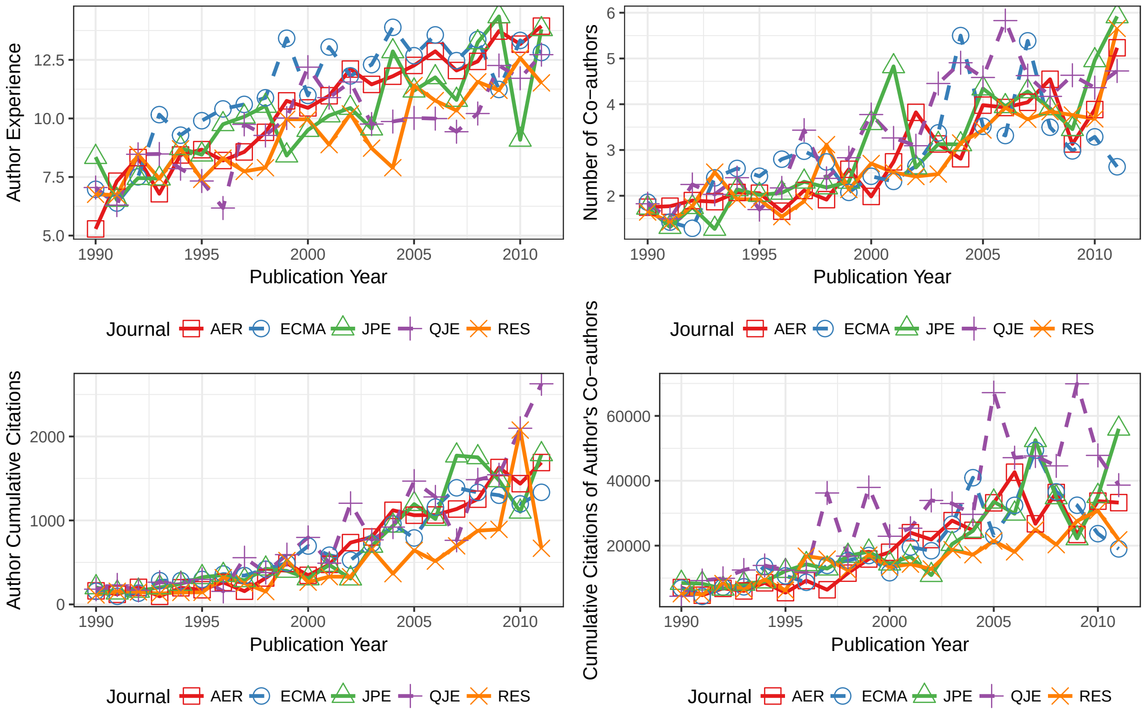
Figure 1.8 shows time series of some selected author information (author experience, author citations, number of co-authors, and total cumulative citations of papers written by co-authors of the author) as of the year of publication. It can be observed that the authors of papers published in the 2000s are more experienced, more highly cited, and have stronger collaboration network as of the year of publication than the authors of papers published in the 1990s.

Table 1.6: Author information at the year of publication

	Mean(Standard Deviation)				
	AER	**ECMA**	**JPE**	**QJE**	**RES**
Num. Authors	1.88	1.83	1.80	1.92	1.76
	(0.79)	(0.76)	(0.73)	(0.83)	(0.74)
Au: Cumulative Citations	782.61	657.19	607.62	776.33	465.58
	(1,628.63)	(1,195.37)	(1,323.77)	(1,606.56)	(1,129.23)
Au: Num. Pub.	22.38	23.95	19.12	21.10	18.85
	(21.22)	(21.42)	(18.75)	(19.50)	(16.88)
Au: Num. Top Field	8.94	10.88	8.71	8.59	8.93
	(8.37)	(9.16)	(7.97)	(7.83)	(7.83)
Au: Num. Top 5	3.75	4.43	4.08	4.35	3.85
	(4.00)	(3.97)	(4.13)	(4.24)	(3.97)
Au: Num. Co-authored Pub.	16.12	16.53	13.78	15.46	13.76
	(16.41)	(15.68)	(15.38)	(16.02)	(13.97)
Au: Experience	10.89	11.19	9.81	9.42	9.36
	(8.01)	(7.12)	(7.68)	(6.85)	(6.40)
Au: Affiliation Score	1,049.20	870.46	1,097.47	1,220.32	792.20
	(661.46)	(587.05)	(619.03)	(687.23)	(510.58)
Au: Country Score	38,838.96	33,489.86	40,661.58	38,433.97	29,406.68
	(15,119.70)	(18,296.01)	(13,959.04)	(15,238.80)	(19,622.98)
Au: Num. Co-authors	3.08	2.92	2.86	3.30	2.83
	(3.93)	(3.78)	(4.44)	(4.28)	(4.98)
Co-au: Cumulative Citations	21,850.72	20,454.21	19,007.95	29,308.25	15,790.74
	(41,497.14)	(35,130.16)	(34,946.07)	(63,785.91)	(24,554.40)
Co-au: Num. Pub.	610.43	620.15	570.55	800.12	523.52
	(1,163.69)	(1,038.50)	(876.80)	(1,960.60)	(890.14)
Co-au: Num. Top Field	144.83	196.19	151.30	176.22	162.39
	(164.25)	(208.04)	(192.93)	(202.22)	(175.20)
Co-au: Num. Top 5	63.33	77.48	69.18	88.97	70.05
	(88.59)	(101.44)	(107.33)	(116.90)	(104.21)

Note: "Au" is an abbreviation of "Author", and "Co-au" is an abbreviation of "Co-author". The measures of author information are constructed by the "No-duplicated" and "Cumulative" ways described in Appendix B.2.

Figure 1.8: Evolution of author profile at the year of publication



1.3 Estimation and Prediction Strategy

In this section, I present the strategy of estimating paper effect and author effect on the 10-year citations and citation paths, as well as the strategy of predicting paper's 10-year citations with the information available as of the year of publication.

1.3.1 Estimation strategy

After a paper i is published, it enters the academic market and is publicly viewed by consumers (researchers) in the market. Researchers observe paper and author information of paper i, and decide whether or not to cite paper i when writing a new paper. The information used by researchers cannot be observed, but the objectively measurable features of paper topic words ($\boldsymbol{W_i}$), presentation style ($\boldsymbol{P_i}$), research field

($\boldsymbol{F_i}$), journal information ($\boldsymbol{J_i}$), and author information ($\boldsymbol{A_{i,t}}$) arguably can be used as imperfect measures of the information that is used in the citing decision process.

If the new paper is subsequently published, paper i accumulates one more citation. The cumulative citations of paper i after t years of publication, $C_{i,t}$, can be viewed as an imperfect measure of the aggregation of researchers' decisions on citing paper i. In this chapter, I use $C_{i,10}$ and $C_{i,t}$ as the dependent variables to investigate paper effect and author effect on paper's citation counts.

To investigate the paper effect and author effect on paper's long-term citations, I estimate the effects of paper and author information as of the year of publication on paper's 10-year citations $C_{i,10}$. Equation 1.2 is used as the main regression equation.

$$C_{i,10} = \boldsymbol{W_i'\beta} + \boldsymbol{P_i'\gamma} + \boldsymbol{A_{i,0}'\eta} + \boldsymbol{F_i'\theta} + \boldsymbol{J_i'\rho} + \varepsilon_i \tag{1.2}$$

where $C_{i,10}$ is the 10-year citations of paper i, $\boldsymbol{W_i}$ is a vector of variables measuring topic words of paper i, $\boldsymbol{P_i}$ is a vector of variables measuring presentation style of paper i, $\boldsymbol{A_{i,0}}$ is a vector of variables measuring author information of paper i as of the year of publication (average value is used in papers having more than one author), $\boldsymbol{F_i}$ is a vector of variables measuring research field information of paper i, $\boldsymbol{J_i}$ is a vector of variables measuring journal information of paper i, and ε_i is the unobserved error term.

To investigate the drivers of paper citation paths, I use Equation 1.3 as the main regression equation. Intuitively, Equation 1.3 assumes the growth of paper's cumulative citations follows a quadratic function of t, and the parameters of the quadratic function are determined by paper and author information of the paper.

$$C_{i,t} = \alpha_0(\boldsymbol{A_{i,t}}) + \alpha_1(\boldsymbol{W_i}, \boldsymbol{P_i}, \boldsymbol{A_{i,t}}, \boldsymbol{F_i})t + \alpha_2(\boldsymbol{W_i}, \boldsymbol{P_i}, \boldsymbol{A_{i,t}}, \boldsymbol{F_i})t^2 + \mu_i + \varepsilon_{i,t} \tag{1.3}$$

where $C_{i,t}$ is the cumulative citations of paper i after t years of publication, t is the number of years after publication, $\alpha_0(\boldsymbol{A_{i,t}})$, $\alpha_1(\boldsymbol{W_i}, \boldsymbol{P_i}, \boldsymbol{A_{i,t}}, \boldsymbol{F_i})$, $\alpha_2(\boldsymbol{W_i}, \boldsymbol{P_i}, \boldsymbol{A_{i,t}}, \boldsymbol{F_i})$ are parameters of the quadratic function weighted by the features of papers and authors, $\boldsymbol{W_i}$ is a vector of variables measuring topic words of paper i, $\boldsymbol{P_i}$ is a vector of variables measuring presentation style of paper i, $\boldsymbol{A_{i,t}}$ is a vector of variables measuring author information of paper i after t years of publication (average value is used in papers having more than one author), $\boldsymbol{F_i}$ is a vector of variables measuring research field information of paper i, μ_i is the unobserved time-invariant individual effect for each paper, and $\varepsilon_{i,t}$ is the unobserved time-varying error term.

To simplify the specification of Equation 1.3, I assume $\alpha_0(\boldsymbol{A_{i,t}})$, $\alpha_1(\boldsymbol{W_i}, \boldsymbol{P_i}, \boldsymbol{A_{i,t}}, \boldsymbol{F_i})$, $\alpha_2(\boldsymbol{W_i}, \boldsymbol{P_i}, \boldsymbol{A_{i,t}}, \boldsymbol{F_i})$ are additive functions as shown in Equations 1.4-1.6.

$$\alpha_0(\boldsymbol{A_{i,t}}) = \boldsymbol{A'_{i,t}\eta_0} \tag{1.4}$$

$$\alpha_1(\boldsymbol{W_i}, \boldsymbol{P_i}, \boldsymbol{A_{i,t}}, \boldsymbol{F_i}) = \boldsymbol{W'_i\beta_1} + \boldsymbol{P'_i\gamma_1} + \boldsymbol{A'_{i,t}\eta_1} + \boldsymbol{F'_i\theta_1} \tag{1.5}$$

$$\alpha_2(\boldsymbol{W_i}, \boldsymbol{P_i}, \boldsymbol{A_{i,t}}, \boldsymbol{F_i}) = \boldsymbol{W'_i\beta_2} + \boldsymbol{P'_i\gamma_2} + \boldsymbol{A'_{i,t}\eta_2} + \boldsymbol{F'_i\theta_2} \tag{1.6}$$

Apart from estimating paper effect and author effect using low-dimensional measures of features of papers and authors, I also investigate the effect of the appearance of topic words and word pairs on paper's 10-year citations. The vectors of word and word pair dummies constructed by textual analysis are high dimensional ($p > n$), which makes Ordinary Least Squares inadequate to estimate the regression equation. To estimate the model, I use Post-Lasso method suggested by Belloni et al. [9]. The Post-Lasso method has two steps:

Step 1: Use Lasso to select variables by solving:

$$\hat{\boldsymbol{\beta}}(\lambda) = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{N}\sum_{i=1}^{N}(Y_i - \boldsymbol{X_i'}\boldsymbol{\beta})^2 + \lambda||\boldsymbol{\beta}||_1 \tag{1.7}$$

where $\lambda \geq 0$ is the penalty level and is chosen by the theoretically grounded plug-in method suggested by Belloni et al. [9], and $||\boldsymbol{\beta}||_1 = \sum_{j=1}^{p}|\beta_j|$.

Step 2: Use the variables selected by Lasso to estimate the model using Ordinary Least Squares.

1.3.2 Prediction strategy

In this subsection, I present the strategy for predicting paper's 10-year citations with the information available as of the year of publication. The dictionary-based textual analysis presented in Section 1.2 has constructed more than 6,000 variables measuring the features of papers and authors. Some of these variables might be powerful predictors of paper citations, while these high dimensional measures increase the risk of overfitting in traditional regression methods (e.g., Ordinary Least Squares). In addition, the large number of available variables makes it hard to construct interaction terms purely by intuition. For instance, it might make sense to construct interactions between some variables measuring paper information and some variables measuring author information, but including all pairwise interactions would produce more than 36,000,000 interaction terms, which would be infeasible in practice. Thus, I use machine learning methods that can handle high dimensional data and create the most predictive interactions between variables.

One possible way is to use regression shrinkage methods (e.g., Lasso, Post-Lasso, Ridge, Elastic Net). In this chapter, I use the objective function of Elastic Net proposed by Zou and Hastie [69] as the general form of objective functions of regression shrinkage methods, and compare the prediction performance of Lasso, Ridge, Elastic Net, and the Post-Lasso method proposed by Belloni et al. [9]. Let $\boldsymbol{X_i}$ denote paper i's vector of variables measuring paper and author information with dimension p, and Y_i denote the 10-year citations. The objective function of Elastic Net for the Gaussian family is shown in Equation 1.8. It sets $\alpha \in (0,1)$, which is a compromise between Lasso ($\alpha = 1$) and Ridge ($\alpha = 0$).

$$\hat{\boldsymbol{\beta}}(\alpha, \lambda) = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{2N} \sum_{i=1}^{N} (Y_i - \boldsymbol{X_i'\beta})^2 + \lambda[\alpha ||\boldsymbol{\beta}||_1 + (1-\alpha)||\boldsymbol{\beta}||_2^2/2] \tag{1.8}$$

where $\alpha \in (0,1)$, $\lambda \geq 0$ is the penalty level, $||\boldsymbol{\beta}||_1 = \sum_{j=1}^{p} |\beta_j|$ and $||\boldsymbol{\beta}||_2^2 = \sum_{j=1}^{p} \beta_j^2$.

The penalty level λ can be chosen by different algorithms. In Section 1.5, I compare the results of two algorithms: the cross-validation algorithm provided by Zou and Hastie [69] and the theoretically grounded plug-in method suggested by Belloni et al. [9].

Regression shrinkage methods can reduce the number of covariates, and detect the most powerful predictors. However, the parameters in regression shrinkage methods are fitted jointly, which makes the computational burden of fitting non-linear models with many higher-order interactions between measures of papers and authors fairly high. Thus, other machine learning methods might be more suitable for adding higher-order interactions into the prediction model.

One way to add higher-order interactions into the prediction model is to use Neural Network. The Neural Network model is normally formed by one input layer, one output layer, and one or several hidden layers. Predictors enter the model via input layer, and they are weighted, combined and transformed by activation functions of each unit in the first hidden layer. Then, the outputs of units in the first hidden layer are weighted, combined and transformed by activation functions of units in the second hidden layer. This process continues until the output layer is reached, and the outputs of units in the last hidden layer are weighted and combined in the output layer to form outputs of the Neural Network. The interactions between variables are created by the combination of outputs of units in hidden layers, and the nonlinearity is added by nonlinear activation functions.

Let M denote the number of units in a representative hidden layer. The combination and transformation of inputs by activation functions of units in a hidden layer can be represented by Equation 1.9.

$$z_{i,m} = \sigma(\nu_{i,m}) = \sigma(\boldsymbol{X_i'\alpha_m}) \tag{1.9}$$

where $z_{i,m} = \sigma(\nu_{i,m})$ is the activation function, $\boldsymbol{\alpha_m} = \alpha_{1,m}, \alpha_{2,m}, ..., \alpha_{p,m}$, and $m = 1, ..., M$.

In this chapter, I use Rectified Linear Unit (ReLU) as the activation function because it is less likely to cause vanishing gradient problem and is less computationally expensive than other activation functions (e.g., Sigmoid). The equation of ReLU is given by Equation 1.10.

$$\sigma(\nu_{i,m}) = \begin{cases} 0 & \text{for} \quad \nu_{i,m} < 0 \\ \nu_{i,m} & \text{for} \quad \nu_{i,m} \geq 0 \end{cases} \tag{1.10}$$

In the output layer, the outputs $(z_{i,1}, ..., z_{i,M})$ of units in the last hidden layer are summed by Equation 1.11 to form output $f(\boldsymbol{X_i})$.

$$f(\boldsymbol{X_i}) = \sum_{m=1}^{M} z_{i,m}\beta_m \tag{1.11}$$

To train the Neural Network, I use backpropagation to compute the gradient descent of the loss function (Equation 1.12), and the unknown parameters $\boldsymbol{\alpha_1}, ..., \boldsymbol{\alpha_M}$ and $\beta_1, ..., \beta_M$ are updated in gradient direction. The update of parameters at the $(r+1)$th iteration is given by Equation 1.13 and 1.14.

$$R = \sum_{i=1}^{N} R_i = \sum_{i=1}^{N} (Y_i - f(\boldsymbol{X_i}))^2 \tag{1.12}$$

$$\alpha_{p,m}^{(r+1)} = \alpha_{p,m}^{(r)} - \gamma_r \sum_{i=1}^{N} \frac{\partial R_i}{\partial \alpha_{p,m}^{(r)}} \tag{1.13}$$

$$\beta_{m}^{(r+1)} = \beta_{m}^{(r)} - \gamma_r \sum_{i=1}^{N} \frac{\partial R_i}{\partial \beta_{m}^{(r)}} \tag{1.14}$$

where γ_r is the learning rate.

To speed up the training process and save memory space, I set the batch size of the network to be a small number, which means only a small subset of samples are propagated through the Neural Network in one forward/backward pass. In addition, I use the dropout method [62] to randomly drop units from the neural network during the model fitting process to lower the risk of overfitting, and test different numbers of epochs to find a suitable number of epochs for training the Neural Network. In Section 1.5, I present the prediction result of an Artificial Neural Network model with two hidden layers built on Tensorflow system [1].

Another way to add higher-order interactions into the prediction model is to use regression tree method. The regression tree method partitions samples $\boldsymbol{X_1}, ..., \boldsymbol{X_N}$ into K regions (terminal nodes), and form the prediction rule for $Y_1, ..., Y_N$ as:

$$f(\boldsymbol{X_i}) = \sum_{k=1}^{K} c_k I(\boldsymbol{X_i} \in S_k) \tag{1.15}$$

where c_k is the prediction for the kth region.

The partition of the space is determined by minimizing the loss function (Equation 1.16), and the algorithm given by Equation 1.17 is used when determining a split point s for a given splitting variable j.

$$R = \sum_{k=1}^{K} \sum_{\boldsymbol{X_i} \in S_k} (Y_i - f(\boldsymbol{X_i}))^2 \tag{1.16}$$

$$\min_{j,s}[\min_{c_1} \sum_{\boldsymbol{X_i} \in \{\omega | \omega_j \leq s\}} (Y_i - c_1)^2 + \min_{c_2} \sum_{\boldsymbol{X_i} \in \{\omega | \omega_j > s\}} (Y_i - c_2)^2] \tag{1.17}$$

where S_k is the set that defines terminal node k.

The complexity of the regression tree model increases with the number of terminal nodes, and the splitting terminates when some predetermined complexity criteria are achieved. The higher-order interactions are added into the regression tree in the process of partitioning samples $\boldsymbol{X_1}, ..., \boldsymbol{X_N}$ into K regions. Considering that a simple regression tree may underfit, while a very complicated regression tree may overfit, I choose Random Forest [12] and Gradient Boosted Trees [25] to fit a number of regression trees, rather than a single tree.

The Random Forest method fits a number of regression trees using $1, ..., B$ random samples with replacement. After the regression trees are fitted, the outputs of the regression trees are bagged using Equation 1.18 to form output $f(\boldsymbol{X_i})$.

$$f(\boldsymbol{X_i}) = \frac{1}{B} \sum_{b=1}^{B} f_b(\boldsymbol{X_i}) \tag{1.18}$$

The Gradient Boosted Trees method fits the parameters in a stage-wise fashion. It starts with initializing $\hat{f}^0(\boldsymbol{X_i}) = 0$ and setting the number of trees as T. Then, for $t = 1, ..., T$, the negative gradient of the loss function ($R = \sum_{i=1}^{N} R_i = \sum_{i=1}^{N}(Y_i - f(\boldsymbol{X_i}))^2$)

is calculated as $-\frac{\partial R_i}{\partial f(\boldsymbol{X_i})}$, and regression tree t is used to fit the negative gradient computed from regression tree $t-1$. In each tree model, the optimal terminal node predictions, $\rho_1^{(t)}, ..., \rho_K^{(t)}$ are computed by Equation 1.19, and they are used by Equation 1.20 to update $\hat{f}^{(t)}(\boldsymbol{X_i})$.

$$\rho_k^{(t)} = \arg\min_{\rho_k^{(t)}} \sum_{\boldsymbol{X_i} \in S_k} (Y_i - \hat{f}^{(t-1)}(\boldsymbol{X_i}) - \rho_k^{(t)})^2 \tag{1.19}$$

$$\hat{f}^{(t)}(\boldsymbol{X_i}) = \hat{f}^{(t-1)}(\boldsymbol{X_i}) + \lambda \sum_{k=1}^{K} \rho_k^{(t)} I(\boldsymbol{X_i} \in S_k) \tag{1.20}$$

where λ is learning rate.

Since Random Forest and Gradient Boosted Trees could not compare the predictive power of each variable before building regression trees, their prediction performance might be negatively affected by variables that were not predictive. Thus, if some methods that could select the variables to be used in tree models were embedded, the prediction ability of Random Forest and Gradient Boosted Trees might be improved.

Based on my evaluation of the advantages and disadvantages of these machine learning methods, I develop a hybrid method that combines dictionary-based textual analysis, regression shrinkage, and Gradient Boosted Trees, in order to combine the advantages and partially overcome the shortcomings of each method. The hybrid method uses dictionary-based textual analysis to construct variables measuring features of papers and authors, uses regression shrinkage to select variables that are powerful predictors of paper citations, and uses Gradient Boosted Trees to fit nonlinear prediction model with higher-order interactions. The steps are:

Step 1: Variable Construction

Use dictionary-based textual analysis to measure unstructured data to construct high dimensional vectors of variables measuring paper and author information.

Step 2: Variable Selection

Use regression shrinkage methods to fit a variety of linear models to predict 10-year citations of each paper, and select the model that gives the smallest out-of-sample Mean Squared Error (MSE). Then, use the variables with non-zero coefficients as predictors in Step 3.

Step 3: Model Fitting

Add the predictors selected in Step 2 sequentially into a variety of Gradient Boosted Trees based on the absolute value of the estimated coefficients of the predictors, and select the Gradient Boosted Trees model that gives the smallest out-of-sample MSE.

In Section 1.4, I present the results of estimating paper effect and author effect on paper citations. In Section 1.5, I compare the proposed hybrid method with other machine learning methods in terms of MSE and out-of-sample fit.

1.4 Estimation Results and Discussion

1.4.1 Effects on paper's 10-year citations

Paper effect and author effect

Table 1.7 presents the estimated coefficients of the variables measuring paper and author information as of the year of publication on paper's 10-year citations. Papers in AER 1990 are used as the baseline. Due to inevitable measurement error and multicollinearity, individual coefficients should be interpreted with caution.

Compared with papers in the other journals, papers in QJE get higher citations even after controlling a variety of paper and author information, though the QJE effect decreases after more control variables are added. The coefficient of QJE in column (5) suggests that papers in QJE are predicted to get 2.33 more citation counts after 10 years of publication, controlling paper and author information. One potential explanation might be that QJE performed better in advertising publications. It is also possible that editors of QJE preferred papers that would be highly cited, while editors of the other journals did not have strong preferences for highly cited papers. In addition, the differences in the pools of submitted manuscripts could be another cause of the QJE effect.

The estimation results confirm the importance of paper research field in determining paper citations, and papers with higher "Macro, Monetary Econ" intensity generally get more citations, while papers with higher "Micro", "Public Econ", "Labor Econ", and "Agricultural, Environmental Econ" intensity get fewer citations. Specifically, the coefficients in column (5) suggest that 0.25% higher "Macro, Monetary Econ" intensity leads to 2 more 10-year citation counts, while 0.25% higher "Micro" intensity leads to 1 less 10-year citation counts. Paper topic and paper presentation style also determine paper's citations, and papers that cover more popular topics and have more pages generally have more citations, while papers with higher complexity get fewer citations. One potential explanation might be that a paper covering more popular topics is related to a broader range of subsequent studies that may cite it, while a paper with higher complexity takes longer for other researchers to read and cite it in subsequently published papers.

Within the variables measuring author information, the number of authors and number of author's co-authored papers are positively correlated with citations, while the coefficients of author experience and number of author's publications are not statistically significant. In addition, the coefficients of some variables measuring the strength of author's collaboration network are also significant. Specifically, higher 10-year citation counts are associated with higher total citations and numbers of the top 5 publications of authors' co-authors, while associated with lower numbers of publications and numbers of top field publications of authors' co-authors. The 10-year citation counts are negatively correlated with numbers of authors' top field publications, and its coefficient in column (5) suggests that papers by authors with 10 more top field publications are predicted to have 2.3 less 10-year citation counts. One possible explanation might be that for an author with more top field publications, the competition on citations between her/his publications in the same field might be fiercer, which reduce average citations of her/his papers. Another possible explanation might be that higher number of top field publications sends a positive signal on the quality of the author's submission to the editor, though the academic contribution of the submission could be low. In Card and DellaVigna [14], the positive effect of the number of author's top publications on paper citation counts is reported. However, the way of measuring the number of author's top publications and some other paper/author information in this chapter is different from their paper.

The adjusted R-squared in all of these regressions is very low. The regression in column (5) gives the largest adjusted R-squared of 0.17, meaning only 17% of the variation of the 10-year citations can be explained by the variables measuring paper and author information as of the year of publication in linear regression. The low adjusted R-squared may be caused by the outliers in the sample. In addition, it may indicate that some potentially important determinants of paper citations may be missing from the model and a simple linear regression model is inadequate to model the variation of paper citation counts. For example, the topic word dummies that could be important determinants of the citation counts are not included in the regression models. To better model the variation of the 10-year citations and predict out-of-sample, I use a quadratic function with a panel of paper information and yearly author information in Section 1.4.2, and use machine learning methods in Section 1.5.

Table 1.7: Effects on 10-year citation counts part I

	Dependent variable:				
	10-year Citations: $C_{i,10}$				
	(1)	(2)	(3)	(4)	(5)
Journal: ECMA	−0.45	0.45	1.01*	0.85	0.96*
	(0.54)	(0.56)	(0.57)	(0.57)	(0.58)
Journal: JPE	0.48	0.43	0.69	0.71	0.78
	(0.51)	(0.50)	(0.52)	(0.52)	(0.53)
Journal: QJE	2.86***	2.61***	2.78***	2.35***	2.33***
	(0.74)	(0.74)	(0.76)	(0.70)	(0.71)
Journal: RES	−0.39	−0.03	0.68	0.96*	1.09*
	(0.54)	(0.54)	(0.57)	(0.57)	(0.58)
Math, Quant Methods	−98.50	61.42	67.39	71.91	86.18
	(102.07)	(102.63)	(148.23)	(156.85)	(152.23)
Micro	−658.43***	−475.74***	−433.87***	−395.67***	−379.44***
	(115.88)	(114.31)	(149.02)	(137.95)	(136.01)
Macro, Monetary Econ	970.38***	837.38***	862.01***	879.23***	858.26***
	(285.27)	(282.25)	(277.93)	(278.75)	(270.71)
International Econ	551.82	612.52*	635.64**	623.45**	583.40**
	(342.37)	(338.80)	(322.63)	(299.39)	(288.01)
Financial Econ	124.38	−136.21	−91.40	−154.95	−179.84
	(266.05)	(267.58)	(260.59)	(257.63)	(252.09)
Public Econ	−468.10***	−544.95***	−529.67***	−501.88***	−503.29***
	(170.94)	(173.13)	(170.05)	(171.21)	(171.95)
Health, Education, Welfare	228.94	388.29	412.09*	348.28	334.79
	(248.22)	(241.86)	(243.46)	(248.04)	(242.35)
Labor Econ	−504.53***	−748.23***	−721.64***	−654.62***	−679.95***
	(141.12)	(141.89)	(140.37)	(139.97)	(141.60)
Law, Econ	−348.54	−908.99	−798.30	−719.92	−863.88
	(558.31)	(560.84)	(583.28)	(598.52)	(582.51)
Industrial Organization	61.58	0.06	−12.97	64.99	69.74
	(150.35)	(146.19)	(150.98)	(147.57)	(146.87)
Business Econ, Marketing	264.53	527.52	526.65	480.23	508.34
	(348.85)	(343.96)	(350.09)	(340.89)	(338.02)
Econ Development, Growth	24.87	6.09	−56.21	−17.75	−18.34
	(244.67)	(225.80)	(243.47)	(227.10)	(225.72)
Econ Systems	275.83	58.63	119.59	105.86	96.84
	(371.29)	(358.93)	(354.46)	(356.68)	(355.04)
Agricultural, Environmental Econ	−726.64***	−1,056.32***	−992.34***	−924.22***	−903.46***
	(243.76)	(262.84)	(252.54)	(238.27)	(233.42)
Urban Econ	−167.87	−290.65	−252.50	−138.49	−84.04
	(289.82)	(296.84)	(283.84)	(277.48)	(266.33)

Effects on 10-year citation counts part II

	Dependent variable:				
	10-year Citations: $C_{i,10}$				
	(1)	(2)	(3)	(4)	(5)
Popular Topic Coverage		18.00***	13.40***	8.50**	7.89**
		(2.84)	(3.09)	(3.46)	(3.50)
Presentation: Descriptiveness			593.37	377.64	376.83
			(596.44)	(591.52)	(596.81)
Presentation: Vocabulary Richness			−10.58	−37.53	−31.45
			(55.73)	(53.49)	(54.02)
Presentation: Complexity			−0.08*	−0.08*	−0.10**
			(0.04)	(0.04)	(0.05)
Num. Pages			0.06**	0.05**	0.05**
			(0.02)	(0.02)	(0.02)
Num. Authors				0.95***	0.93***
				(0.29)	(0.28)
Au: Cumulative Citations(×1000)				1.46**	0.71
				(0.62)	(0.65)
Au: Num. Pub.				0.02	0.03
				(0.05)	(0.05)
Au: Num. Top Field				−0.27**	−0.23*
				(0.13)	(0.13)
Au: Num. Top 5				0.25*	0.10
				(0.15)	(0.15)
Au: Num. Co-authored Pub.				0.11**	0.14**
				(0.05)	(0.06)
Au: Experience				−0.04	−0.03
				(0.04)	(0.04)
Au: Num. Co-authors					−0.11
					(0.08)
Co-au: Cumulative Citations(×1000)					0.06***
					(0.02)
Co-au: Num. Pub.					−0.002***
					(0.0005)
Co-au: Num. Top Field					−0.01***
					(0.002)
Co-au: Num. Top 5					0.01***
					(0.004)
Control: Pub. Year	Yes	Yes	Yes	Yes	Yes
Observations	3,472	3,472	3,472	3,472	3,472
R^2	0.12	0.13	0.14	0.17	0.18
Adjusted R^2	0.11	0.12	0.13	0.16	0.17

Note: *p<0.1; **p<0.05; ***p<0.01. Papers in AER at the year of 1990 are used as the baseline. "Au" is an abbreviation of "Author", and "Co-au" is an abbreviation of "Co-author". The measures of paper information are constructed by parsing papers' full texts. The measures of author information are constructed by "No-duplicated" and "Cumulative" ways described in Appendix B.2. Robust standard errors in parentheses.

Effect of topic word appearance

Table 1.7 has shown that some of the research field measures are significantly correlated with papers' 10-year citation counts, and the coefficient of popular topic coverage is positive in all of these regressions. To investigate the effect of each popular topic word on papers' 10-year citation counts, I estimate the effect of the appearance of each popular topic word and word pair on papers' 10-year citation counts. Since adding topic word dummies, word pair dummies, and control variables to Equation 1.2 yields more than 6,000 variables, I use Post-Lasso method to estimate the regression equation.

Table 1.8 presents the top topic words ranked by the absolute values of their estimated coefficients. It can be seen that some topic words and topic word pairs related to macroeconomics (e.g., "gdp", ("capit","share"), ("product","develop","growth"), and ("distribut","incom")), quantitative methods (e.g., "correl"), financial economics (e.g., "bank"), and educational economics (e.g., "school") have the largest positive coefficients, and some topic words and topic word pairs related to microeconomics (e.g., ("ration","inform")), law and economics (e.g., ("prison","sentenc")), and transportation economics (e.g., ("vehicl","drive","port")) have the largest negative coefficients. However, some words with similar meaning have opposite estimated coefficients (e.g., "correl" and "regress"), and some topic words can represent multiple research topics (e.g., "optim" is possible to be stripped from "optimal contract", "optimal taxation", or any other words that contain "optim"). The standard errors of the coefficents are not estimated because estimating the distribution of post-model-selection estimators is not trivial (for discussion on this issue, see Leeb and Pötscher [46, 47]).

Table 1.8: Effect of word appearance on 10-year citation counts

Top 10 words with positive coefficients		Top 10 words with negative coefficients	
Topic Word	**Coefficient**	**Topic Word**	**Coefficient**
"gdp" (Full Text)	23.07	"optim" (First 200 Sent.)	-12.88
"correl" (First 100 Sent.)	12.76	("consum","consumpt") (First 100 Sent.)	-7.45
"road" (Full Text)	9.87	("ration","inform") (First 100 Sent.)	-5.35
"bank" (Full Text)	9.12	"compat" (First 100 Sent.)	-5.26
("capit","share") (Full Text)	8.64	("prison","sentenc") (First 10 Sent.)	-3.57
"school" (Full Text)	8.60	("competit","price") (First 10 Sent.)	-3.32
("product","develop","growth") (Full Text)	7.73	("vehicl","drive","port") (First 200 Sent.)	-3.03
("distribut","incom") (Full Text)	5.83	"regress" (Full Text)	-2.93
"intern" (Full Text)	5.37	("popul","distribut","aid") (First 10 Sent.)	-2.36
("revenu","taxat","taxsystem") (First 10 Sent.)	2.03	"growth" (Full Text)	-1.93

Note: The topic words are standardized by Porter Stemmer. For instance, the word "compat", which is a standardized keyword in JEL code L (Industrial Organization), can be "compatibility" or other topic words having root "compat". "Sent." is an abbreviation of "Sentences". The "First 10 Sent.", "First 100 Sent.", "First 200 Sent.", and "Full Text" in brackets show the place of the appearance of the topic words and topic word pairs. Variables measuring paper's research field information, popular topic coverage, presentation style, author information, and journal information are used as control variables. The regression equation is estimated by Post-Lasso, and 68 of 6,234 variables are selected.

1.4.2 Effects on paper citation path

In this subsection, I present the results of estimating paper effect and author effect on paper citation path with a panel of paper information and yearly author information. After checking the regression results of the quadratic function (Equation 1.3), I find the coefficient of the quadratic term is quite small in most of the regressions, and the linear term is the main determinant of the slope of paper citation path. Thus, I focus on discussing the coefficient of the linear term in this subsection.

Paper effect

Table C.4 presents the estimated coefficients of paper's research field on its citation path. The results suggest that a steeper slope of paper citation path is associated with higher "Math, Quant Methods", "Econ Development, Growth", and "Econ Systems" intensity, while associated with lower "Macro, Monetary Econ", "Public Econ", "Labor Econ", and "Industrial Organization" intensity.

Table C.5 presents the coefficients of paper topic and presentation style on its citation path. It shows that a steeper slope of paper citation path is associated with higher popular topic coverage and number of pages in all of these regressions, though the coefficients decrease after the variables measuring author information are controlled for. However, the coefficients of complexity, descriptiveness, and vocabulary richness of paper's presentation style are not statistically significant in most of these regressions.

Author effect

Table C.6 presents the coefficients of variables measuring author information on paper citation paths. The results show that a steeper slope of paper citation path is associated with higher number of authors and author citations, while associated with lower number of author's top field publications, number of author's top 5 publications, number of author's co-authored publications, and number of publications written by co-authors of the author.

These results indicate that the papers written by a bigger team of highly cited authors generally have a steeper slope of citation path. The explanations to the negative effect of the number of author's top field publications on the 10-year citation counts could help explain the negative coefficients of the number of author's top publications. Another explanation might be that for an author with more subsequent top publications, the theory or method proposed in the original paper is extended in the subsequent top publications, which leads researchers to cite the author's subsequent top publications, instead of the original paper.

Heterogeneity in effects on paper citation path

Table C.7 compares the estimation results of paper effect and author effect on paper citation path between the top 5 economics journals. It shows that papers in RES are the most negatively affected by higher "Macro, Monetary Econ" intensity, and QJE has extremely large positive coefficients of "Econ Development, Growth" and "Econ Systems" intensity. Regarding paper presentation style, AER and QJE are negatively affected by higher paper complexity, while coefficients of the other three journals are not statistically significant. The coefficients of variables measuring author information also show heterogeneity among different journals. Notably, papers in JPE benefit the most from bigger teams of highly cited authors.

There is heterogeneity among journals in the adjusted R-squared. The regression for JPE gives the largest adjusted R-squared of 0.54, while the regression for RES gives the smallest adjusted R-squared of 0.16, meaning that a much smaller portion of the variation of RES papers' citations can be explained by the variables in the fixed effect model. One potential explanation for the heterogeneity in the adjusted R-squared among journals could be that some important determinants of RES papers' citations were not captured. For example, it is possible that the topic word dictionary could only capture a small share of the topic words that are important determinants of the citation counts of papers in RES.

Table C.8 compares the estimation results of paper effect and author effect on paper citation path between different author groups. The authors were grouped according to the ranking of author's most recent institution, and the authors without a matchable institution are excluded. Since more than half of the observations did not have matched institution information, the regression results in Table C.8 should be interpreted with caution. Table C.8 shows that the coefficients seem to have no clear trend across author groups, and most of them are insignificant.

The estimation results could help deepen our understanding of the drivers of paper citations, while the low R-squared shows that a simple regression model might be inadequate to model the variation of paper citation counts and predict out-of-sample. In the next section, I present the results of using machine learning methods to predict papers' 10-year citations with the information available as of the year of publication.

1.5 Prediction Results and Discussion

1.5.1 Paper citation out-of-sample prediction

In this subsection, I test the ability of machine learning methods to predict papers' 10-year citation deciles with the information available as of the year of publication. I use the 10-year citation deciles instead of the 10-year citation counts as the variable to be predicted, in order to reduce the influence of the outliers. Even though predicting citation deciles reduces the difficulty of prediction, the low adjusted R-squared (less than 0.5) in Table C.9 shows it is still hard for OLS models to explain the variation of paper citations. In addition, as will be shown in this section, even the prediction model that gives the smallest Mean Squared Error (MSE) cannot predict the citation counts of the highest and lowest deciles well.

The 3,472 papers that have 10-year citation data are used in this test. I sort papers into 10 parts based on the deciles of papers' 10-year citations and create variable $D_{i,10}$, the 10-year citation decile of paper i, which has values ranges from 1 to 10. Then, I randomly select 70% of the papers as the training sample and 30% of the papers as the testing sample and use prediction models to predict papers' $D_{i,10}$ using paper and author information available as of the year of publication. To compare models' out-of-sample prediction performance, I use Mean Squared Error (MSE):

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(D_{i,10} - \hat{D}_{i,10})^2 \tag{1.21}$$

where $D_{i,10}$ is the observed 10-year citation decile of paper i, and $\hat{D}_{i,10}$ is the predicted 10-year citation decile of paper i.

The vectors of variables measuring paper and author information are added sequentially into prediction models. The vectors of variables in prediction models are shown in Table C.10. Model (1) only includes variables measuring journal information, Model (2) adds variables measuring field information, Model (3) adds variables measuring popular topic coverage, Model (4) adds variables measuring presentation style, Model (5) adds variables measuring author information at the publication year, Model (6) adds lagged author variables, and Model (7) to Model (9) add high dimensional measures of paper topics.

The linear model fitted by Ordinary Least Squares is used as the baseline model, and regression shrinkage methods (Post-Lasso, Lasso, Ridge, Elastic Net), Neural Network, Random Forest, Gradient Boosted Trees, and the hybrid methods developed in Section 1.3.2 are compared with each other. The implementation and the values of the key parameters of machine learning methods are presented in Table C.11. The parameters of these machine learning methods are determined after testing many different combinations of parameters.

Table 1.9 presents the out-of-sample prediction results. It shows that the MSE of prediction methods, except for OLS, generally decreases after adding more predictors (from Model (1) to Model (9)). The results from regression shrinkage methods, Random Forest, and Gradient Boosted Trees show that the MSE decreases by more than 10% after adding variables measuring research fields (from Model (1) to Model

(2)) and decreases by another roughly 10% after adding variables measuring author information (from Model (4) to Model (5)). However, after adding the high dimensional vector of topic word dummies (from Model (6) to Model (7)), some machine learning methods begin to overfit, which causes their out-of-sample prediction performance to deteriorate.

The Neural Network has lower MSE than OLS, but it is not as good as the other machine learning methods in this prediction test. Instead, the result shows that it suffers from overfitting issue when many predictors are added, and the MSE increases from 7.04 in Model (7) to 13.08 in Model (9). One possible reason might be that Neural Network has a large number of parameters, and the sample size in this chapter is not big enough for Neural Network to show its merit. It is also possible that the prediction ability of Neural Network can be improved by more complex network structure and other combinations of parameters.

The prediction performance of the proposed Shrinkage-Gradient Boosted Trees Hybrid method is much better than OLS and marginally better than regression shrinkage methods, Random Forest, and Gradient Boosted Trees. In addition, the Shrinkage-Gradient Boosted Trees Hybrid Model (6), which gives the smallest MSE, only uses 221 predictors. Whereas, the Gradient Boosted Trees Model (7), which gives the second smallest MSE, uses 1,836 predictors. This property of the Shrinkage-Gradient Boosted Trees Hybrid method significantly reduces the cost of data collection and computation for using it to predict 10-year citation deciles of new papers.

The row "SGBT Hybrid (No JID)" in Table 1.9 reports the results of Shrinkage-Gradient Boosted Trees Hybrid model without using journal ID variable. It shows that in prediction models with many predictors, adding journal ID variable only marginally improves the prediction performance.

Table 1.9: Out-of-sample Mean Squared Error

	Mean Squared Error								
Method	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**
Ordinary Least Squares (OLS)	9.23	10.48	10.87	10.92	11.36	11.60	14.80	15.84	15.84
Plug-in Post-Lasso	6.98	6.12	6.14	6.09	5.39	5.39	5.43	5.33	5.38
Plug-in Lasso	6.98	6.07	5.91	5.87	5.32	5.32	5.17	5.19	5.20
Cross-Validation Lasso	6.84	6.08	5.82	5.74	5.23	5.22	5.27	5.16	5.22
Cross-Validation Ridge	6.85	6.03	5.77	5.77	5.22	5.23	5.23	5.31	5.30
Cross-Validation Elastic Net	6.82	5.96	5.79	5.80	5.24	5.22	5.27	5.18	5.15
Neural Network	7.27	12.30	11.29	11.23	9.88	19.19	7.04	7.64	13.08
Random Forest	6.83	5.81	5.68	5.65	5.03	5.08	5.39	5.51	5.58
Gradient Boosted Trees	6.60	5.58	5.46	5.45	4.81	4.83	4.80	4.81	4.81
Shrinkage-Random Forest Hybrid	6.83	4.91	4.91	4.91	4.92	4.90	4.90		
Shrinkage-Gradient Boosted Trees Hybrid	6.60	4.82	4.83	4.83	4.81	**4.79**	4.80		
SGBT Hybrid (No JID)	7.09	4.86	4.84	4.84	4.83	4.81	4.83		

Note: "SGBT Hybrid (No JID)" represents Shrinkage-Gradient Boosted Trees Hybrid model without using journal ID variable. The 243 predictors selected by Cross-Validation Elastic Net model (9) are added sequentially into hybrid models following the rank of the absolute value of coefficients. The hybrid model (1)-(6) have the same numbers of predictors as the model (1)-(6) of the other prediction models. The hybrid model (7) uses all the 243 predictors. The hybrid model (6), which gives the smallest MSE, uses 221 predictors.

The top 50 predictive variables selected by the regression shrinkage model with the smallest MSE (Cross-Validation Elastic Net Model (9)) are shown in Table 1.10. The result shows that the "QJE dummy" predicts higher 10-year citations, while the "RES dummy" predicts lower 10-year citations. However, since there are more than 200 variables selected by the regression shrinkage model, the importance of journal dummies is much decreased. The top predictive variables measuring paper information are "Popular Topic Coverage in the Full Text" and "Number of Pages", and their positive coefficients indicate that longer papers with higher coverage of popular topics are predicted to have higher 10-year citations. The top predictive variables measuring author information are "Total Number of the Top 5 Publications of Authors' Co-authors", "Author Citations", and "Number of Co-authors within the Top 5 Publications". These variables measure author citations and the strength of author's collaboration network, and their positive coefficients show that papers written by highly cited authors with stronger collaboration network are predicted to have higher 10-year citations.

A big portion of the variables in Table 1.10 are topic words and word pairs. The coefficients of these variables confirm that the appearance of some research topics are powerful predictors of papers' 10-year citations. In addition, some variables measuring paper research fields are powerful predictors of 10-year citations, and "Micro in Full Text" negatively predicts 10-year citations, while "Math, Quant Methods in the First 200 Sentences" positively predicts 10-year citations.

Figure 1.9 shows the predicted citation distribution given by the preferred hybrid method and actual citation distribution of the papers in the testing set. As shown in Figure 1.9, the preferred hybrid method can partially capture the skewness of the citation distribution of the top 5 economics journals. However, it predicts a distribution concentrated around its mean value, while cannot predict the citation counts of the highest and lowest deciles well. One potential explanation could be that some important features of papers in the highest and lowest deciles are not well captured by the variables used in the analysis.

Figure 1.10 shows the correct rate of decile prediction. In this prediction, I use the preferred hybrid method to predict the exact decile of the citations of each paper. The result shows that the citation deciles of the papers in the middle of the citation distribution are easier to be predicted than that of the papers in the tails. For the papers in the 4th to 7th deciles, the preferred method can predict their deciles correctly by more than 20% of the chance. However, the correct rate drops sharply when the method is used to predict papers in the tails. Moreover, the preferred hybrid method cannot give the correct prediction for papers in the 1st and 10th deciles of the citation distribution.

Table 1.10: The top 50 predictors selected by Elastic Net Model (9)

Variable	**Coefficient**	**Variable**	**Coefficient**
Journal: RES	-0.531	("factor product","product","develop") (Full Text)	0.059
Popular Topic Coverage (Full Text)	0.227	"school" (First 10. Sent.)	0.059
Num. Pages	0.170	"develop" (Full Text)	0.059
Co-authors: Num. Top 5 Papers (Dup Cum)	0.159	"equiti" (First 200. Sent.)	0.058
Author: Cumulative Citations	0.128	"litig" (Full Text)	-0.057
Author: Num. Top 5 Co-authors (Dup)	0.123	"bank" (Full Text)	0.056
"Micro" (Full Text)	-0.122	("product","econom growth","develop") (Full Text)	0.055
"cluster" (Full Text)	0.114	"deriv" (First 10. Sent.)	-0.054
Num. Words (Full Text)	0.106	"educ attain" (Full Text)	0.054
"Math, Quant Methods" (First 200. Sent.)	0.104	("budget","receiv") (First 200. Sent.)	-0.053
Co-authors: Cumulative Citations (Nodup Cum)	0.104	Journal: QJE	0.052
("capit","share") (First 200. Sent.)	0.087	"correl" (First 200. Sent.)	0.052
Co-authors: Cumulative Citations (Nodup Point)	0.081	"intern" (First 100. Sent.)	0.050
("author","motiv") (Full Text)	0.081	"polic" (Full Text)	0.048
("consum","expenditur") (First 100. Sent.)	-0.079	("rule","associ") (First 100. Sent.)	-0.048
Num. Authors	0.079	("econometr model","model") (First 100. Sent.)	-0.048
("endogen","estim") (Full Text)	0.077	"build" (First 200. Sent.)	0.047
"foundat" (First 100. Sent.)	0.077	"gdp" (First 200. Sent.)	0.046
"Math, Quant Methods" (First 100. Sent.)	0.075	("statist","indic") (First 10. Sent.)	-0.046
"liquid" (Full Text)	0.074	"administr" (First 200. Sent.)	-0.045
"psycholog" (Full Text)	0.069	("method","experi") (First 200. Sent.)	0.043
("product","organ") (Full Text)	0.062	("point estim","estim","coeffici") (Full Text)	0.042
"optim" (First 100. Sent.)	-0.062	("labor demand","employ") (First 200. Sent.)	-0.042
Author: Num. Top Field Co-authors (Dup)	0.061	("consum","consumpt") (First 100. Sent.)	-0.042
"road" (Full Text)	0.060	("privat","land") (First 200. Sent.)	-0.042

Note: The top 50 variables are ranked by the absolute value of their estimated coefficients. The categorical variables constructed from publication years are not shown in this table, because they are not very informative. Nevertheless, some of these categorical variables are top predictors. "Au" is an abbreviation of "Author", "Co-au" is an abbreviation of "Co-author", and "Sent." is an abbreviation of "Sentences". The "First 10 Sent.", "First 100 Sent.", "First 200 Sent.", and "Full Text" in brackets show the place of the appearance of the topic words and topic word pairs. "Nodup" is an abbreviation of "Non-duplicated", "Dup" is an abbreviation of "Duplicated", and "Cum" is an abbreviation of "Cumulative". "Nodup", "Dup", "Cum", and "Point" indicate the variables are constructed by the specific ways described in Appendix B.2.

Figure 1.9: Predicted distribution v.s. actual distribution

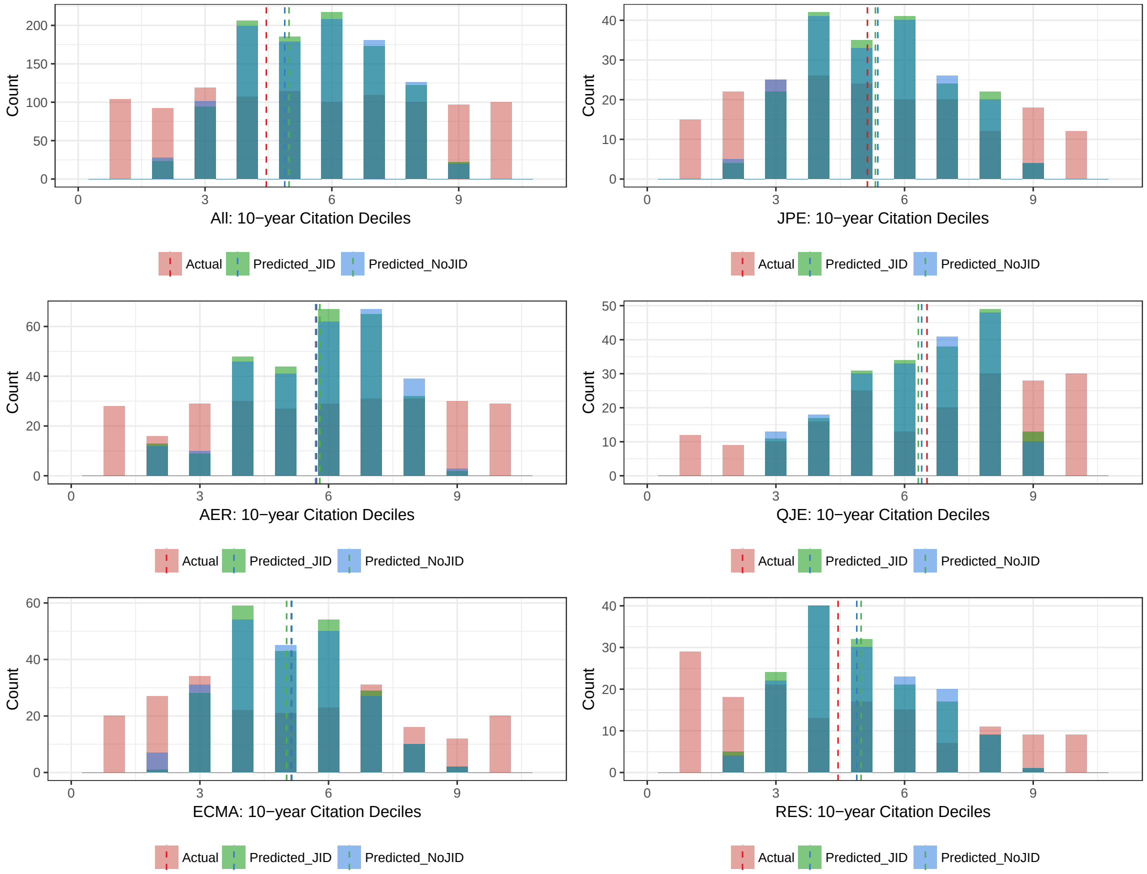
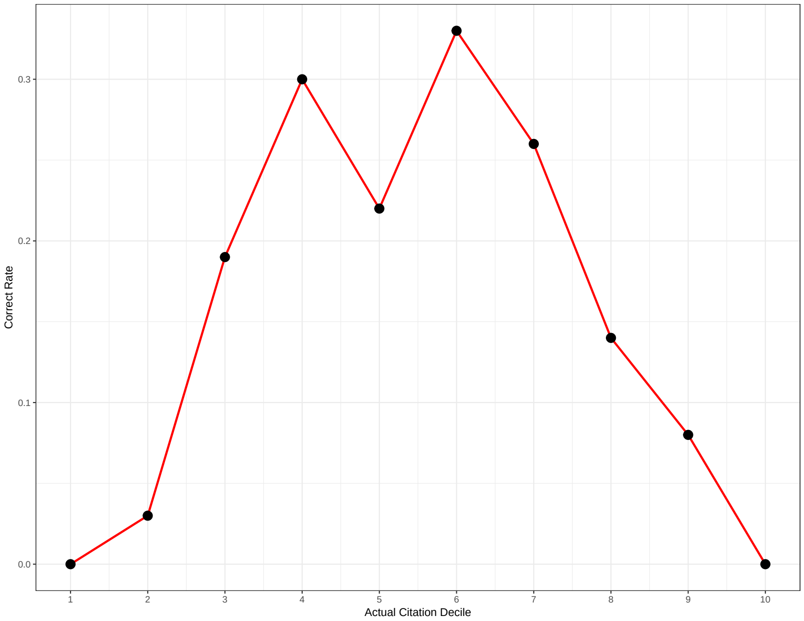


Table 1.11: Kolmogorov-Smirnov test for predicted distribution with and without journal ID

	All Top 5	**AER**	**ECMA**	**JPE**	**QJE**	**RES**
Kolmogorov-Smirnov statistic	0.012	0.050	0.049	0.031	0.052	0.067
P-value	1.000	0.875	0.952	1.000	0.958	0.891

Figure 1.10: Correct rate of decile prediction



1.5.2 Application to the academic publishing process

In this subsection, I discuss the potential of the prediction models to be used as first stage screening tool in the academic publishing process. The preferred Ordinary Least Squares model (OLS model (1)) is used as the baseline model to be compared with the preferred shrinkage model (Cross-Validation Elastic Net model (9)), tree model (Gradient Boosted Trees model (7)), and hybrid model (Shrinkage-Gradient Boosted Trees Hybrid method (6)).

Two-category case

In the first test, I test prediction models' ability to identify the papers in the upper half of the citation distribution. First, I separate papers into two citation categories: the "highly cited" category if a paper is in the upper half of the citation distribution ($D_{i,10} > 5$) and the "lowly cited" category if a paper is in the lower half of the citation distribution ($D_{i,10} \leq 5$). Then, I use each preferred prediction model to predict 10-year citation decile of papers in the testing set. Last, I code the predicted citation category of each paper in the testing set based on paper's predicted 10-year citation decile.

Define Positive (P) as the number of real "highly cited" papers, Negative (N) as the number of real "lowly cited" papers, True Positive (TP) as the number of correctly predicted "highly cited" papers, True Negative (TN) as the number of correctly predicted "lowly cited" papers, False Positive (FP) as the number of incorrectly predicted "highly cited" papers, and False Negative (FN) as the number of incorrectly predicted "lowly cited" papers in the testing set. Then, I use "$Precision$", "$Recall$", "$Accuracy$", "$Pr(High|Predicted\ high)$", and "$Pr(Low|Predicted\ low)$" to assess the prediction performance of these models. Their formulas are shown in Equations 1.22-1.26.

$$Precision = \frac{TP}{TP + FP} \tag{1.22}$$

$$Recall = \frac{TP}{TP + FN} \tag{1.23}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1.24}$$

$$Pr(High|Predicted\ high) = \frac{N(High \cdot Predicted\ high)}{N(Predicted\ high)} \tag{1.25}$$

$$Pr(Low|Predicted\ low) = \frac{N(Low \cdot Predicted\ low)}{N(Predicted\ low)} \tag{1.26}$$

Table 1.12 shows that the prediction result of Ordinary Least Squares model is almost equivalent to a random guess, whereas the Elastic Net, Gradient Boosted Trees, and Shrinkage-Gradient Boosted Trees Hybrid model have much higher precision rate, recall rate, and accuracy rate. In addition, Shrinkage-Gradient Boosted Trees Hybrid model has marginally better prediction performance than the Elastic Net model and

Gradient Boosted Trees model. Within the papers being predicted by the Shrinkage-Gradient Boosted Trees Hybrid model to be highly cited, 72.7% of them turn out to be highly cited, and within the papers being predicted to be lowly cited, 76.7% of them turn out to be lowly cited after 10 years of publication. Suppose an editor used paper's expected citation as one of the criteria in editorial decision making. Then, the Shrinkage-Gradient Boosted Trees Hybrid model may be a useful tool by giving "Accept" suggestion if a paper is predicted to be highly cited, and giving "Reject" suggestion if a paper is predicted to be lowly cited.

Table 1.12: Prediction performance test: 2 categories

	OLS		**Elastic Net**		**GB Trees**		**Hybrid**	
	High	Low	High	Low	High	Low	High	Low
Predicted high	186	177	368	146	383	144	387	145
Predicted low	320	359	138	390	123	392	119	391
Precision	51.24%		71.60%		72.68%		72.74%	
Recall	36.76%		72.73%		75.69%		76.48%	
Accuracy	52.30%		72.74%		74.38%		74.66%	
Pr(High|Predicted high)	51.24%		71.60%		72.68%		72.74%	
Pr(Low|Predicted high)	48.76%		28.40%		27.32%		27.26%	
Pr(Low|Predicted low)	52.87%		73.86%		76.12%		76.67%	
Pr(High|Predicted low)	47.13%		26.14%		23.88%		23.33%	

Note: OLS result is from OLS model (1), Elastic Net result is from Cross-Validation Elastic Net model (9), GB Trees result is from Gradient Boosted Trees model (7), and Hybrid result is from Shrinkage-Gradient Boosted Trees Hybrid method (6).

Three-category case

In the second test, I test the prediction models' ability to identify papers in three citation distribution categories: "highly cited" ($D_{i,10} > 7$), "middle" ($4 \leq D_{i,10} \leq 7$), and "lowly cited" ($D_{i,10} < 4$). Then, I test each prediction model's ability to label the correct citation category of each paper in the testing set.

Table 1.13 shows that the Ordinary Least Squares model cannot identify the "highly cited" papers, and only can identify the "lowly cited" papers correctly less than 50% of the time. Whereas, the other three models can identify the "highly cited" papers and the "lowly cited" papers more than 64% of the time. Within the papers being predicted by the Shrinkage-Gradient Boosted Trees Hybrid model to be "highly cited", 65.0% of them turn out to be "highly cited", and only 4.7% of them turn out to be "lowly cited". Within the papers being predicted to be "lowly cited", 66.7% of them turn out to be "lowly cited", and only 2.6% turn out to be "highly cited" after 10 years of publication. Given that the chance of predicting "highly cited" papers as "lowly cited" and the chance of predicting "lowly cited" papers as "highly cited" are low, the hybrid method proposed in this chapter may be helpful in identifying articles that are sufficiently below the acceptance threshold of a journal to enable editors to reject a significant fraction of inappropriate or low-quality submissions, as well as preventing rejection of submissions that will turn out to be highly cited.

Table 1.13: Prediction performance test: 3 categories

	OLS			**Elastic Net**			**GB Trees**			**Hybrid**		
	High	Middle	Low	High	Middle	Low	High	Middle	Low	High	Middle	Low
Predicted high	0	0	0	119	56	10	151	68	9	152	71	11
Predicted middle	253	337	217	176	328	200	141	300	163	139	289	152
Predicted low	44	93	98	2	46	105	5	62	143	6	70	152
Pr(High|Predicted high)		-			64.32%			66.23%			64.96%	
Pr(Low|Predicted high)		-			5.41%			3.95%			4.70%	
Pr(Low|Predicted low)		41.70%			68.63%			68.10%			66.67%	
Pr(High|Predicted low)		18.72%			1.31%			2.38%			2.63%	

Note: OLS result is from OLS model (1), Elastic Net result is from Cross-Validation Elastic Net model (9), GB Trees result is from Gradient Boosted Trees model (7), and Hybrid result is from Shrinkage-Gradient Boosted Trees Hybrid method (6).

The prediction performance of the proposed Shrinkage-Gradient Boosted Trees Hybrid method is much better than OLS. The proposed hybrid method cannot predict the citation counts of the highest and lowest deciles well. Instead, it performs better in the middle of the distribution. Despite this, the hybrid method shows practical value by predicting (based on information only available at time of publication) whether a paper's citations after 10 years of publication will be in the top 30% of the distribution correctly 65% of the time, and only 4.7% of the predicted top 30% papers turn out to be in the bottom 30% of the distribution.

1.6 Conclusion

The 10-year citations of papers in the top 5 economics journals is a highly right-skewed distribution, and the upper tail of the distribution is well approximated by a power law. I use some new measures of features of papers and authors to estimate paper effect and author effect on the 10-year citation distribution and citation paths. The estimation results show that papers that have higher 10-year citations are associated with higher popular topic coverage, numbers of authors, and total citations of authors' co-authors, while associated with lower "Micro" intensity, paper complexity, and numbers of authors' top field publications.

I also use the measures of features of papers and authors as predictors in machine learning models to predict papers' 10-year citations. The hybrid method developed in this chapter performs much better than Ordinary Least Squares in 10-year citation out-of-sample prediction test while using a relatively small number of variables compared to other machine learning methods. This property of the hybrid method significantly reduces the cost of data collection and computation for using it to predict 10-year citation counts of a new paper. In addition, the estimation and prediction strategies of analyzing large-scale high dimensional data has potential to be used to investigate the drivers of decision making and predict decision making in other places, where large-scale high dimensional data are produced, such as media market, financial market, online shopping, and online social network.

The hybrid method has shown its potential to be used to help find highly cited papers, as well as being used as a first stage screening tool in the academic publishing process to more efficiently direct the scarce time of editors and referees. However, it cannot predict the citation counts of the highest and lowest deciles well. It would be interesting for future study to identify additional features of papers and authors that predict paper citations, as well as exploring other types of hybrid methods, such as the hybrid of regression shrinkage and deep neural network. It would also be interesting to see whether prediction performance can be further improved by embedding unsupervised learning methods.

Papers in QJE get higher citations even after controlling a variety of paper and author information, though the QJE effect becomes much less important in prediction models with many variables. The QJE effect could either be caused by the differences in editors' preferences for papers that would be highly cited, or be caused by the differences in the pools of submitted manuscripts. The data used in this chapter only contain the papers that are published, and rejected ones are not observed. In the second chapter, I use data on manuscript submissions (including rejected ones) and records of decision making for four academic journals, linked with yearly paper citation data and author data to investigate the drivers of editorial decision making, duration from submission to decision, and paper citations. The data in journal database might help deepen our understanding of journal effect on paper citations.

Chapter 2

What Drives Editor Decisions?

2.1 Introduction

Editorial decision making has significant consequence for academics, as the phrase "publish or perish" indicates. Since editors' decisions can have a big impact on the promotion of professors and structure of academia, it is natural to ask: What is the objective of an editor or a journal? How do they decide which papers to publish and which to reject? Even though many journals have transferred their journal databases to online submission and editorial tracking systems where the records of editorial decision making can be tracked, editorial decision making is still unclear to researchers who are not involved in the process. Some empirical studies find recently published economics papers are written by bigger teams of more experienced authors [13, 37] and are longer [13], and the research field and style has evolved to be more empirical [3, 37]. Since the decision making of editors has an important impact on the evolution of the field and style of academic papers, analyzing the process of editorial decision making may help deepen our understanding of the process of editorial decision making.

Researchers have developed models of editorial decision making to investigate the increasing tendency of academic journals to require multiple revisions of articles [21], the relationship between evolving social norms and first response times [5], and the decision process of journal editors [7, 8, 14]. However, constrained by the availability of data on manuscripts and the technical challenge of analyzing textual data, none of these studies has included detailed measures of paper content in their analysis. In addition, previous studies, except for Bandeh-Ahmadi [7] and Card and DellaVigna

[14], did not have access to the data that links referee recommendations and editorial decisions.

In this study, I analyze data on the keywords and abstracts of submissions (including both accepted papers and rejected papers) and records of decision making, including referee recommendations, actions of authors, and the accept, reject and revise and resubmit decisions by the editor in charge of the submissions of four academic journals. The data provided by these journals offers a unique opportunity to study editorial decision making, and the constructed database that links paper and author information with decisions of editors and referees can improve our understanding of the academic publishing process. To measure the content of each paper, I use dictionary-based textual analysis algorithms to analyze each paper's abstract and keywords to construct variables measuring its topics and research fields. The dictionary-based textual analysis used in this chapter has been used in other economic studies, including the measurement of investor sentiment [63], media slant [30], tone in financial text [52], and economic policy uncertainty [6].

Apart from the study on editorial decision making in the academic publishing process, some studies have discussed the motivation of serving as referees [24] and referees' opinions of what a good paper was [68]. Other studies have analyzed the effects of submission fees and time delays in the academic publishing process and propose ways of using submission fees and time delays to maximize journal quality [4, 18, 48]. Using the data of economic journals, Ellison [22] shows that despite that following the adoption of online editorial software, the duration of academic publishing process has actually increased. However, the factors affecting the lags from submission to editorial decision process have yet to be examined.

Some researchers have evaluated the value added of the academic publishing process, and Card and DellaVigna [14] have shown that one aspect of the value added of journal editors in identifying and publishing papers: The average number of citations of accepted papers is higher than rejected ones. Bandeh-Ahmadi [7] has shown the language from referee comments can help predict paper citations, and Laband [45] has shown the value added of the peer review system in improving paper quality. However, it is unclear whether the relatively higher average number of citations of published papers is a result of editors' accepting papers that are more likely to be highly cited or by selecting relatively higher quality papers, which tend to also have higher citations. Thus, it is unclear how much the editors weigh expected citations when they decide which paper to publish. As will be shown in this chapter, there is high skewness in the citation distribution even in papers that are published in the same year by the same journal. In addition, it is not uncommon to find a published paper that is not highly cited, as well as rejected submissions that are subsequently published and highly cited.

In the first chapter of this thesis, I showed that the 10-year citation distribution of papers in top economics journals is highly right-skewed, and the upper tail of the distribution is well approximated by a power law. To understand the causes of the skewed citation distribution, the first chapter investigated the drivers of paper citations. This chapter further investigates the drivers of the citation distribution of published papers and tries to answer the following question: Do editors and referees prefer to publish papers that are expected to be highly cited or is citation correlated with "quality"? Bayar and Chemmanur [8] assume a journal editor maximizes his journal's expected payoff from publishing high-quality papers, net of costs due to (mistakenly) publishing low-quality papers, and Card and DellaVigna [14] assume that editors maximize the expected quality of accepted papers and citations are unbiased measures of quality. This chapter assumes that publishing high-quality papers is one of the objectives of journal editors, and "quality" or "scientific impact" is measured by citations. Admittedly, the model developed in this chapter cannot really distinguish an editor who is trying to maximize citations of the papers he/she accepts or the quality of the papers he/she accepts if quality is correlated with citations.

A paper's quality cannot be directly measured. However, the number of citations provides a quantitative measure of the impact of the paper on the subsequently published papers. Price [56] and Redner [57] find the distribution of citations of papers in various academic journals are highly skewed, and paper citation distribution follows a power law distribution. In addition, there are differences in the citation paths of published papers, and statistical models are developed to investigate the effects of paper's topic and field on determining citation path [2, 65, 67].

To understand the skewness of citation distribution, I link the data from the journal databases and the citation data from Google Scholar for both accepted papers and rejected papers (i.e., papers that were not published by one journal, though they may have been published later by some other journal). Compared to the citations of rejected papers, the citations of accepted papers are much higher. However, there is overlap between the citation distribution of accepted and rejected papers, meaning that a proportion of accepted papers have lower cumulative citations than some of rejected ones. The overlap may indicate that it is hard to use the information available at the time of submission to identify papers that will be highly cited in the long run assuming that was the editor's objective. Alternatively, it may indicate that there are factors other than expected citations that drive editorial decision making.

To investigate the drivers of editorial decision making, I use regression methods to estimate the effects of features measuring paper information, author information, and referee recommendation on editorial decision making. Since most of the submissions to the journals used in this study only had one round of peer review, I focus on the decision making of the desk review round and the first round of peer review.

Empirical results confirm the effects of referee recommendation on editorial decision making. The results suggest that papers with higher referee recommendation scores and scientific contribution scores are more likely to be published. Recommendation score and scientific contribution score reflect referees' evaluation of the quality of the paper that is reviewed, and the positive effect of these scores indicate that editors do rely on their evaluations when making their decisions. In addition, papers with a higher measure of disagreement by referees are less likely to be published. One explanation could be that the papers with more disagreement are more "risky", and some of them may turn out to be papers that are not scientifically impactful. I also use the fractions of referee recommendation and scientific evaluation scores as alternative measures of signals from referees. The results suggest that a higher share of positive referee recommendations is associated with a higher probability of passing the first round of peer review.

I find there are particular paper attributes/features that make them more likely to be accepted. For example, papers with higher coverage of "popular" research topics are more likely to be published. One explanation could be that popular topics attract the interest of a broad range of researchers. Another explanation could be popular topics are areas where high-quality work is being done. In addition, the papers that do not cover enough popular topics are either too narrow or too far away from the scope of the target journal, which reduces their chance of being published.

The results for editorial decision making also suggest that papers by authors with a higher number of submissions and lower rejection rate are more likely to be published. Card and DellaVigna [14] find that referees and editors tend to be more supportive of less prolific authors. However, we find that editors are more inclined to publish papers by authors with higher number of submissions and lower rejection rate. One possible explanation can be that the authors with these characteristics have decent records in the journal database, which makes it easier for editors to evaluate the quality of their work. Another possible explanation can be that the authors with these characteristics could be better researchers, and the "author effect" is a proxy for paper quality.

I also investigate the effects of paper and author information, editors' decisions, and referee recommendations on the duration from submission to decision and paper citations. For non-desk-rejected papers, the paper and author information does not have a significant effect on the duration of the first round of review. However, the papers with higher referee recommendation scores and lower standard deviation of the scores have shorter durations of the first round of review. The results for paper citations suggest that accepted papers on average get higher citations than rejected ones, and higher paper citation counts are associated with higher coverage of popular research topics, referee recommendation scores, and scientific contribution scores.

In the prediction part, I use a variety of state-of-the-art machine learning methods, including regression shrinkage models (Lasso, Post-Lasso, Ridge, and Elastic Net) in Zou and Hastie [69] and Belloni et al. [9], Random Forest [12], and Gradient Boosted Trees [25, 26] to predict paper citations with the information available at the time of submission. The model that uses Random Forest method, measures of publication information, measures of research fields and topics, and high dimensional measures of the appearance of popular topic words gives the best out-of-sample prediction performance.

Using the preferred prediction model, I test the possibility of combining artificial intelligence (AI) and human experts in the academic publishing process. I compare four alternative academic publishing processes that use different levels of human intelligence and artificial intelligence. The experiment shows that the average number of cumulative citations of papers accepted by editors is 24% higher than all submissions. This result suggests that papers selected for publication turn to have higher average citations than rejected ones, even though editors may not use paper's expected citations as one of the criteria when they decide which paper to publish. As an exercise, I use the citation prediction model to decide which papers to publish based on maximizing citations. For a comparable acceptance rate as the human-based editorial process, the papers published by the algorithm have 2% higher citation counts. In addition, the average number of cumulative citations of the papers of both human choices and artificial intelligence choices is 22% higher than all publishable papers. Admittedly, there are other factors that affect editors' decision on which paper to publish. However, the artificial intelligence based prediction model may help editor to identify the papers that are more likely to be highly cited from publishable papers.

In Section 2.2, I present theoretical model. In Section 2.3, data collection and textual analysis are presented. In Section 2.4, I discuss estimation and prediction strategy. In Section 2.5, results are presented and discussed. In Section 2.6, I conclude.

2.2 Model

2.2.1 An overview of the academic publishing process

Nowadays, most of the submissions enter peer review system before being published. In the peer review system, the editor and peer reviewers evaluate the quality of the paper, and the editor decides whether or not to accept a paper for publication. In the meanwhile, they provide comments on the paper that is reviewed.

Editors are constrained by the number of papers that can be published in an issue and the time they can devote to editorial service. Under these constraints, editors select a subset of papers from the submitted papers to publish. A representative editor's decision tree in the academic publishing process is presented in Figure 2.1. After a new submission is received, the editor decides whether to desk reject a paper or not. If a submission is not desk rejected, it is sent to referees for peer review. After all (or some) of the review reports are received, the editor decides whether to reject the paper, ask for a revision, or accept the paper as it is. If a paper is not rejected in the first round of peer review, the editor is most likely to send a revision request to the corresponding author. If the corresponding author chooses to revise and resubmit the paper, the paper will re-enter the peer review system, and the editor will decide whether to accept the paper as it is, send it to reviewers for another round of peer review, or reject it.

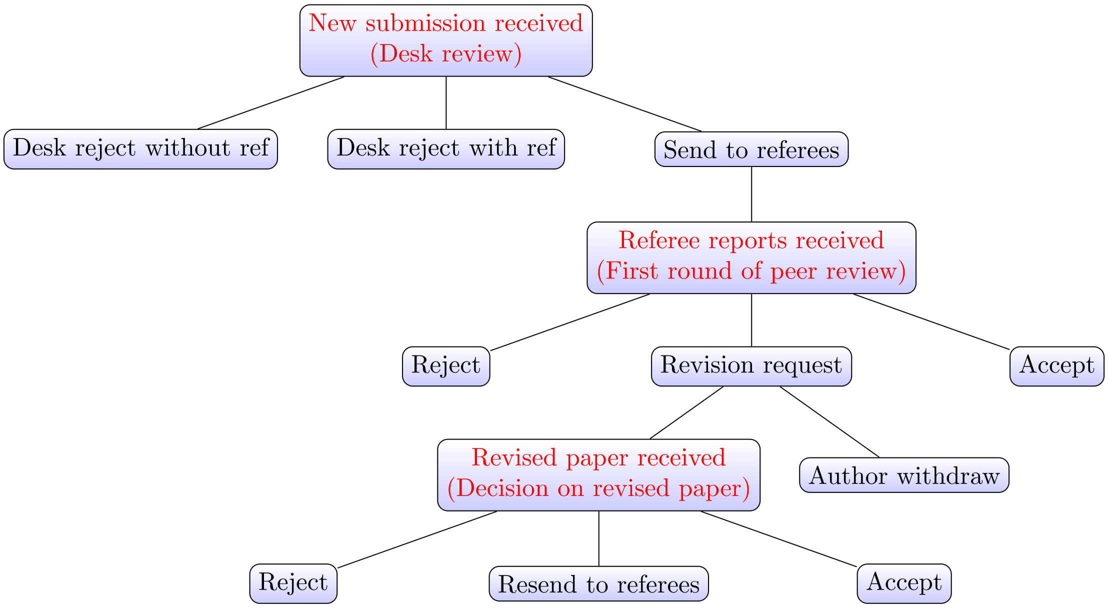


Figure 2.1: Editor's decisions in the academic publishing process

In each round of decision making, the editor imperfectly assesses the quality of the paper by reading the paper, and knowing the research records of the authors, as well as signals from referees once referee reports are received. To analyze editorial decision-making process, the model makes the following assumptions: 1. Editors evaluate a paper's quality using information related to the paper and its authors and makes editorial decisions based on his/her evaluation of paper quality; 2. The number of citations is an indicator of the paper's quality.

As will be shown in Section 2.5.1, most of the submissions to the journals used in this study only had one round of peer review. Thus, I focus on the decision making in the desk review round and the first round of peer review.

2.2.2 Decision making in the academic publishing process

Editor's decision in the desk review round

Decision making in the academic publishing process starts when a new submission is received by the editor. After receiving a new submission, the editor reads the paper quickly and imperfectly observes paper information, author information and existing literature related to the new submission. Based on the observed information, the editor chooses between "Desk reject without referee input", "Desk reject with referee input" and "Send to referees".[1] In my model, I assume referee input in the desk reject decision is negligible and combine the two desk rejection categories as "Desk reject".

Based on paper and author information, the editor evaluates the true quality of new submission i as Q_i. Q_i cannot be observed by the econometrician. I assume Q_i can be written as:

$$Q_i = f(W_i, A_i, F_i) + \epsilon_i \tag{2.1}$$

where $f(W_i, A_i, F_i)$ is a function of new submission's objectively measurable features including topic words (W_i), author information (A_i), and field information (F_i), and

[1] Both "Desk reject without referee input" and "Desk reject with referee input" are considered as "Desk Reject". "Desk reject with referee input" is also known as "Summary reject". In the journal databases used in this study, no paper is desk accepted.

ϵ_i is an unobservable error term, which is observed by the editor but not by the econometrician.

Define the editor's decision on new submission i in the desk review round as d_i^0, and code the possible outcomes as: $d_i^0 = 0$ if "Desk reject" and $d_i^0 = 1$ if "Send to referees". The probability of observing editorial decision in the desk review round for ordered logit is shown in Equations 2.2 - 2.3.

$$\begin{aligned} & Pr(d_i^0 = 0|Q_i) \\ =& Pr(f(W_i, A_i, F_i) + \epsilon_i < \delta_1^0) \\ =& \frac{1}{1 + e^{-\delta_1^0 + f(W_i, A_i, F_i)}} \end{aligned} \tag{2.2}$$

$$\begin{aligned} & Pr(d_i^0 = 1|Q_i) \\ =& Pr(\delta_1^0 < f(W_i, A_i, F_i) + \epsilon_i) \\ =& 1 - \frac{1}{1 + e^{-\delta_1^0 + f(W_i, A_i, F_i)}} \end{aligned} \tag{2.3}$$

where δ_1^0 is the cutoff in the desk review round.

Signals from referees

This chapter does not model referee's decision in referee report writing process. Instead, I focus on investigating the effect of the signals from referees on editorial decision making. If the editor chose "Send to referees" in the desk review round, paper i would enter peer review round. Let S_i^r represent the signal sent by referee r. The signal S_i^r includes referee recommendation score and scientific contribution score. In the peer review system, most papers receive more than one referee report. I assume the editor weighs signals from multiple referees following a deterministic function that can be written as:

$$S_i = f(S_i^1, ..., S_i^R) \tag{2.4}$$

Editor's decision on new submission in the peer review round

After receiving referee reports, the editor updates her/his evaluation of new submission i to $Q_i^{'}$. I assume $Q_i^{'}$ can be written as:

$$Q_i^{'} = f(W_i, A_i, F_i, S_i) + \epsilon_i^{'} \tag{2.5}$$

where $f(W_i, A_i, F_i, S_i)$ is a function of new submission's topic words (W_i), author information (A_i), field information (F_i), and weighted signal from referees (S_i), and $\epsilon_i^{'}$ is an "unobservable" error term.

Define the editor's decision on new submission i in the first round of peer review as d_i^1, and code the possible outcomes in the desk review round as: $d_i^1 = 1$ if "Reject", $d_i^1 = 2$ if "Revision request", and $d_i^1 = 3$ if "Accept". The probability of observing an editor's decision on new submission in peer review round for ordered logit model is shown in Equations 2.6 - 2.8.

$$\begin{aligned} & Pr(d_i^1 = 1 | Q_i^{'}) \\ =& Pr(f(W_i, A_i, F_i, S_i) + \epsilon_i^{'} < \delta_1^1) \\ =& \frac{1}{1 + e^{-\delta_1^1 + f(W_i, A_i, F_i, S_i)}} \end{aligned} \tag{2.6}$$

$$
\begin{aligned}
&Pr(d_i^1 = 2|Q_i^{'}) \\
=&Pr(\delta_1^1 < f(W_i, A_i, F_i, S_i) + \epsilon_i^{'} < \delta_2^1) \\
=&\frac{1}{1 + e^{-\delta_2^1 + f(W_i, A_i, F_i, S_i)}} - \frac{1}{1 + e^{-\delta_1^1 + f(W_i, A_i, F_i, S_i)}}
\end{aligned} \tag{2.7}
$$

$$
\begin{aligned}
&Pr(d_i^1 = 3|Q_i^{'}) \\
=&Pr(\delta_2^1 < f(W_i, A_i, F_i, S_i) + \epsilon_i^{'}) \\
=&1 - \frac{1}{1 + e^{-\delta_2^1 + f(W_i, A_i, F_i, S_i)}}
\end{aligned} \tag{2.8}
$$

where δ_1^1 and δ_2^1 are the cutoffs in the first round of peer review.

Since the journals used in this study only have one editor each, the editor fixed effect is not specified. The estimation for decision making in the peer review rounds may suffer from selection bias since only the papers that are not desk-rejected enter the peer review rounds. For the journals used in this study, the editor-in-chief of each journal makes the desk reject decision and if he/she does not desk reject it, a co-editor may be assigned to handle the submission, though it may also be the same editor. However, even if an editor is assigned to make both desk review decision and peer review decision, there is additional information in the cover letters and referee reports as well as additional information the editor obtains from reading the paper more closely than the first time when making desk review decision.

I assume that the decisions made by editors in the peer review rounds are independent of the decisions in the desk review rounds, and ϵ_i and $\epsilon_i^{'}$ are independently distributed. However, in future work I will consider error structures that have unobserved "random effect" components in the error terms that would allow the shocks to be correlated and then the selection bias would be explicitly controlled for via a maximum likelihood estimator that integrates out the unobserved random effect term entering the unobserved quality of the paper as perceived by the editor.

2.2.3 Efficiency of the academic publishing process

The objective of the academic publishing process is to identify and publish high-quality papers, as well as to improve the quality of papers that are reviewed even if they are rejected. However, the review process takes time. There is unavoidable trade-off between quick time to decision and quality of decisions, but the long durations to receive referee reports could be associated with more uncertain, intermediate quality papers where there is also disagreement by the referees. It may take longer to read/understand a poorly written paper than a well written one and so intermediate quality papers may take longer for referees to read and understand than either very bad papers or very good ones. In this chapter, I assess both the duration of the academic publishing process and its performance in publishing high-quality papers[2] in order to evaluate the efficiency of the academic publishing process.

The delays in the academic publishing process arise from the time needed for editors and referees to evaluate the quality of submissions and the time needed for the authors to revise their papers. In Section 2.4.2, I discuss the strategy for estimating the effects of the objectively measurable information in papers, authors, and referee recommendations on the duration of the first round of peer review. Even though there are unobservable factors affecting the duration, it is interesting to see how much of the variation can be explained by the objectively measurable features.

To assess editorial performance, I compare the cumulative citations of accepted and rejected papers. The number of citations is driven by readers' decision to cite the paper when he/she writes a new paper, which provides a measure of the impact of the paper on subsequently published papers. In Section 2.4.3, I discuss a strategy for eval-

[2] As discussed before, I assume that the number of citations can be used as a measure of paper's quality.

uating the academic publishing process's performance of identifying and publishing high-quality papers.

2.3 Data Collection and Textual Analysis

2.3.1 Data sources

I analyzed data from the editorial databases of four academic journals, and linked to citation data from Google Scholar, and available information in the public domain. The constructed database has data on 13,517 papers submitted to these journals between 2005 and 2017.

The academic journal database is constructed using editorial databases of four academic journals under a Non-Disclosure Agreement (NDA) signed by individual journals and the author. The editorial database contains records of the actions of editors, referees, and authors, as well as full-text copies of paper manuscripts, referee reports, editor decision letters, and each time stamped and related to each stage of the academic publishing process. The details on extracting data from the journal database are documented in Appendix A.3.

I collected paper citation data from Google Scholar based on match of title. The detailed algorithms for collecting data from Google Scholar are described in Appendix A.1.2.

For each paper, I searched for its title on Google Scholar. Then, I compared the returned author names with the author names in the academic journal database to check whether the returned paper was correct or not. If the returned paper was correct, I recorded the number of its cumulative citations as of the end of November 2017.[3] The citation data of 10,693 out of 13,517 papers (including rejected ones) could be

[3] The citation data was collected in the last week of November 2017.

retrieved from Google Scholar. Within the papers that have citation data, 5,128 of them are accepted papers, and 5,565 of them are rejected papers.

2.3.2 Textual analysis and variables

Paper textual information

I used dictionary-based textual analysis algorithms to analyze each paper's abstract and keywords to construct variables measuring its topics and research fields. The dictionaries were composed of the frequent topic words appeared in each of the journals used in this study. Due to the non-disclosure agreement, I cannot disclose the names or the research fields of these journals. The algorithms for measuring paper abstract and keywords are documented in Appendix B.1.3.

For each paper, I constructed a vector of variables that measured the paper's topic words including coverage of popular topic words, coverage of popular two-word pairs, coverage of popular three-word pairs, coverage of popular four-word pairs, and dummy variables for popular topic words and word pairs in the keywords part and abstract part of each paper.[4]

To measure the research fields of each paper, I used research field dictionaries to construct a vector of variables including research field dummy and research field intensity in the keywords part and abstract part.[5]

[4] Each of the dictionaries for measuring the coverages of popular topic words and word pairs was composed of about 500 keywords (or keyword pairs). The differences in the numbers of keywords (or keyword pairs) were due to the words with the same frequency. The coverage of popular topic words measured the percentage of the appearance of 532 highly frequent keywords. The coverage of popular two-word pairs measured the percentage of the appearance of 531 highly frequent two-keyword pairs. The coverage of popular three-word pairs measured the percentage of the appearance of 505 highly frequent three-keyword pairs. The coverage of popular four-word pairs measured the percentage of the appearance of 451 highly frequent four-keyword pairs.

[5] The words in field dictionaries were selected according to the frequency of word appearance in the keywords part of the papers used in this study. Each of the field dictionaries

Information about authors, editors, and referees

The data on the information about authors, editors, and referees were extracted from the academic journal database. The data in the academic journal database is stored under each paper ID, and I reorganized the database by assigning separate records to the author, editor, and referees associated with it. Then I constructed the information about authors, editors, and referees for each paper.

For each paper, I constructed a vector of variables that measured author information including the number of submissions to each journal and their rejection rate in each journal. The information about authors' publication records and collaboration network were not used for two reasons. First, most of the authors of the papers used in this study did not have personal websites, and their resumes were not available. Thus, when editors decided which submission to accept for publication, and readers decided which paper to cite in a new publication, they might not be able to observe author's publication record or use that information in decision making. Second, it was challenging to identify these authors' publication records and construct their collaboration network using the information from the public domain. It was not uncommon to find two scholars having the identical first name and last name, which made it more difficult to identify the correct author. Given these difficulties, I decided not to use the information about author's publication record and collaboration network in this study.

The variables summarizing the editorial decision-making process include the editorial decision on new submissions in the desk review round, the editorial decision on new submissions in the first round of peer review, duration of the desk review round,

was composed of about 200 words. The differences in the numbers of keywords were due to the words with the same frequency. Field dictionary A was composed of 205 words, field dictionary B was composed of 207 words, field dictionary C was composed of 208 words, and field dictionary D was composed of 193 words.

and duration of each round of peer review. To measure the signals from referees, I constructed statistics of the referee recommendation scores and scientific contribution scores for each paper. The statistics include mean, standard deviation, minimum, and maximum of the scores each paper received.

2.4 Estimation and Prediction Strategy

2.4.1 Estimation strategy for decision making

I estimate the decision making in the desk review round and the first round of peer review with the information available to editors in each stage. The estimation equations for estimating editorial decision on new submission in the desk review round are given by Equations 2.9 - 2.10, and the estimation equations for estimating editorial decision on new submission in the first round of peer review are given by Equations 2.11 - 2.13.

$$\begin{aligned} &Pr(d_i^0 = 0|Q_i) \\ =&\frac{1}{1+e^{-\delta_1^0+\boldsymbol{W_i'\beta_d}+\boldsymbol{A_i'\eta_d}+\boldsymbol{F_i'\theta_d}+\boldsymbol{J_i'\rho_d}}} \end{aligned} \tag{2.9}$$

$$\begin{aligned} &Pr(d_i^0 = 1|Q_i) \\ =&1-\frac{1}{1+e^{-\delta_1^0+\boldsymbol{W_i'\beta_d}+\boldsymbol{A_i'\eta_d}+\boldsymbol{F_i'\theta_d}+\boldsymbol{J_i'\rho_d}}} \end{aligned} \tag{2.10}$$

$$\begin{aligned} &Pr(d_i^1 = 1|Q_i^{'}) \\ =&\frac{1}{1+e^{-\delta_1^1+\boldsymbol{W_i'\beta_p}+\boldsymbol{A_i'\eta_p}+\boldsymbol{F_i'\theta_p}+\boldsymbol{S_i'\kappa_p}+\boldsymbol{J_i'\rho_p}}} \end{aligned} \tag{2.11}$$

$$\begin{aligned}
&Pr(d_i^1 = 2|Q_i^{'}) \\
=&\frac{1}{1+e^{-\delta_2^1+\boldsymbol{W_i'\beta_p}+\boldsymbol{A_i'\eta_p}+\boldsymbol{F_i'\theta_p}+\boldsymbol{S_i'\kappa_p}+\boldsymbol{J_i'\rho_p}}} \\
&-\frac{1}{1+e^{-\delta_1^1+\boldsymbol{W_i'\beta_p}+\boldsymbol{A_i'\eta_p}+\boldsymbol{F_i'\theta_p}+\boldsymbol{S_i'\kappa_p}+\boldsymbol{J_i'\rho_p}}}
\end{aligned} \tag{2.12}$$

$$\begin{aligned}
&Pr(d_i^1 = 3|Q_i^{'}) \\
=&1-\frac{1}{1+e^{-\delta_2^1+\boldsymbol{W_i'\beta_p}+\boldsymbol{A_i'\eta_p}+\boldsymbol{F_i'\theta_p}+\boldsymbol{S_i'\kappa_p}+\boldsymbol{J_i'\rho_p}}}
\end{aligned} \tag{2.13}$$

where d_i^0, d_i^1 are records of editorial decision making in the desk review round and the first round of peer review, $\boldsymbol{W_i}$ is a vector of variables measuring topic words of paper i, $\boldsymbol{A_i}$ is a vector of variables measuring author information of paper i, $\boldsymbol{F_i}$ is a vector of variables controlling research field information of paper i, $\boldsymbol{S_i}$ is a vector of variables measuring signals from referees for paper i, $\boldsymbol{J_i}$ is a vector of variables controlling journal information of paper i, δ_1^0 is the cutoff in the desk review round, and δ_1^1 and δ_2^1 are the cutoffs in the first round of peer review.

I use Maximum likelihood Estimation (MLE) to estimate equation 2.9 - 2.13. The generic form of the log-likelihood is presented by Equation 2.14.

$$ln\mathcal{L} = \sum_{i=1}^{N}\sum_{j=1}^{J} Z_{i,j} ln[\Phi_{i,j} - \Phi_{i,j-1}] \tag{2.14}$$

where $Z_{i,j} = 1$ if $d_i = j$ and 0 otherwise, $\Phi()$ is the link function of the ordered logit model, $\Phi_{i,j} = \Phi(\delta_j - \boldsymbol{X_i\beta})$, and $\Phi_{i,j-1} = \Phi(\delta_{j-1} - \boldsymbol{X_i\beta})$.

Without some constraints on parameters, $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ are unidentified. To estimate the model, I fix the intercept β_0 as 0 and σ_i (the standard deviation of $Q_i^{'}$) as 1 to identify $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$.

2.4.2 Estimation strategy for the duration of the academic publishing process

As discussed by Ellison [21] and Ellison [22], the academic publishing process has been taking longer in recent years. To investigate the drivers of the duration of the academic publishing process, I estimate the effects of paper and author information, and signals from referees on the duration of the first round of peer review. Considering that the durations of the review of desk-rejected papers are very shorter, I focus on estimating the duration of the first round of peer review for non-desk-rejected papers. The estimation equation for the duration of the first round of peer review of non-desk-rejected papers is given by Equation 2.15.

$$T_i = \alpha_{dp} + \boldsymbol{W_i'}\boldsymbol{\beta_{dp}} + \boldsymbol{A_i'}\boldsymbol{\eta_{dp}} + \boldsymbol{F_i'}\boldsymbol{\theta_{dp}} + \boldsymbol{S_i'}\boldsymbol{\kappa_{dp}} + \boldsymbol{J_i'}\boldsymbol{\rho_{dp}} + \varepsilon_{dp,i} \qquad (2.15)$$

where T_i is the duration of the first round of peer review, $\boldsymbol{W_i}$ is a vector of variables measuring topic words of paper i, $\boldsymbol{A_i}$ is a vector of variables measuring author information of paper i, $\boldsymbol{F_i}$ is a vector of variables controlling field information of paper i, $\boldsymbol{S_i}$ is a vector of variables measuring signals from referees for paper i, $\boldsymbol{J_i}$ is a vector of variables controlling journal information of paper i, and $\varepsilon_{dp,i}$ is the error term.

2.4.3 Estimation strategy for paper citations

To investigate the drivers of paper citations, I estimate the paper effect, author effect, and referee effect on paper's cumulative citations. In addition, I evaluate the performance of the academic publishing process in identifying high quality submissions by comparing the cumulative citations of the accepted and rejected papers. The estimation equation is given by Equation 2.16.

$$C_i = \alpha + \boldsymbol{W_i'}\boldsymbol{\beta} + \boldsymbol{A_i'}\boldsymbol{\eta} + \boldsymbol{F_i'}\boldsymbol{\theta} + \boldsymbol{S_i'}\boldsymbol{\kappa} + \boldsymbol{J_i'}\boldsymbol{\rho} + \varepsilon_i \tag{2.16}$$

where C_i is the cumulative citations of paper i, $\boldsymbol{W_i}$ is a vector of variables measuring topic words of paper i, $\boldsymbol{A_i}$ is a vector of variables measuring author information of paper i, $\boldsymbol{F_i}$ is a vector of variables controlling field information of paper i, $\boldsymbol{S_i}$ is a vector of variables measuring signals from referees for paper i, $\boldsymbol{J_i}$ is a vector of variables controlling journal information of paper i, ε_i is the error term.

2.4.4 Prediction models

To predict paper citations, I use the OLS linear regression as the benchmark, and use a variety of state-of-the-art machine learning methods, including regression shrinkage methods (Lasso, Post-Lasso, Ridge, and Elastic Net) in Zou and Hastie [69] and Belloni et al. [9], Random Forest [12], and Gradient Boosted Trees [25, 26]. The details of these methods are described in the first chapter. In Section 2.5.3, I present prediction results.

2.5 Results and Discussion

2.5.1 Summary statistics

Editorial decision

Table 2.1 presents the categories of editorial decision by round. Decision round 1 includes one round of desk review and one round of peer review if the paper passes desk review, and the other decision rounds only have one round of peer review. Table 2.2 decomposes the decision making in decision round 1. It can be observed that about half of the submissions are desk rejected in decision round 1. For the submissions that are not desk-rejected, most of them are sent to referees and then revised and resubmitted to the journal. Similar to the situation in the economic field, only very few submissions are "accepted as is" in decision round 1. For most of the submissions, after they are revised and resubmitted, the editor makes a new round of decision, and more than 85% of them are accepted in decision round 2.

Table 2.3 presents the rejection ration rate in each round, and the average rejection rate of these journals is 60.4%. The rejection rate of the desk review round varies between 39.6% and 59.1% in these journals, and on average 50.7% of the submissions are rejected in the desk review round. However, the rejection rate drops sharply in the peer review rounds. Only 17.6% of the submissions are rejected in the first round of peer review, and more than 90% of the revised and resubmitted papers are accepted in the second to fourth round of peer review. Since the rejection rate is on average 50.7%, which means the pool of submissions is about twice as large as the pool of the papers being published, the possibility of not having enough submissions for the editor to select from is not a major concern for this chapter.

Table 2.1: Editorial decision by round

	Accept	Reject	Desk Reject	R&R	Withdrawn
All Journals					
Decision Round 1	14	1,173	6,854	5,476	
Decision Round 2	4,316	124		538	20
Decision Round 3	471	10		32	3
Decision Round 4	23	2		5	
Journal A					
Decision Round 1	6	421	1,751	2,246	
Decision Round 2	1,733	57		272	9
Decision Round 3	250	7		6	1
Decision Round 4	5	1			
Journal B					
Decision Round 1	6	387	2,456	1,544	
Decision Round 2	1,188	35		149	6
Decision Round 3	129	2		7	2
Decision Round 4	5	1			
Journal C					
Decision Round 1	1	278	2,175	1,229	
Decision Round 2	1,063	17		58	5
Decision Round 3	52	4			
Decision Round 4	3				
Journal D					
Decision Round 1	1	87	472	457	
Decision Round 2	332	15		59	
Decision Round 3	40	1		15	
Decision Round 4	10			5	

Note: Decision round 1 includes one round of desk review and one round of peer review if the paper passes desk review, and the other decision rounds only have one round of peer review.

Table 2.2: Decision making in the first round of review

	Accept	Reject	Desk Reject	R&R	Send to Referees
All Journals					
Desk Review			6,854 (50.7%)		6663 (49.3%)
Peer Review Round 1	14 (0.2%)	1173 (17.6%)		5476 (82.2%)	
Journal A					
Desk Review			1,751 (39.6%)		2673 (60.4%)
Peer Review Round 1	6 (0.2%)	421 (15.8%)		2246 (84.0%)	
Journal B					
Desk Review			2,456 (55.9%)		1937 (44.1%)
Peer Review Round 1	6 0.3%	387 (20.0%)		1544 (79.7%)	
Journal C					
Desk Review			2,175 (59.1%)		1508 (40.9%)
Peer Review Round 1	1 (0.1%)	278 (18.4%)		1229 (81.5%)	
Journal D					
Desk Review			472 (46.4%)		545 (53.6%)
Peer Review Round 1	1 (0.2%)	87 (16.0%)		457 (83.9%)	

Note: Percentages in parentheses.

Table 2.3: Rejection rate by round

Round	Rejected	Not Rejected	Rejection Rate
All Journals			
Desk Review	6,854	6,663	50.7%
Peer Review Round 1	1,173	5,490	17.6%
Peer Review Round 2	124	4,854	2.5%
Peer Review Round 3	10	503	1.9%
Peer Review Round 4	2	28	6.7%
Total	8,163	5,354	60.4%
Journal A			
Desk Review	1,751	2,673	39.6%
Peer Review Round 1	421	2,252	15.8%
Peer Review Round 2	57	2,005	2.8%
Peer Review Round 3	7	256	2.7%
Peer Review Round 4	1	5	16.7%
Total	2,237	2,187	50.6%
Journal B			
Desk Review	2,456	1,937	55.9%
Peer Review Round 1	387	1,550	20.0%
Peer Review Round 2	35	1,337	2.6%
Peer Review Round 3	2	136	1.4%
Peer Review Round 4	1	5	16.7%
Total	2,881	1,512	65.6%
Journal C			
Desk Review	2,175	1,508	59.1%
Peer Review Round 1	278	1,230	18.4%
Peer Review Round 2	17	1,121	1.5%
Peer Review Round 3	0	56	0.0%
Peer Review Round 4	0	3	0.0%
Total	2,470	1,213	67.1%
Journal D			
Desk Review	472	545	46.4%
Peer Review Round 1	87	458	16.0%
Peer Review Round 2	15	391	3.7%
Peer Review Round 3	1	55	1.8%
Peer Review Round 4	0	15	0.0%
Total	575	442	56.5%

Figure 2.2 shows the trend of each journal's rejection rate. In 2006-2016, the rejection rates in all of these journals have significantly increased. This result is consistent with the result for top economics journals reported by Card and DellaVigna [13]. Figure 2.3 shows the trend of each journal's number of submissions. Limited by the number of papers that can be published, the higher rejection rate appears to be caused by the increase in the number of submissions.

Figure 2.2: Rejection rate by journal

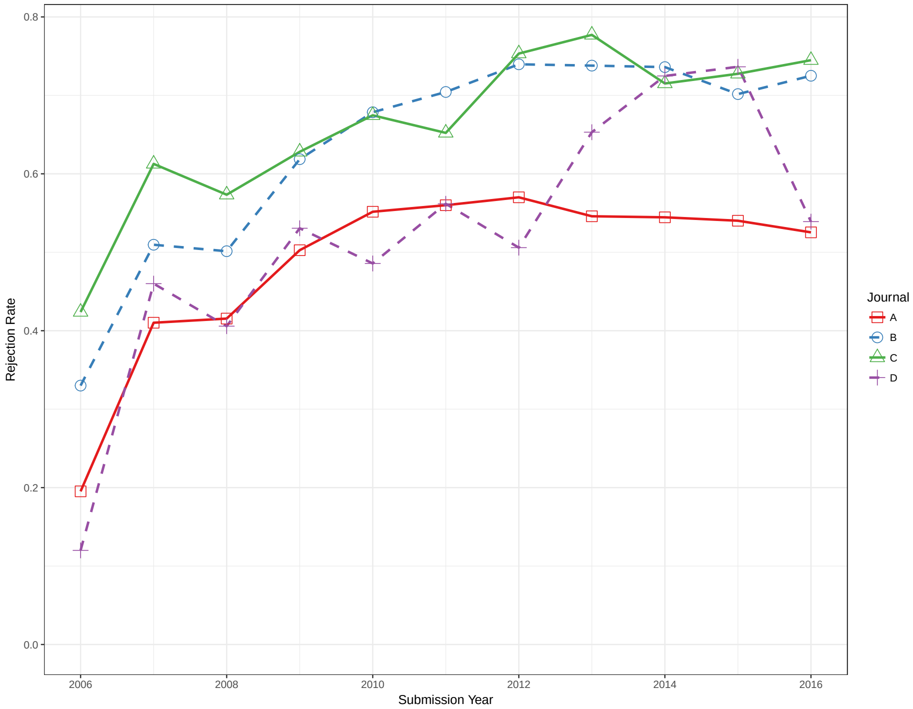
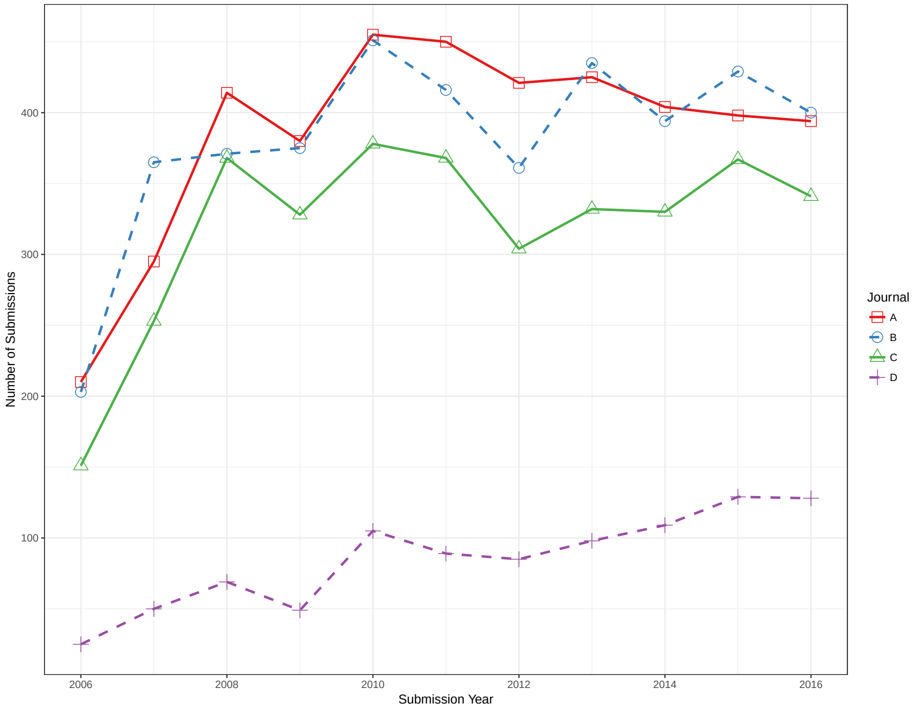


Figure 2.3: Number of submissions by journal



Referee evaluation

The referees evaluate the quality of the submitted paper and provide two referee evaluation scores: 1. Referee recommendation score; 2. Scientific contribution score. Table 2.4 presents the categories of referee recommendation score. The smallest score is 1 indicating "Definitely reject or Reject", and the largest score is 5 indicating "Accept as is". Table 2.5 presents the categories of scientific contribution score. The smallest score is 1 indicating "No scientific value or incremental contribution", and the largest score is 6 indicating "Potential seminal contribution".

Table 2.6 presents the distribution of referee recommendation and scientific contribution scores. In general, higher referee recommendation score is associated with higher scientific contribution score. However, some papers receive decent scientific contribution scores, while receiving relatively low recommendation scores. In addition, it is possible for a paper to receive recommendation scores that reflect disagreement by referees. These situations are expected to increase the difficulty of editorial decision making. In the estimation part, I create measures of referee recommendation scores and scientific evaluation scores to investigate the effect of referee recommendations on the decision making of editors.

Table 2.7 presents the statistics of referee recommendation scores, and Table 2.8 presents the statistics of scientific contribution scores. The average of referee recommendation scores is the highest in the second round, so is the average of scientific contribution scores. The relatively higher score in the second round may indicate that the quality of the revised papers that are sent for the second round of peer review is higher than that of the new submissions in the first round. It may also reveal that the referees appreciate the effort of the authors on revising the papers in response to referee reports so that higher referee evaluation scores are given. Besides, the lower average score of referee recommendation for papers in the third round of peer review may indicate that the papers that require two rounds of revisions have lower inner quality than the papers that only require one round of revision. To investigate the decision making of referees, data on referee reports and multiple versions of paper manuscripts would be necessary.

Table 2.4: Categories of referee recommendation score

Recommendation Score	Category
1	Definitely reject or reject
2	Return, major revisions
3	Accept with medium revisions
4	Accept with minor revisions
5	Accept as is

Table 2.5: Categories of scientific contribution score

Scientific Contribution Score	Category
1	No scientific value or incremental contribution
2	Only of minimal scientific value
3	Small scientific contribution
4	Average or typical scientific contribution
5	Strong scientific contribution
6	Potential seminal contribution

Table 2.6: Distribution of referee recommendation and scientific contribution scores

	Scientific Contribution Score					
Referee Recommendation Score	**1**	**2**	**3**	**4**	**5**	**6**
1	281	778	702	259	21	14
2	22	225	916	1,590	175	21
3	4	41	463	2,078	415	20
4	0	6	15	28	9	2
5	0	26	294	2,939	1,704	113

Table 2.7: Referee recommendation score by round

	Mean(Standard Deviation)				
	All Journals	**Journal A**	**Journal B**	**Journal C**	**Journal D**
Peer Review Round 1					
Mean Ref. Rec.	3.18	3.22	3.11	3.22	3.18
	(1.20)	(1.18)	(1.21)	(1.21)	(1.24)
S.D. of Ref. Rec.	0.99	0.98	0.99	1.01	0.95
	(0.83)	(0.83)	(0.83)	(0.84)	(0.85)
Maximum Ref. Rec.	3.90	3.94	3.83	3.94	3.87
	(1.36)	(1.31)	(1.41)	(1.35)	(1.40)
Minimum Ref. Rec.	2.48	2.52	2.40	2.51	2.47
	(1.36)	(1.36)	(1.32)	(1.40)	(1.40)
Peer Review Round 2					
Mean Ref. Rec.	3.64	3.84	3.39	3.14	3.69
	(1.42)	(1.36)	(1.42)	(1.41)	(1.50)
S.D. of Ref. Rec.	0.97	0.97	1.06	1.00	0.64
	(0.94)	(0.95)	(0.96)	(0.91)	(0.83)
Maximum Ref. Rec.	3.96	4.16	3.71	3.55	3.88
	(1.42)	(1.33)	(1.49)	(1.52)	(1.49)
Minimum Ref. Rec.	3.33	3.52	3.05	2.73	3.53
	(1.63)	(1.62)	(1.59)	(1.59)	(1.64)
Peer Review Round 3					
Mean Ref. Rec.	3.10	3.09	2.79	2.50	3.50
	(1.66)	(1.80)	(1.63)	(1.91)	(1.51)
S.D. of Ref. Rec.	0.53	0.59	0.71		0.00
	(0.63)	(0.70)			
Maximum Ref. Rec.	3.17	3.24	2.86	2.50	3.50
	(1.66)	(1.82)	(1.57)	(1.91)	(1.51)
Minimum Ref. Rec.	3.02	2.94	2.71	2.50	3.50
	(1.70)	(1.85)	(1.70)	(1.91)	(1.51)

Note: Papers in peer review round 4 are omitted since there were very few papers receiving referee reports in peer review round 4. Ref. is short for Referee. Rec. is short for Recommendation.

Table 2.8: Scientific contribution score by round

	Mean(Standard Deviation)				
	All Journals	**Journal A**	**Journal B**	**Journal C**	**Journal D**
Peer Review Round 1					
Mean Sci. Con.	3.78	3.83	3.69	3.80	3.81
	(0.77)	(0.73)	(0.80)	(0.79)	(0.80)
S.D. of Sci. Con.	0.62	0.61	0.63	0.65	0.63
	(0.57)	(0.57)	(0.56)	(0.60)	(0.56)
Maximum Sci. Con.	4.20	4.24	4.10	4.22	4.23
	(0.81)	(0.76)	(0.84)	(0.82)	(0.85)
Minimum Sci. Con.	3.35	3.41	3.26	3.37	3.38
	(0.98)	(0.96)	(0.99)	(1.01)	(0.99)
Peer Review Round 2					
Mean Sci. Con.	3.80	3.87	3.70	3.57	3.88
	(0.77)	(0.80)	(0.70)	(0.86)	(0.69)
S.D. of Sci. Con.	0.52	0.50	0.61	0.46	0.53
	(0.57)	(0.58)	(0.60)	(0.57)	(0.43)
Maximum Sci. Con.	3.92	3.96	3.85	3.69	4.03
	(0.77)	(0.77)	(0.73)	(0.87)	(0.75)
Minimum Sci. Con.	3.66	3.68	3.58	3.51	3.79
	(0.86)	(0.91)	(0.77)	(0.92)	(0.73)
Peer Review Round 3					
Mean Sci. Con.	3.58	3.20	3.75	3.67	3.89
	(1.03)	(1.48)	(0.96)	(0.58)	(0.33)
S.D. of Sci. Con.	0.42	0.24	0.71		0.71
	(0.39)	(0.41)			
Maximum Sci. Con.	3.69	3.36	3.80	3.67	4.00
	(1.04)	(1.50)	(0.84)	(0.58)	(0.47)
Minimum Sci. Con.	3.59	3.27	3.60	3.67	3.90
	(0.98)	(1.42)	(0.89)	(0.58)	(0.32)

Note: Papers in peer review round 4 are omitted since there were very few papers receiving referee reports in peer review round 4. Sci. is short for Scientific. Con. is short for Contribution.

Table 2.9 compares referee recommendation of accepted and rejected papers submitted in 2005-2017. The average of referee recommendation scores of accepted papers is 3.70, which is higher than the score for "Revise and resubmit" shown in Table 2.4. The scores of rejected papers and desk-rejected papers are 1.81 and 1.41 respectively, which are between the score for "Definitely reject or reject" and the score for "Weak revise and resubmit" shown in Table 2.4. In addition, the scientific evaluation of accepted papers is much higher than that of the rejected and desk-rejected papers as well. The relatively higher evaluation scores of accepted papers may indicate that the signals from referees are used by the editor to decide which paper to publish. It may also reflect that editor's evaluation of paper quality is generally consistent with referees' evaluation. Thus, the high-quality papers get higher evaluation scores, as well as being accepted by the editor.

Table 2.9: Referee recommendation and scientific evaluation: accepted v.s. rejected

Statistic	N	Mean	St. Dev.	25% Quantile	75% Quantile
Accepted					
Mean of Ref. Rec. Round 1	4,805	3.648	0.992	3.000	4.333
S.D. of Ref. Rec. Round 1	4,497	0.967	0.827	0.000	1.414
Max. of Ref. Rec. Round 1	4,805	4.338	1.045	3	5
Min. of Ref. Rec. Round 1	4,805	2.927	1.304	2	3
Mean of Sci. Con. Round 1	4,132	4.033	0.579	3.667	4.500
S.D. of Sci. Con. Round 1	3,666	0.543	0.532	0	0.707
Max. of Sci. Con. Round 1	4,132	4.398	0.632	4	5
Min. of Sci. Con. Round 1	4,132	3.650	0.809	3	4
Rejected					
Mean of Ref. Rec. Round 1	1,297	1.805	0.666	1.500	2.000
S.D. of Ref. Rec. Round 1	1,192	0.987	0.854	0.500	1.414
Max. of Ref. Rec. Round 1	1,297	2.605	1.393	2	3
Min. of Ref. Rec. Round 1	1,297	1.156	0.407	1	1
Mean of Sci. Con. Round 1	1,143	2.978	0.786	2.500	3.500
S.D. of Sci. Con. Round 1	953	0.887	0.637	0.577	1.414
Max. of Sci. Con. Round 1	1,143	3.554	0.947	3	4
Min. of Sci. Con. Round 1	1,143	2.407	0.947	2	3
Desk-rejected					
Mean of Ref. Rec. Round 1	102	1.407	0.935	1.000	1.000
S.D. of Ref. Rec. Round 1	7	0.909	1.058	0	1.414
Max. of Ref. Rec. Round 1	102	1.451	1.011	1	1
Min. of Ref. Rec. Round 1	102	1.363	0.920	1	1
Mean of Sci. Con. Round 1	63	2.730	1.288	1.500	4.000
S.D. of Sci. Con. Round 1	6	0.943	0.577	0.707	1.414
Max. of Sci. Con. Round 1	63	2.794	1.346	1.5	4
Min. of Sci. Con. Round 1	63	2.667	1.270	1.5	4

Note: Ref. is short for Referee. Rec. is short for Recommendation. Sci. is short for Scientific. Con. is short for Contribution.

Duration of the academic publishing process

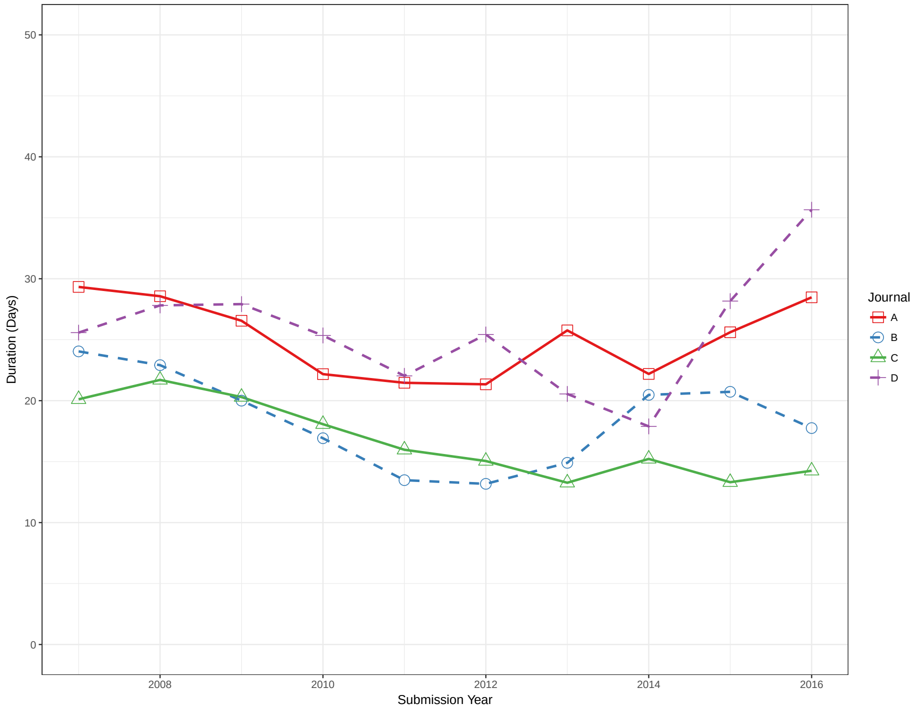
The trend of each journal's duration of the first round is shown in Figure 2.4. For desk-rejected papers, the duration of the first round only includes the duration of the desk review round. For non-desk-rejected papers, the duration of the first round includes the duration of the desk review round and the first round of peer review. Figure 2.4 shows that the average duration of the first round is relatively stable in 2007-2016. However, the relatively stable duration of the first round may be caused by the coexistence of the slowing down of the first round of peer review and the increasing number of papers that are quickly desk-rejected.

To further investigate the trends of the average duration of the first found, I present the average duration of the first round for non-desk-rejected papers and desk-rejected papers separately in Figure 2.5. The increasing duration of non-desk-rejected papers is seen in three of the four journals. Whereas, the average duration of desk-rejected papers is relatively short and frequently fluctuates.

Table 2.10 presents the average duration of each stage of the academic publishing process. The average duration of the first round is 23.29 days for all papers, 44.28 days for non-desk-rejected papers, and 2.95 days for desk-rejected papers. The non-negligible duration of the non-desk-rejected papers may either caused by the delays of editorial decision making or the time needed for referees to provide referee reports. Compared to the first round, the average durations of the second round and third round are not long, which may indicate that most of the papers in the second the third round only require minor revisions, and the editorial decision making for these papers is relatively easier.

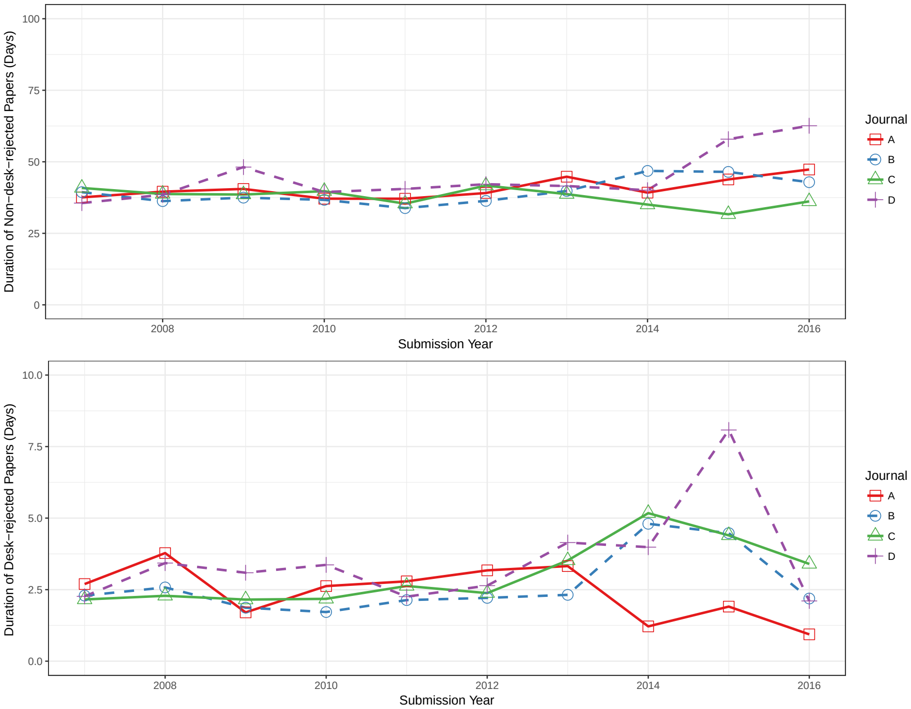
The non-negligible delays of the first round for non-desk-rejected papers raise the following question: Is it possible to train artificial intelligence algorithm to "read" papers and provide recommendations in order to reduce the duration of the first round for non-desk-rejected papers?

Figure 2.4: Duration of the first round of review



Note: This figure presents the durations for papers submitted between 2007 and 2016. The records of the durations for papers submitted before 2007 were not available.

Figure 2.5: Duration of the first round of review: Non-desk-rejected papers v.s. desk-rejected papers



Note: This figure presents the durations for papers submitted between 2007 and 2016. The records of the durations for papers submitted before 2007 were not available.

Table 2.10: Duration of the academic publishing process

Statistic	N	Mean	St. Dev.	25% Quantile	75% Quantile
All Journals					
Round 1 (Days)	13,495	23.29	30.48	1.00	38.96
Round 1 Non-desk-rejected (Days)	6,641	44.28	31.24	29.00	51.00
Round 1 Desk-rejected (Days)	6,854	2.95	6.57	0.00	3.00
Round 2 (Days)	4,975	8.28	18.37	1.00	7.00
Round 3 (Days)	514	6.31	13.09	0.00	6.00
Journal A					
Round 1 (Days)	4,412	28.36	35.42	1.00	42.96
Round 1 Non-desk-rejected (Days)	2,661	45.32	36.37	29.00	52.00
Round 1 Desk-rejected (Days)	1,751	2.58	6.99	0.00	2.00
Round 2 (Days)	2,059	9.67	21.24	1.00	9.00
Round 3 (Days)	264	6.78	15.78	0.00	6.00
Journal B					
Round 1 (Days)	4,391	20.92	28.62	1.00	36.50
Round 1 Non-desk-rejected (Days)	1,935	43.91	29.44	28.00	51.00
Round 1 Desk-rejected (Days)	2,456	2.80	6.09	0.00	3.00
Round 2 (Days)	1,375	7.85	16.48	1.00	8.00
Round 3 (Days)	140	4.64	7.86	0.00	6.00
Journal C					
Round 1 (Days)	3,676	18.97	25.49	1.00	33.00
Round 1 Non-desk-rejected (Days)	1,501	41.76	25.55	28.00	47.04
Round 1 Desk-rejected (Days)	2,175	3.24	6.45	0.00	3.00
Round 2 (Days)	1,136	6.13	16.02	0.00	5.00
Round 3 (Days)	54	4.95	9.82	0.00	3.00
Journal D					
Round 1 (Days)	1,016	27.15	28.00	1.00	44.00
Round 1 Non-desk-rejected (Days)	544	47.42	22.97	32.00	57.00
Round 1 Desk-rejected (Days)	472	3.80	7.80	0.00	3.00
Round 2 (Days)	405	8.68	13.80	1.00	10.00
Round 3 (Days)	56	9.61	11.73	1.00	13.00

Note: Papers in peer review round 4 are omitted since there were very few papers having four rounds of peer review.

Paper citations

Table 2.11 presents the statistics of the citation counts of papers submitted in 2006-2010. I choose papers published in 2006-2010 to ensure that the papers used for the statistics have at least 7 years to accumulate citations.[6] The average number of citations of accepted papers is much higher than that of rejected and desk-rejected papers, and the difference between rejected and desk-rejected papers is only marginal.

[6] For most of the papers used in this study, the number of citations becomes stable after 7 years of publication.

The higher number of citations of accepted papers would reflect the performance of editors in identifying high-quality papers, if high-quality papers are more likely to be highly cited. However, by looking at the 25% quantile and the 75% quantile cutoffs of each group, we can find that there exist a few rejected papers that have more citations than a fraction of accepted papers.

Figure 2.6 presents the citation distributions of papers published by the four journals in 2006-2010. The distributions show that most of the papers get less than 30 citations, while some seminal papers get more than 400 citations. This finding is consistent with the situation for the papers in the top 5 economics journals studied in the first chapter. Figure 2.7 compares the citation distributions of accepted papers and rejected papers of each journal. The overlap between the distribution of accepted papers and rejected papers of each journal confirms the existence of rejected papers that have more citations than a fraction of accepted papers. The overlap may indicate that it is hard to use the information available at the time of submission to identify papers that will be highly cited in the long run. It may also indicate that there are factors other than expected citations that drive editorial decision making.

Table 2.11: Paper citations by journal

Statistic	N	Mean	St. Dev.	25% Quantile	75% Quantile
All Journals					
Accepted	2,267	28.21	30.27	12	35
Rejected	437	15.13	19.34	4	19
Desk-rejected	1,629	16.76	20.92	4	21
Journal A					
Accepted	879	27.50	30.06	11	34
Rejected	157	12.78	12.95	4	16
Desk-rejected	410	13.96	20.70	3	17
Journal B					
Accepted	707	24.57	25.98	10	32
Rejected	131	13.34	16.99	3	15
Desk-rejected	583	14.78	19.36	4	19
Journal C					
Accepted	531	35.14	36.11	15	43.5
Rejected	128	19.40	26.45	5	23.2
Desk-rejected	565	21.13	21.96	7	28
Journal D					
Accepted	150	25.08	23.02	12	33
Rejected	21	17.81	18.41	5	26
Desk-rejected	71	14.45	20.91	2	19

Note: The statistics are based on the citation counts of papers submitted in 2006-2010.

Figure 2.6: Citation distribution of accepted papers as of the end of November 2017

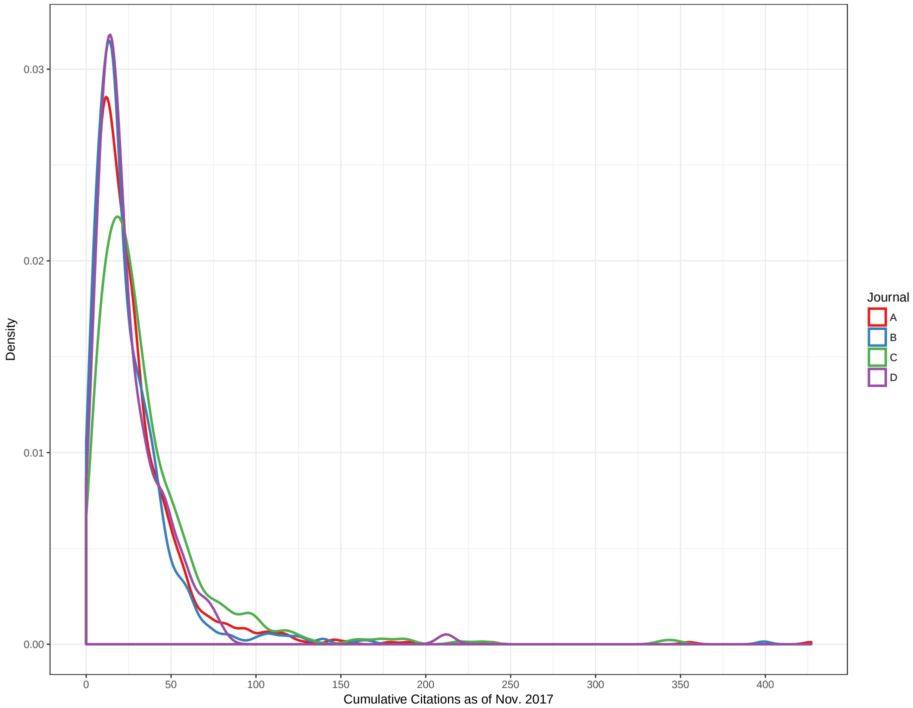
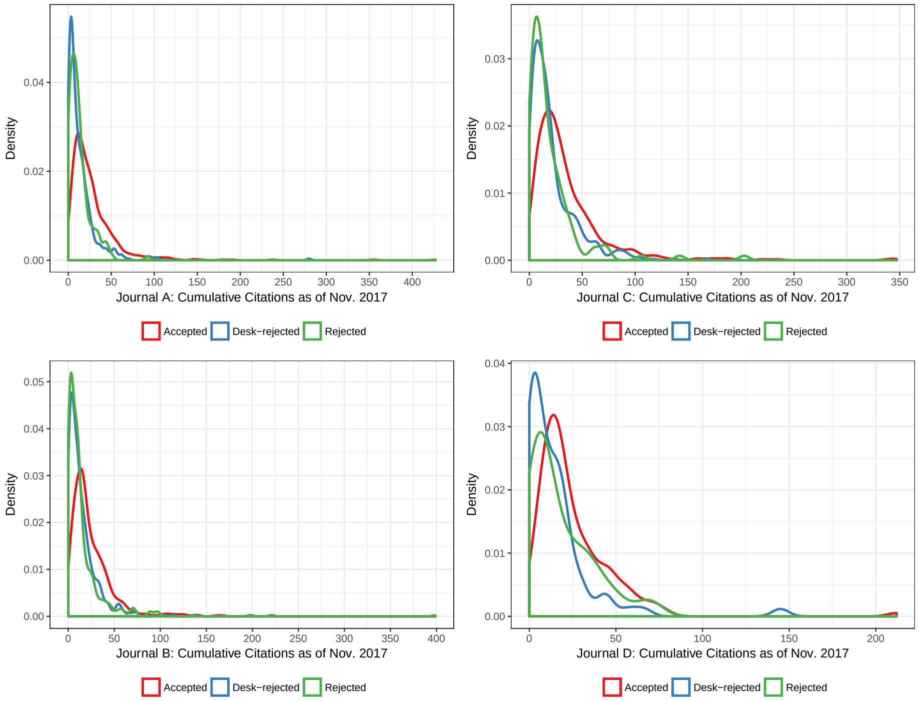


Figure 2.7: Citation distribution of papers as of the end of November 2017: accepted v.s. rejected



Paper and author information

Table 2.12 compares the paper and author information of accepted papers and rejected papers that were submitted in 2005-2017. Regarding author information, the number of author's submissions of accepted papers is marginally higher than that of rejected papers and is much higher than that of desk-rejected papers. It may indicate that authors with longer submission records are easier to pass the desk review. In addition, the average rejection rate of the authors of accepted papers is 22.3%, whereas the average rejection rates of the authors of rejected papers and desk-rejected papers are 38.5% and 54.5% respectively. The lower average rejection rate of the authors of accepted papers may indicate that authors whose papers were often rejected in history are harder to get a new paper published.

Regarding the variables measuring research fields and topic words, the frequencies of the appearance of research fields and topic words for accepted papers and rejected papers are similar, and the frequency for desk-rejected papers is relatively lower. The difference in the frequency of word appearance between non-desk-rejected papers and desk-rejected papers may indicate that desk-rejected papers are either too narrow or too far away from the scope of the target journal.

Table 2.12: Paper and author information

Statistic	N	Mean	St. Dev.	25% Quantile	75% Quantile
Accepted					
Author Num. Submission	2,332	4.660	6.128	1	5.2
Author Rejection Rate	2,332	0.223	0.329	0	0.333
Keyword Field A Dummy	4,829	0.235	0.424	0	0
Keyword Field B Dummy	4,829	0.222	0.416	0	0
Keyword Field C Dummy	4,829	0.199	0.399	0	0
Keyword Field D Dummy	4,829	0.187	0.390	0	0
Abstract Field A Intensity	4,829	0.031	0.016	0.020	0.041
Abstract Field B Intensity	4,829	0.031	0.019	0.017	0.042
Abstract Field C Intensity	4,829	0.025	0.016	0.013	0.032
Abstract Field D Intensity	4,829	0.025	0.017	0.013	0.034
Abstract Popular Topics	4,829	0.023	0.010	0.015	0.030
Abstract Popular Two-word Pairs	4,829	0.009	0.010	0.002	0.011
Abstract Popular Three-word Pairs	4,829	0.001	0.002	0	0
Abstract Popular Four-word Pairs	4,829	0.00003	0.0004	0	0
Rejected					
Author Num. Submission	535	4.508	6.277	1	5
Author Rejection Rate	535	0.385	0.386	0	0.667
Keyword Field A Dummy	1,309	0.230	0.421	0	0
Keyword Field B Dummy	1,309	0.235	0.424	0	0
Keyword Field C Dummy	1,309	0.205	0.404	0	0
Keyword Field D Dummy	1,309	0.199	0.399	0	0
Abstract Field A Intensity	1,309	0.032	0.016	0.020	0.042
Abstract Field B Intensity	1,309	0.031	0.018	0.018	0.043
Abstract Field C Intensity	1,309	0.026	0.017	0.014	0.034
Abstract Field D Intensity	1,309	0.025	0.017	0.013	0.035
Abstract Popular Topics	1,309	0.023	0.011	0.015	0.030
Abstract Popular Two-word Pairs	1,309	0.008	0.010	0.002	0.011
Abstract Popular Three-word Pairs	1,309	0.001	0.003	0	0
Abstract Popular Four-word Pairs	1,309	0.00004	0.001	0	0
Desk-rejected					
Author Num. Submission	1,824	3.651	5.548	1	4
Author Rejection Rate	1,824	0.545	0.414	0	1.000
Keyword Field A Dummy	6,854	0.168	0.374	0	0
Keyword Field B Dummy	6,854	0.197	0.398	0	0
Keyword Field C Dummy	6,854	0.184	0.388	0	0
Keyword Field D Dummy	6,854	0.136	0.342	0	0
Abstract Field A Intensity	6,854	0.028	0.016	0.016	0.038
Abstract Field B Intensity	6,854	0.031	0.018	0.018	0.041
Abstract Field C Intensity	6,854	0.027	0.019	0.014	0.036
Abstract Field D Intensity	6,854	0.021	0.016	0.010	0.030
Abstract Popular Topics	6,854	0.021	0.010	0.013	0.026
Abstract Popular Two-word Pairs	6,854	0.008	0.010	0	0.009
Abstract Popular Three-word Pairs	6,854	0.001	0.003	0	0
Abstract Popular Four-word Pairs	6,854	0.0001	0.005	0	0

In the next subsection, I use logistic regressions to estimate the effect of paper and author information, and referee recommendation on editorial decision making in the desk review round and the first round of peer review.

2.5.2 Estimation results

Editorial decision making

Table 2.13 presents the results for decision making in the desk review round. Papers that cover more popular topics are more likely to pass desk review. One explanation could be that popular topics attract the interest of a broad range of researchers. Another explanation could be popular topics are areas where high-quality work is being done. In addition, the papers that do not cover enough popular topics are either too narrow or too far away from the scope of the target journal, which reduces their chance of being published.

The results also suggest that papers by authors with higher numbers of submissions and lower rejection rates are more likely to pass the desk review round. One possible explanation can be that the authors with these characteristics have decent records in the journal database, which makes it easier for editors to evaluate the quality of their work. Another possible explanation can be that the authors with these characteristics could be better researchers, and the "author effect" is a proxy for paper quality.

Table 2.13: Decision making in the desk review round

	All Journals				
	Not Rejected in the Desk Review Round				
	(1)	(2)	(3)	(4)	(5)
Author Num. Submission	0.02***				0.02***
	(0.006)				(0.006)
Author Rejection Rate	−1.62***				−1.57***
	(0.08)				(0.09)
Abstract Field A Intensity		31.06***		24.38***	13.54***
		(2.07)		(2.13)	(3.62)
Abstract Field B Intensity		−26.40***		−36.76***	−30.70***
		(2.14)		(2.26)	(3.90)
Abstract Field C Intensity		−18.10***		−28.65***	−17.81***
		(1.71)		(1.90)	(3.26)
Abstract Field D Intensity		31.23***		21.60***	20.63***
		(1.98)		(2.11)	(3.81)
Abstract Popular Topics			38.92***	58.05***	43.32***
			(2.40)	(3.48)	(6.02)
Abstract Popular Two-word Pairs			−3.64	6.91**	3.06
			(2.89)	(3.20)	(5.41)
Abstract Popular Three-word Pairs			−49.38***	−12.39	5.58
			(10.33)	(12.22)	(19.22)
Abstract Popular Four-word Pairs			6.00	−50.91	−68.89*
			(13.60)	(31.97)	(36.40)
Control: Publication Information	Yes	Yes	Yes	Yes	Yes
Observations	4,899	13,517	13,517	13,517	4,899
Log Likelihood	-2,849.32	-8,631.53	-8,789.13	-8,444.65	-2,750.74

Note: *p<0.1; **p<0.05; ***p<0.01.

In the first round of peer review, the editor chooses between "Reject", "Revision request", and "Accept". Table 2.14 presents the results for decision making in the first round of peer review. The coefficients of the variables measuring author information are not significant in the decision making in the first round of peer review. One possible explanation could be that the recommendation from referees provide additional information about the quality of the paper, which makes author information not as important as it is in the desk review round. However, the coefficient of "Abstract Popular Topics" is positive and significant in all of these regressions, which suggests the positive effect of popular topic words on receiving a revision request at least.

The regression results also suggest that the papers with higher referee recommendation scores and scientific contribution scores are more likely to receive a revision request at least. The positive effect of referee recommendation indicates that editors do rely on their evaluations of the quality of the paper that is reviewed. In addition, the papers with a higher standard deviation of referee recommendation scores are less likely to pass the first round of peer review. One explanation could be that the papers with a higher standard deviation of referee recommendation scores are more "risky", and some of them may turn out to be papers that are not scientifically impactful. Admittedly, the mean and standard deviation of referee recommendation scores and scientific contribution scores are not perfect measures of the signal from referees, individual coefficients of the recommendation scores and scientific contribution scores should be interpreted with caution.

Table 2.15 presents the results of the regressions using the fractions of referee recommendation and scientific evaluation scores as measures of signals from referees. The fractions of recommendations are also used by Card and DellaVigna [14]. The results in Table 2.15 are consistent with the results in Table 2.14 where the mean and standard deviation of referee recommendations are used. The coefficients of the variables measuring author information are not significant, indicating the author effect is not significant in the first round of peer review. The coefficients of the shares of the positive referee recommendations are generally larger than the coefficients of the shares of the negative referee recommendations, which suggests that a higher share of positive referee recommendations is associated with a higher probability of passing the first round of peer review.

The regressions using the fractions are less restrictive and result in a better fit even controlling for the larger number of parameters estimated by comparing AIC "per observation". However, the results in Table 2.14 and Table 2.15 suggest that the mean and standard deviation is a convenient measure of referee recommendations that does not distort results too much. Thus, I use the mean and standard deviation as the measure of referee recommendations to estimate the effect of referee recommendations on the duration of the first round of peer review and paper citations.

Table 2.14: Decision making in the first round of peer review

	All Journals				
	Decision Making in Peer Review Round 1				
	(1)	(2)	(3)	(4)	(5)
Author Num. Submission				0.004	0.0001
				(0.01)	(0.01)
Author Rejection Rate				−0.34	−0.33
				(0.24)	(0.24)
Abstract Field A Intensity					17.34***
					(0.0002)
Abstract Field B Intensity					−12.64***
					(0.003)
Abstract Field C Intensity					−0.25***
					(0.004)
Abstract Field D Intensity					2.68***
					(0.001)
Abstract Popular Topics					3.43***
					(0.001)
Abstract Popular Two-word Pairs					−2.22***
					(0.001)
Abstract Popular Three-word Pairs					37.30***
					(0.0001)
Abstract Popular Four-word Pairs					−43.08***
					(0.00004)
Mean of Ref. Rec. Round 1	2.66***	4.29***	4.38***	4.25***	4.31***
	(0.11)	(0.09)	(0.05)	(0.06)	(0.07)
S.D. of Ref. Rec. Round 1		−1.92***	−2.09***	−1.94***	−1.95***
		(0.07)	(0.06)	(0.08)	(0.09)
Mean of Sci. Eval. Round 1			0.88***	0.81***	0.82***
			(0.10)	(0.15)	(0.15)
Cutoff1	278.11***	287.47***	396.16***	454.00***	452.03***
	(0.04)	(0.04)	(0.02)	(0.01)	(0.01)
Cutoff2	290.91***	306.76***	417.57***	474.86***	473.06***
	(0.37)	(0.44)	(0.41)	(0.55)	(0.55)
Control: Publication Information	Yes	Yes	Yes	Yes	Yes
Observations	6,623	6,187	5,451	2,542	2,542
AIC	3349.99	2523.43	1932.28	879.09	887.88

Note: *p<0.1; **p<0.05; ***p<0.01. "Cutoff1" denotes the cutoff between "Reject" and "Revision request", and "Cutoff2" denotes the cutoff between "Revision request" and "Accept".

Table 2.15: Decision making in the first round of peer review

	All Journals				
	Decision Making in Peer Review Round 1				
	(1)	(2)	(3)	(4)	(5)
Author Num. Submission				0.01	0.003
				(0.01)	(0.01)
Author Rejection Rate				−0.33	−0.31
				(0.25)	(0.25)
Abstract Field A Intensity					21.63***
					(0.002)
Abstract Field B Intensity					−18.36***
					(0.003)
Abstract Field C Intensity					−4.15***
					(0.005)
Abstract Field D Intensity					8.64***
					(0.002)
Abstract Popular Topics					9.77***
					(0.002)
Abstract Popular Two-word Pairs					−4.19***
					(0.002)
Abstract Popular Three-word Pairs					47.32***
					(0.0003)
Abstract Popular Four-word Pairs					−83.14***
					(0.0001)
Ref. Rec. Rejection Share	−6.04***		−5.42***	−3.42***	−3.37***
	(0.07)		(0.07)	(0.12)	(0.11)
Ref. Rec. Major Rev. Share	1.81***		1.89***	3.90***	4.07***
	(0.13)		(0.13)	(0.21)	(0.21)
Ref. Rec. Medium Rev. Share	5.65***		5.53***	7.48***	7.79***
	(0.16)		(0.15)	(0.24)	(0.24)
Ref. Rec. Accept Share	6.65***		6.29***	8.40***	8.78***
	(0.16)		(0.16)	(0.24)	(0.24)
Sci. Con. No Value Share		−4.73***	−1.01***	−0.27***	−0.40***
		(0.01)	(0.01)	(0.02)	(0.02)
Sci. Con. Minimal Value Share		−4.05***	−1.31***	−0.96***	−1.06***
		(0.10)	(0.15)	(0.21)	(0.22)
Sci. Con. Small Value Share		−1.48***	−0.24*	−0.08	−0.14
		(0.10)	(0.14)	(0.23)	(0.23)
Sci. Con. Average Value Share		1.31***	0.87***	1.37***	1.31***
		(0.09)	(0.14)	(0.22)	(0.22)
Sci. Con. Strong Value Share		2.96***	1.42***	2.00***	1.87***
		(0.03)	(0.07)	(0.11)	(0.11)
Cutoff1	307.18***	139.34***	335.56***	357.13***	348.15***
	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)
Cutoff2	319.73***	149.02***	348.50***	370.20***	361.41***
	(0.37)	(0.36)	(0.38)	(0.51)	(0.51)
Control: Publication Information	Yes	Yes	Yes	Yes	Yes
Observations	5,742	5,742	5,742	2,666	2,666
AIC	1984.73	3701.86	1917.77	857.09	862.07

Note: *p<0.1; **p<0.05; ***p<0.01. "Cutoff1" denotes the cutoff between "Reject" and "Revision request", and "Cutoff2" denotes the cutoff between "Revision request" and "Accept".

Duration of the academic publishing process

Table 2.16 presents the results for the paper and author effect on the duration of the first round for non-desk-rejected papers. Most of the coefficients are not significant. The effect of author's rejection rate on the duration is positive and significant, but the coefficient is only significant at 10% and 5% level.

Table 2.17 presents the effect of referee recommendation on the duration of the first round for non-desk-rejected papers. The papers with higher referee recommendation scores and a lower standard deviation of the scores have shorter durations of the first round of review. The shorter durations for these papers may indicate that the quality of these papers is more "clear" to editors and referees, which makes it easier for editors to make editorial decision and referees to provide referee reports. It would be interesting for future study to investigate whether it is the editor or the referee that causes delays in the peer review round, as well as the reasons for the delays.

Table 2.16: Paper and author effect on the duration of the first round (non-desk-rejected)

	All Journals				
	Duration (Days)				
	(1)	(2)	(3)	(4)	(5)
Author Num. Submission	−0.05				−0.06
	(0.06)				(0.06)
Author Rejection Rate	3.06*				3.21**
	(1.58)				(1.59)
Abstract Field A Intensity		−4.70		−30.78	−59.25
		(38.63)		(39.36)	(50.34)
Abstract Field B Intensity		−3.57		−30.83	−79.66
		(37.37)		(37.57)	(49.18)
Abstract Field C Intensity		24.40		−3.72	20.76
		(34.62)		(37.59)	(50.60)
Abstract Field D Intensity		−85.25**		−111.05***	−59.91
		(39.30)		(42.44)	(53.57)
Abstract Popular Topics			2.31	190.91***	241.92***
			(49.94)	(61.53)	(81.38)
Abstract Popular Two-word Pairs			−53.61	−24.69	11.34
			(49.27)	(51.93)	(72.76)
Abstract Popular Three-word Pairs			307.10	138.95	−34.30
			(210.91)	(212.83)	(284.29)
Abstract Popular Four-word Pairs			−179.29	44.00	−576.01
			(1,009.49)	(1,008.66)	(856.90)
Control: Publication Information	Yes	Yes	Yes	Yes	Yes
Observations	3,064	6,641	6,641	6,641	3,064
R^2	0.03	0.03	0.03	0.03	0.04
Adjusted R^2	0.03	0.03	0.03	0.03	0.03

Note: *p<0.1; **p<0.05; ***p<0.01. For the non-desk-rejected papers, the duration of the first round includes the duration of the desk review and the first round of peer review. Robust standard errors in parentheses.

Table 2.17: Referee effect on the duration of the first round (non-desk-rejected)

	All Journals				
	Duration (Days)				
	(1)	(2)	(3)	(4)	(5)
Mean of Ref. Rec. Round 1	-2.05^{***}	-2.08^{***}	-1.63^{***}	-1.29^{***}	-1.29^{***}
	(0.30)	(0.29)	(0.25)	(0.34)	(0.34)
S.D. of Ref. Rec. Round 1		0.77^{*}	1.20^{***}	1.27^{***}	1.25^{***}
		(0.44)	(0.38)	(0.40)	(0.40)
Mean of Sci. Eval. Round 1			1.15^{***}	0.64	0.65
			(0.38)	(0.55)	(0.55)
Control: Publication Information	Yes	Yes	Yes	Yes	Yes
Control: Author Information				Yes	Yes
Control: Field Information					Yes
Control: Topic Information					Yes
Observations	6,601	6,169	5,443	2,540	2,540
R^2	0.03	0.02	0.03	0.04	0.04
Adjusted R^2	0.03	0.02	0.03	0.03	0.04

Note: *p<0.1; **p<0.05; ***p<0.01. For the non-desk-rejected papers, the duration of the first round includes the duration of the desk review and the first round of peer review. "Mean of Ref. Rec. Round 1" denotes the mean of referee recommendation scores in round 1. "S.D. of Ref. Rec. Round 1" denotes the standard deviation of referee recommendation scores in round 1. "Mean of Sci. Eval. Round 1" denotes the mean of scientific contribution scores in round 1. Robust standard errors in parentheses.

Paper citations

Table 2.18 presents the results for assessing the effect of editorial decision making on paper citations. The desk-rejected papers are used as the reference group. Compared to desk-rejected papers, accepted papers on average have 10 more citations after 7-12 years of publication. One explanation is that the number of citations is associated with the scientific impact of the paper, and paper's scientific impact is one the criteria when editors decide whether to publish a paper or not. Thus, the papers with higher scientific impact get published and accumulate more citations than rejected ones. Another explanation is that the journals used in this study may have performed better in advertising publications than the journals where rejected papers are subsequently published.

Table 2.19 presents the results for assessing the effect of referee evaluations on paper citations. The results suggest that papers with higher referee recommendation scores on average get higher citation counts, which may indicate that the papers recommended by the referees generally have higher quality and are more likely to be highly cited. Table 2.20 presents the results for paper and author effect on paper citations. The papers covering more popular research topics are more likely to be highly cited. However, the effect of author information is not significant.

Table 2.18: Assessing the effect of editorial decision making on paper citations

	All Journals				
	Cumulative Citations				
	(1)	(2)	(3)	(4)	(5)
Accepted	10.32***	11.35***	10.74***	10.16***	10.89***
	(0.79)	(1.39)	(0.80)	(0.81)	(1.33)
Rejected	−2.02*	−1.17	−2.18**	−2.70**	−2.48
	(1.08)	(2.03)	(1.09)	(1.08)	(1.91)
Revision Request	−0.37	−1.10	−0.18	−0.29	−0.95
	(1.64)	(2.38)	(1.59)	(1.55)	(2.11)
Control: Publication Information	Yes	Yes	Yes	Yes	Yes
Control: Author Information		Yes			Yes
Control: Field Information			Yes	Yes	Yes
Control: Topic Information				Yes	Yes
Observations	4,483	1,718	4,483	4,483	1,718
R^2	0.13	0.17	0.15	0.16	0.21
Adjusted R^2	0.13	0.17	0.15	0.16	0.20

Note: *p<0.1; **p<0.05; ***p<0.01. The cumulative citations denote the number of citations as of the end of November 2017. Robust standard errors in parentheses.

Table 2.19: Assessing the effect of referee evaluations on paper citations

	All Journals				
	Cumulative Citations				
	(1)	(2)	(3)	(4)	(5)
Mean of Ref. Rec. Round 1	3.08***	3.08***	1.88***	2.22**	2.47***
	(0.41)	(0.44)	(0.53)	(0.87)	(0.84)
S.D. of Ref. Rec. Round 1		−0.87	−1.03*	−0.80	−0.86
		(0.60)	(0.62)	(1.01)	(1.00)
Mean of Sci. Eval. Round 1			3.20***	1.55	1.69
			(1.05)	(1.49)	(1.45)
Control: Publication Information	Yes	Yes	Yes	Yes	Yes
Control: Author Information				Yes	Yes
Control: Field Information					Yes
Control: Topic Information					Yes
Observations	2,844	2,589	1,928	911	911
R^2	0.12	0.13	0.16	0.21	0.26
Adjusted R^2	0.12	0.12	0.16	0.21	0.25

Note: *p<0.1; **p<0.05; ***p<0.01. The cumulative citations denote the number of citations as of the end of November 2017. "Mean of Ref. Rec. Round 1" denotes the mean of referee recommendation scores in round 1. "S.D. of Ref. Rec. Round 1" denotes the standard deviation of referee recommendation scores in round 1. "Mean of Sci. Eval. Round 1" denotes the mean of scientific contribution scores in round 1. Robust standard errors in parentheses.

Table 2.20: Assessing paper and author effect on paper citations

	All Journals				
	Cumulative Citations				
	(1)	(2)	(3)	(4)	(5)
Author Num. Submission	0.12				−0.006
	(0.15)				(0.15)
Author Rejection Rate	−1.56				−2.02
	(2.40)				(2.37)
Abstract Field A Intensity		113.53***		48.22	−38.60
		(39.90)		(40.92)	(68.81)
Abstract Field B Intensity		−92.59**		−156.59***	−156.83**
		(40.69)		(42.17)	(73.78)
Abstract Field C Intensity		223.53***		101.43***	139.39**
		(38.48)		(38.10)	(63.83)
Abstract Field D Intensity		24.65		−40.38	29.59
		(37.86)		(39.96)	(79.23)
Abstract Popular Topics			241.80***	373.89***	307.13**
			(51.68)	(71.44)	(121.26)
Abstract Popular Two-word Pairs			284.43***	232.56***	300.27***
			(65.20)	(65.93)	(116.42)
Abstract Popular Three-word Pairs			−2.40	−98.89	345.92
			(160.93)	(155.78)	(376.87)
Abstract Popular Four-word Pairs			16.46	56.65	−619.21*
			(73.41)	(67.55)	(351.44)
Control: Publication Information	Yes	Yes	Yes	Yes	Yes
Observations	1,718	4,483	4,483	4,483	1,718
R^2	0.14	0.11	0.12	0.13	0.18
Adjusted R^2	0.14	0.11	0.12	0.12	0.17

Note: *p<0.1; **p<0.05; ***p<0.01. The cumulative citations denote the number of citations as of the end of November 2017. Robust standard errors in parentheses.

2.5.3 Prediction results

The 4,322 papers that were submitted to these journals in 2006-2010 and had identifiable citation counts are used in the citation prediction experiment. I sort these papers into 10 parts based on the deciles of papers' cumulative citations as of the end of November 2017 and create variable D_i, the citation decile of paper i, which has values from 1 to 10. Then, I randomly select 70% of the papers as training samples and 30% of the papers as testing samples and use prediction models to predict paper's D_i with paper information available at the time of submission. To compare models' out-of-sample prediction performance, I use Mean Squared Error (MSE):

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(D_i - \hat{D}_i)^2 \tag{2.17}$$

where D_i is the observed citation decile of paper i, and $\hat{D}_i$ is the predicted citation decile of paper i.

I use the following models which have the vectors of variables measuring paper information added sequentially: Model (1) only includes variables measuring publication information (including submission year, journal information, and indicator for being a review), Model (2) adds low dimensional measures of paper fields and topics, and Model (3) to Model (6) add high dimensional measures of the appearance of popular topic words and word pairs. Model (1) has 3 variables, Model (2) has 27 variables, Model (3) has 1,091 variables, Model (4) has 2,153 variables, Model (5) has 3,163 variables, and Model (6) has 4,065 variables.

The linear model fitted by Ordinary Least Squares (OLS) is used as the baseline model, and regression shrinkage methods (Post-Lasso, Lasso, Ridge, Elastic Net), Random Forest, and Gradient Boosted Trees are compared with each other. The values of the key parameters of the machine learning methods used in this test are the same as that of the methods used in the first chapter and are presented in Table C.11.

Table 2.21 presents the out-of-sample prediction results. It shows that the MSE of prediction methods, except for OLS, generally decreases after adding more predictors (from Model (1) to Model (6)). The model that uses Random Forest method, measures of publication information, measures of paper fields and topics, and high dimensional measures of the appearance of popular topic words gives the smallest MSE. In Section 2.5.4, I use the preferred model (Model (3) + Random Forest) as the prediction model to investigate the possibility of adding artificial intelligence into the academic publishing process.

Table 2.21: Out-of-sample prediction: Papers submitted to the four journals 2006-2010

	Mean Squared Error (S.D. of Mean Squared Error)					
Method	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Ordinary Least Squares (OLS)	8.60	9.22	11.92	13.42	13.72	13.72
	(0.14)	(0.17)	(0.25)	(0.28)	(0.29)	(0.29)
Plug-in Post-Lasso	8.02	7.37	7.34	7.46	7.42	7.44
	(0.20)	(0.20)	(0.20)	(0.21)	(0.20)	(0.20)
Plug-in Lasso	8.06	7.51	7.58	7.62	7.64	7.65
	(0.19)	(0.19)	(0.18)	(0.18)	(0.19)	(0.19)
Cross-Validation Lasso	8.07	7.35	7.27	7.32	7.28	7.28
	(0.19)	(0.19)	(0.19)	(0.19)	(0.19)	(0.19)
Cross-Validation Ridge	8.09	7.34	7.43	7.41	7.36	7.39
	(0.19)	(0.19)	(0.19)	(0.19)	(0.19)	(0.19)
Cross-Validation Elastic Net	8.15	7.34	7.28	7.20	7.26	7.34
	(0.19)	(0.19)	(0.19)	(0.19)	(0.19)	(0.19)
Random Forest	7.98	7.50	**7.11**	7.14	7.19	7.25
	(0.20)	(0.20)	**(0.19)**	(0.19)	(0.19)	(0.19)
Gradient Boosted Trees	7.95	7.33	7.15	7.16	7.16	7.16
	(0.20)	(0.19)	(0.19)	(0.19)	(0.19)	(0.19)

2.5.4 Adding artificial intelligence into the academic publishing process

In this subsection, I investigate the possibility of using artificial intelligence to help editors identify the papers that are more likely to be highly cited with the information available at the time of submission. Admittedly, there are other factors that affect editors' decision on which paper to publish. However, the artificial intelligence based prediction model may be able to help editor decide which paper to publish when the number of publishable papers is higher than the number of papers that can be published in an issue. The papers in the testing group of the out-of-sample prediction test of Section 2.5.3 are used to test the performance of the alternative academic publishing processes that utilize different levels of human intelligence and artificial intelligence. There are 1,345 papers in the testing group of out-of-sample prediction test, and 701 of them were published papers. Specifically, I compare four alternative academic publishing processes:

(1). **Alternative 1:** Academic publishing process without reviewing. The Alternative 1 does not use any human intelligence or artificial intelligence, and all submissions are accepted.

(2). **Alternative 2:** Human intelligence based academic publishing process. The Alternative 2, which is the current academic publishing process, only uses the intelligence of editors and referees in paper selection. Among the 1,345 papers in the testing group, 701 papers were selected by editors for publication.

(3). **Alternative 3:** Artificial intelligence based academic publishing process. The Alternative 3 uses the preferred prediction model to decide which paper to publish. It uses artificial intelligence to "read" each paper and publishes the top 700 papers that have the highest number of predicted cumulative citations.

(4). **Alternative 4:** Academic publishing process integrating human intelligence and artificial intelligence. The Alternative 4 publishes the papers that are in the overlap of the selection of human editors and the artificial intelligence algorithm. It trains artificial intelligence to "read" each paper and identifies the top 700 papers that have the highest number of predicted cumulative citations. Then, among the papers selected by the prediction model, if a paper is one of the 701 papers selected by the editor, it is marked as "eventually published". In this process, only the papers that are both selected by the editor and the prediction model get published, and the artificial intelligence is used as a screening tool to identify the papers that are expected to be highly cited among the publishable papers. After the selection of the editor and the artificial intelligence based prediction model, 406 papers were "eventually published".

Table 2.22 presents the comparison of the cumulative citations of the papers selected by each alternative academic publishing process. The Alternative 1 has no delay in the academic publishing process, as it publishes all of the 1,345 submissions. However, the average number of cumulative citations of these papers is 22.54, which is the lowest among the four alternatives.

The Alternative 2, the current human intelligence based academic publishing process, has non-negligible delays in the process. The average number of cumulative citations of the published papers is 28.08, which is more than 24% higher than that of the Alternative 1. The results suggest that the papers selected by the human intelligence based academic publishing process turn to have higher average citations, even though editors and referees may not use paper's expected citations as one of the criteria when they decide which paper to publish.

The Alternative 3, the artificial intelligence based process, has no delay in the academic publishing process. The average number of cumulative citations of the top 700 papers selected by the artificial intelligence based process is 28.52, which is 2% higher than the human intelligence based process (Alternative 2). The impressive performance of artificial intelligence shows potential to be used as a tool to help editors to identify papers that are more likely to be highly cited. However, we cannot conclude that artificial intelligence should replace human editors and referees, since editors and referees may be able to identify the papers that are lowly cited but make important contributions to knowledge.

The Alternative 4, the academic publishing process that integrates human intelligence and artificial intelligence, has the same amount of delays in the process. For the 406 papers that were both selected by human experts and artificial intelligence, the average number of cumulative citations is 34.22, which is 22% higher than Alternative 2. Admittedly, the selection rule of the Alternative 4 is stricter, which contributes to the higher average number of cumulative citations of the papers selected by the Alternative 4. Nevertheless, the academic publishing process that combines human and artificial intelligence has shown its potential to be used to identify papers that are more likely to be highly cited among the publishable papers.

Table 2.22: Identifying highly cited papers: artificial intelligence v.s. human

Statistic	N	Mean	St. Dev.	25% Quantile	75% Quantile
Alternative 1: Publishing without reviewing					
Cumulative Citations of Non-rejected Papers	1,345	22.54	28.19	7	28
Cumulative Citations of Rejected Papers	0				
Alternative 2: Human intelligence					
Cumulative Citations of Non-rejected Papers	701	28.08	32.48	11	35
Cumulative Citations of Rejected Papers	644	16.50	21.03	4	21
Alternative 3: Artificial intelligence					
Cumulative Citations of Non-rejected Papers	700	28.52	33.40	10	35
Cumulative Citations of Rejected Papers	645	16.04	19.13	5	20
Alternative 4: Combining human and artificial intelligence					
Cumulative Citations of Non-rejected Papers	406	34.22	38.52	14	40.8
Cumulative Citations of Rejected Papers	939	17.48	20.32	5	22

2.6 Conclusion

This chapter analyzes editorial decision making in the academic publishing process. I find that papers with higher referee recommendation scores, higher scientific contribution scores, lower standard deviation of referee recommendation scores, higher share of positive referee recommendations, higher coverage of popular research topics, and written by authors with longer and more solid submission history (higher number of submissions and lower rejection rate) are more likely to be published. Papers with lower coverage of popular research topics and written by authors with shorter and weaker submission history are more likely to be desk rejected.

I also investigate the effects of paper and author information, editorial decisions, and referee recommendations on the duration from submission to decision and paper citations. For non-desk-rejected papers, the ones with higher referee recommendation scores and lower standard deviation of the scores have shorter durations of the first round of review. The results for paper citations suggest that accepted papers on average get higher citations than rejected ones, and higher paper citation counts are associated with higher coverage of popular research topics, referee recommendation scores, and scientific contribution scores.

In the prediction part, I use machine learning methods (regression shrinkage methods, Random Forest, and Gradient Boosted Trees) to predict paper citations with the information available at the time of submission. The model that uses Random Forest method, measures of publication information, measures of paper fields and topics, and high dimensional measures of the appearance of popular topic words gives the best out-of-sample prediction performance. Then, I use the preferred prediction model to test the possibility of adding artificial intelligence (AI) to the academic publishing process in order to help editors identify the papers that are more likely to be highly cited among publishable papers.

The experiment shows that the average number of cumulative citations of papers accepted by editors is more than 24% higher than all submissions. This result suggests that papers published by human editors turn to have higher average citations than rejected ones, even though editors may not use paper's expected citations as one of the criteria when they decide which paper to publish. As an exercise, I use the citation prediction model to decide which papers to publish based on maximizing citations. For a comparable acceptance rate as the human-based editorial process, the papers published by the algorithm have 2% higher citation counts. In addition, the average number of citations of the papers selected from the publishable papers by the artificial intelligence is 22% higher than all publishable papers. Admittedly, there are other factors that affect editors' decision on which paper to publish. However, the artificial intelligence based prediction model may help editor to identify the papers that are more likely to be highly cited from publishable papers.

The prediction part of this chapter focuses on predicting paper citations. It would be interesting for future study to test the possibility of predicting editorial decisions with the expected citations in order to deepen our understanding of the importance of expected citations in editorial decision making. It would also be interesting to see whether it is possible to train the machine learning method to identify articles that are sufficiently below the acceptance threshold of a journal to help editors "desk reject" a significant fraction of inappropriate or low-quality submissions.

Appendix A

Data Collection Details

A.1 Academic Data

A.1.1 Microsoft Academic data

The steps of collecting data from Microsoft Academic (MA) database:

1. Search for paper ID using the paper title.

2. For all the returned search results, compare the returned author names with the author names on the paper. If the last names of all of the authors are matched, then the paper is marked as identified paper.

3. Retrieve title, abstract, author list, journal information, and citation list of each of the identified papers in step 2.

4. Use author IDs to retrieve each author's publication list, author's co-author list, publication lists of co-authors of the author, and author affiliation[1] in MA database.

5. Use textual analysis to measure the retrieved paper and author information to construct quantitative variables.

A.1.2 Citation data from Google Scholar

The steps of collecting citation data from Google Scholar:

1. Use the paper title in the journal database to search articles in Google Scholar and keep the returned results (papers) shown on the first page.

[1] For most of the authors, MA database only contains one record of author affiliation, instead of a complete history of author affiliations. This limitation causes measurement error in the variables related to author affiliation.

2. Search the author list of each of the papers collected in Step 1 for the name of the corresponding author in the journal database, and keep the papers that have correct name matches.

3. Compare the paper title in the journal database with the paper(s) left after Step 2, and select the paper that matches the most words with the paper title in the journal database.

4. Use the total citations of the paper selected in Step 3 as the total citations of the paper in the journal database.

A.1.3 Comparison of Microsoft Academic database and Google Scholar database

The objective of this subsection is not to compare the data quality of Microsoft Academic (MA) database and Google Scholar (GS) database. Instead, I discuss some differences between collecting academic data from these two databases. These differences led me to use MA database as the main source of academic data.

First, MA database provides Application Programming Interface (API) for sending automatic queries to its database, and the data collection process does not require human intervention. Whereas, GS has built-in barriers required by publishers to foil automatic queries, which impedes the efficiency of collecting data from their websites.

Second, paper ID and author ID are accessible in MA database. This feature provides more flexibility in designing algorithms to reduce the measurement error of paper citation lists and author publication lists. Whereas, paper IDs and author IDs are not accessible in GS database. A scrapping algorithm may mix up publication lists of authors who have the same name, which causes miscounting.

Third, the rate limit of MA API is fairly high. This feature makes it possible to send a large number of queries to the database to collect a large group of researchers' collaboration network information.

Fourth, MA database allows retrieving paper's full citation list. Compared with MA database, GS website exhibits at most 100 pages of the citation list of a paper, which only contain 1,000 of the citing papers. This limitation truncates the citation lists of papers with more than 1,000 citations.

A.2 Paper Full Text

The steps of collecting paper full text of the top 5 economics journals:

1. Download the full texts of all of the papers in the top 5 economics journals during 1990-2011.[2]

2. Transfer the downloaded full texts of papers into plain txt files using "Poppler" library in Python.

3. Delete the headers and digital library information using Regular Expression.

4. Delete the sentences which contain any of "thank", "thanks", "comment", "comments", "gratitude", "grateful", "research assistance", "are those of the authors", "do not necessarily reflect" from paper texts to mitigate the disturbance of the acknowledgment of paper.

5. Use textual analysis to measure paper texts to construct quantitative variables.

A.3 Journal Databases

The steps of analyzing records of editorial decision making and other data from journal databases:

[2] The full texts of papers are downloaded from digital journal libraries (ScienceDirect and JSTOR) between January and April 2017 by a group of students to avoid mass downloading

1. Extract data on paper information (including paper ID, abstract, original submission date, decision date, decision status, editor name) and author information (corresponding author's name, institute of the corresponding author, author list) for papers having complete decision information (decision status and decision date) from journal databases, and create the paper table.

2. Link referee action table and referee account information table to get a table with complete referee information linked to each paper, create statistics of referee recommendations, and link the referee table to the paper table.

3. Link editorial decision table and editor account information table to get a table with complete editor information, and link the editor table to the paper table.

4. Construct historical records of editorial decisions for each editor, records of referee evaluations for each referee, and submission records for each corresponding author, and link these records to the paper table.

5. Use textual analysis algorithms to analyze abstract and keywords of each paper in the paper table to create variables measuring the quantified paper information.

A.4 List of Top Field Journals

Table A.1: List of top field journals

Journal Name	Journal Name
Econometrica	Journal of Economic Geography
Quarterly Journal of Economics	Journal of Marketing Research
The Review of Economic Studies	Journal of Time Series Analysis
Journal of Political Economy	Journal of Human Resources
Journal of Finance	World Bank Economic Review
Journal of Monetary Economics	Journal of Applied Econometrics
American Economic Review	Journal of Economic Behavior and Organization
Journal of Economic Theory	European Economic Review
Journal of Econometrics	Journal of Financial and Quantitative Analysis
Games and Economic Behavior	The Journal of Law and Economics
International Economic Review	Journal of Marketing
Journal of Financial Economics	Accounting Organizations and Society
Review of Financial Studies	Journal of Environmental Economics and Management
Journal of Economic Growth	Journal of Development Economics
Journal of International Economics	Economic Inquiry
The Review of Economics and Statistics	Financial Management
Journal of Labor Economics	Management Science
Journal of Business and Economic Statistics	International Journal of Forecasting
Journal of Public Economics	National Tax Journal
Economic Journal	Journal of Corporate Finance
Economic Theory	Industrial Relations
The RAND Journal of Economics	Journal of Urban Economics
Econometric Theory	Journal of Industrial Economics
Journal of Economic Dynamics and Control	Contemporary Accounting Research
Journal of Mathematical Economics	The Journal of Business
Journal of Risk and Uncertainty	Journal of the American Statistical Association
Journal of Money Credit and Banking	Explorations in Economic History
Marketing Science	The Scandinavian Journal of Economics
Accounting Review	Oxford Bulletin of Economics and Statistics
Review of Accounting Studies	Economica
Journal of Accounting Research	Oxford Economic Papers
Journal of Financial Intermediation	Canadian Journal of Economics
Review of Economic Dynamics	Journal of Comparative Economics
Macroeconomic Dynamics	International Journal of Industrial Organization
Journal of Financial Markets	Journal of Population Economics
Social Choice and Welfare	Economics Letters
Journal of Consumer Research	

Appendix B

Textual Analysis Algorithms

B.1 Algorithms for Measuring Paper Information

B.1.1 Measuring the topic words of a paper

The steps of measuring the topic words of a paper:

1. Use the keywords in JEL codes on American Economic Association website to get a dictionary of topic words by field.[1]

2. Use "Sentence Segmentation" function of "NLTK" module [10] in Python to segment paper texts into sentences, and separate each paper text into the following sections: the first 10 sentences, the first 100 sentences, the first 200 sentences, and full text.

3. Use Porter stemming algorithm to standardize the keywords in each dictionary and each section of paper texts, and tokenize paper texts to unigrams, bigrams, and trigrams.

4. Use the keywords dictionaries to parse the full texts of papers published in the top 5 economics journals during 1990-2011, and count the frequency of each keyword, the frequency of co-appearance of two keywords in a field, and the frequency of co-appearance of three keywords in a field.

5. Keep the top 40 frequent keywords, pairs of two keywords, and pairs of three keywords in each JEL code, and merge them into keyword list, two-keyword pair list,

[1] The keywords in field A (General Economics and Teaching), B (History of Economic Thought, Methodology, and Heterodox Approaches), N (Economic History), Y (Miscellaneous Categories), and Z (Other Special Topics) are excluded due to the small number of keywords in these codes.

and three-keyword pair list respectively.[2] After removing duplicated words, the topic word list contains 405 topic words, the two-word pair list contains 566 two-word pairs, and the three-word pair list contains 594 three-word pairs.

6. Use the word lists constructed in step 5 to parse the first 10 sentences, the first 100 sentences, the first 200 sentences, and the full texts of each paper, and create dummy equals one if a word (or word pair) appears.

B.1.2 Measuring the presentation style of a paper

The steps of measuring the presentation style of a paper:

1. Measure the number of pages of each paper.

2. Use "Sentence Segmentation" function of "NLTK" module in Python to segment the plain text files into sentences, and separate each paper texts into the following sections: the first 10 sentences, the first 100 sentences, the first 200 sentences, and full text.

3. Measure the number of sentences in each section, and tokenize text in each section to unigrams, bigrams, and trigrams.

4. Count the number of words (uni-grams) in each section, and calculate the average length (number of words) of sentences in each section.

5. Measure the frequency of adjective words in each section using an adjective-word dictionary.

6. Measure the frequency of advanced words in each section using an advanced-word dictionary.

[2] The words "paper", "research", "journal", "univers", "colleg", "graduat school", "economist", "book", "author", "mine" are excluded, because the frequencies of these words are disturbed by the words in papers' footnotes.

B.1.3 Measuring the abstract and keywords of papers in journal database

The steps of measuring the abstract and keywords of papers in journal database:

1. Extract the keywords and abstracts of papers (including both accepted papers and rejected papers) in the journal database.

2. Use "Sentence Segmentation" function of "NLTK" module [10] in Python to segment paper abstracts into sentences.

3. Use Porter stemming algorithm to standardize the keywords in each dictionary and paper abstracts, and tokenize paper abstracts to unigrams, bigrams, trigrams, and four-grams.

3. Parse the keywords part of each paper, and count the frequency of the appearance of each keyword.

4. Use about 200 most frequent topic words in each research field[3] to make research field dictionaries, and use about 500 most frequent topic words and word pairs[4] to make research topic dictionaries.

5. Use the research field word lists constructed in step 4 to parse the keywords part and abstract part of each paper to create research field dummy and measures of research field intensity of each paper.

6. Use the topic word lists constructed in step 4 to parse the keywords part and the abstract part of each paper, and create dummy equals one for the appearance of each topic word (or word pair).

[3] The difference in the number of words in each dictionary is caused by the words with same frequency.

[4] The difference in the number of words and word pairs in each dictionary is caused by the words and word pairs with same frequency.

B.2 Algorithms for Measuring Author Information

B.2.1 Measuring author publication information

The steps of measuring author publication information:

1. Drop the publications without abstract from author a's publication list.[5]

2. Count citations by year for each of the remaining papers in a's publication list.

3. Tokenize titles of the remaining papers in author a's publication list, and standardize the unigrams in paper titles using Porter stemming algorithm.

4. Compare the titles of the remaining papers in author a's publication list with each other, and check whether there exist multiple papers having "very similar" titles.[6]

5. Merge the papers with "very similar" titles in author a's publication list, and add up cumulative citations of these papers.

6. Calculate the number of author's publications, number of author's top field publications, and number of author's top 5 publications by parsing journal information of the remaining papers.

7. Calculate author's yearly cumulative citations by adding up the yearly citations of the remaining papers.

B.2.2 Measuring publication information of author's co-authors

The steps of measuring publication information of author's co-authors:

[5] By going through the publication lists of some authors, I find the publications without abstract are usually book chapters, conference speeches, or announcements. These publications cause miscounting of the number of an author's publications.

[6] Papers with "very similar" titles in an author's publication list are almost sure to be different versions of the same paper in the academic database. If these papers are not merged, the number of papers by author a is falsely inflated. The algorithm to compare titles of papers uses function "SequenceMatcher" in Python module "difflib", and the similarity ratio in "SequenceMatcher" is set as 0.8.

1. Use author a's publication list to obtain a list of author a's co-authors, and calculate the number of co-authors of author a.[7]

2. Drop the publications without journal information and the publications without citations.[8]

3. Count citations by year for each of the remaining papers of each co-author's publication list.

4. Tokenize titles of the remaining papers in each co-author's publication list, and standardize the unigrams in paper titles using Porter stemming algorithm.

5. Compare titles of the remaining papers in each co-author's publication list with each other, and check whether there exist multiple papers having "very similar" titles.[9]

6. Merge the papers with "very similar" titles in each co-author's publication list, and add up cumulative citations of these papers.

[7] The number of co-authors is counted in two ways – the "non-duplicated" way and the "duplicated" way. The "non-duplicated" way counts co-author who collaborates with the author in multiple papers within one year as one co-author. The "duplicated" way counts co-author who collaborates with the author in multiple papers within one year as multiple co-authors. The "non-duplicated" way indicates the extensiveness of the author's collaboration with co-authors, and the "duplicated" way indicates the intensiveness of the author's collaboration with co-authors.

[8] Due to the large number of papers in the publication lists of co-authors, I drop the publications without journal information and the publications without citations to reduce the computational burden. A big portion of the publications without journal information and the publications without citations are not academic papers, which are not in the scope of this study.

[9] Papers with "very similar" titles in one author's publication list are almost sure to be different versions of the same paper in the academic database. If these papers are not merged, the number of papers by the co-author is falsely inflated. The algorithm to compare titles of papers uses function "SequenceMatcher" in Python module "difflib", and the similarity ratio in "SequenceMatcher" is set as 0.8.

7. Calculate the total number of publications, number of publications in top field journals in economics, and number of publications in the top 5 economics journals written by co-authors of author a.[10]

8. Calculate the total yearly cumulative citations of papers by co-authors of author a.

B.2.3 Measuring economic research score of author's institution

The steps of measuring economic research score of author's institution:

1. Link author's affiliation name with the school name in Tilburg University Economics Schools Research Ranking, and get the economic research score of author's affiliation.[11]

2. Retrieve country of author's affiliation from OpenStreetMap.

3. Calculate the economic research score of each country using the total score of universities in that country, and link the country scores with the country of author's affiliation.

[10] The numbers of co-authors' publications are counted in two ways – the "cumulative" way and the "point" way. The "cumulative" way counts the number of each co-author's papers by the end of the year of the co-authorship. The "point" way only counts the number of papers written by co-authors in the year of the co-authorship. The "cumulative" way measures the "publication background" of the author's co-authors, and the "point" way measures the "activeness" of the author's co-authors.

[11] The default selection of journals is used, and the selected time range is 1990-2016.

Appendix C

Additional Figures and Tables

Figure C.1: Average citations of papers in the top 5 economics journals at the end of 2016

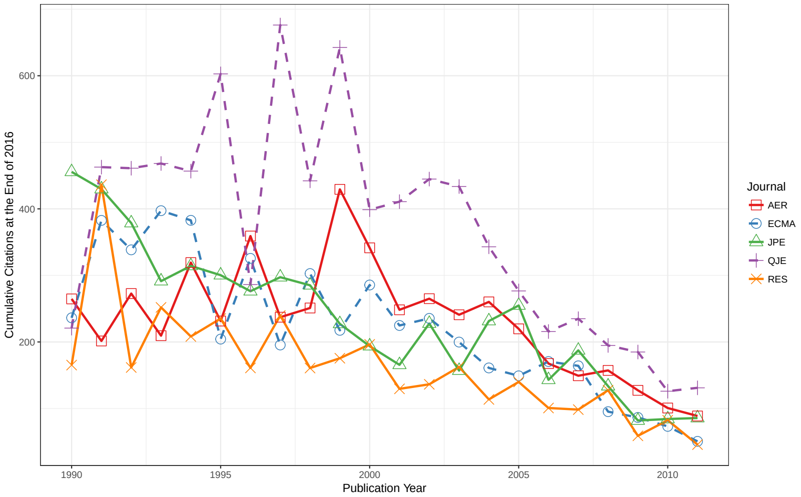


Table C.1: Cumulative citations: Top 5 economics journals

	Observations	Mean	Standard Deviation	Min	Max
All Top 5					
2016 Cumulative Citations (All)	4,482	245.07	455.16	1	11,567
10-year Citations (All)	3,472	133.74	191.77	1	2,771
10-year Citations (1990-1994)	1,046	80.31	130.01	1	1,586
10-year Citations (1995-2000)	1,105	137.82	189.15	1	2,292
10-year Citations (2001-2006)	1,321	172.63	222.72	2	2,771
AER					
2016 Cumulative Citations (All)	1,299	222.87	324.00	1	3,896
10-year Citations (All)	923	142.66	189.70	1	2,153
10-year Citations (1990-1994)	254	69.60	106.20	1	1,175
10-year Citations (1995-2000)	259	152.29	202.37	1	1,547
10-year Citations (2001-2006)	410	181.83	208.33	2	2,153
ECMA					
2016 Cumulative Citations (All)	941	225.97	444.62	1	5,857
10-year Citations (All)	759	118.08	182.68	3	2,771
10-year Citations (1990-1994)	214	85.40	135.97	3	1,146
10-year Citations (1995-2000)	235	109.46	129.30	3	762
10-year Citations (2001-2006)	310	147.17	234.12	3	2,771
JPE					
2016 Cumulative Citations (All)	754	254.67	399.02	2	5,435
10-year Citations (All)	636	122.10	157.99	2	2,314
10-year Citations (1990-1994)	208	87.50	106.60	2	689
10-year Citations (1995-2000)	226	118.35	129.22	6	1,117
10-year Citations (2001-2006)	202	161.92	213.58	6	2,314
QJE					
2016 Cumulative Citations (All)	767	375.23	617.91	3	7,718
10-year Citations (All)	609	197.95	263.09	1	2,292
10-year Citations (1990-1994)	203	105.50	188.77	1	1,586
10-year Citations (1995-2000)	208	216.66	285.54	2	2,292
10-year Citations (2001-2006)	198	273.07	276.71	4	1,949
RES					
2016 Cumulative Citations (All)	721	161.49	492.73	2	11,567
10-year Citations (All)	545	82.27	112.88	1	854
10-year Citations (1990-1994)	167	50.53	77.57	1	553
10-year Citations (1995-2000)	177	86.48	114.21	1	691
10-year Citations (2001-2006)	201	104.94	129.65	3	854

Table C.2: The top 5 highly cited papers in the top 5 economics journals 1990-2011

Journal	Title	Author	Year	GS CC	MA CC
AER	A sensitivity analysis of cross-country growth regressions	Ross Levine and David Renelt	1992	7,400	3,896
AER	Financial dependence and growth	Raghuram Rajan and Luigi Zingales	1998	7,740	3,418
AER	Gravity with gravitas: A solution to the border puzzle	James Anderson and Eric Van Wincoop	2003	5,910	3,002
AER	The twin crises: The causes of banking and balance-of-payments problems	Graciela Kaminsky and Carmen Reinhart	1999	5,780	2,750
AER	Does trade cause growth	Jeffrey Frankel and David Romer	1999	5,300	2,742
ECMA	Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models	Søren Johansen	1991	10,200	5,857
ECMA	The impact of trade on intra-industry reallocations and aggregate industry productivity	Marc Melitz	2003	10,300	4,779
ECMA	Conditional heteroskedasticity in asset returns: A new approach	Daniel Nelson	1991	9,550	4,570
ECMA	A model of growth through creative destruction	Philippe Aghion and Peter Howitt	1992	9,490	4,014
ECMA	Instrumental variables regression with weak instruments	Douglas Staiger and James Stock	1997	5,810	3,256
JPE	Increasing returns and economic geography	Paul Krugman	1991	13,000	5,435
JPE	Performance pay and top-management incentives	Michael Jensen and Kevin Murphy	1990	7,470	3,207
JPE	Property rights and the nature of the firm	Oliver Hart and John Moore	1990	6,440	2,811
JPE	A theory of fads, fashion, custom, and cultural change as informational cascades	Sushil Bikhchandani, David Hirshleifer and Ivo Welch	1992	6,270	2,790
JPE	Noise trader risk in financial markets	Bradford De Long, Andrei Shleifer, Lawrence Summers and Robert Waldmann	1990	5,500	2,502
QJE	A contribution to the empirics of economic growth	Gregory Mankiw, David Romer, and David Weil	1992	14,400	7,718
QJE	Economic growth in a cross section of countries	Robert Barro	1991	14,200	6,946
QJE	A theory of fairness, competition, and cooperation	Ernst Fehr and Klaus Schmidt	1999	9,010	4,969
QJE	Why do some countries produce so much more output per worker than others	Robert Hall and Charles Jones	1999	8,600	4,388
QJE	Corruption and growth	Paolo Mauro	1995	8,530	4,268
RES	Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations	Manuel Arellano and Stephen Bond	1991	17,900	11,567
RES	Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme	James Heckman, Hidehiko Ichimura, and Petra Todd	1997	3,900	2,329
RES	Identification of endogenous social effects: The reflection problem	Charles Manski	1993	4,863	2,303
RES	Income distribution and macroeconomics	Oded Galor and Joseph Zeira	1993	4,370	2,191
RES	Job creation and job destruction in the theory of unemployment	Dale Mortensen and Christopher Pissarides	1994	3,340	1,829

Note: The ranking in this table is based on the citation data of 4,482 papers used in this chapter. Some highly cited papers that did not return correct paper IDs from Microsoft Academic database are not included. The column "GS CC" reports the cumulative citations by the end of 2016 in Google Scholar database. The column "MA CC" reports the cumulative citations by the end of 2016 in Microsoft Academic database.

Table C.3: Paper research field intensities

	Mean(Standard Deviation)				
	AER	**ECMA**	**JPE**	**QJE**	**RES**
Math, Quant Methods	0.45%	0.43%	0.46%	0.46%	0.37%
	(0.20%)	(0.22%)	(0.21%)	(0.18%)	(0.26%)
Micro	0.72%	0.52%	0.79%	0.71%	0.65%
	(0.24%)	(0.26%)	(0.28%)	(0.26%)	(0.28%)
Macro, Monetary Econ	0.31%	0.17%	0.32%	0.33%	0.23%
	(0.18%)	(0.10%)	(0.16%)	(0.19%)	(0.14%)
International Econ	0.27%	0.17%	0.26%	0.28%	0.21%
	(0.14%)	(0.08%)	(0.11%)	(0.18%)	(0.13%)
Financial Econ	0.19%	0.11%	0.20%	0.20%	0.14%
	(0.10%)	(0.08%)	(0.09%)	(0.12%)	(0.08%)
Public Econ	0.18%	0.09%	0.18%	0.19%	0.13%
	(0.12%)	(0.07%)	(0.11%)	(0.16%)	(0.09%)
Health, Education, Welfare	0.16%	0.10%	0.17%	0.19%	0.11%
	(0.09%)	(0.06%)	(0.10%)	(0.13%)	(0.08%)
Labor Econ	0.25%	0.13%	0.27%	0.30%	0.16%
	(0.15%)	(0.10%)	(0.16%)	(0.18%)	(0.11%)
Law, Econ	0.04%	0.02%	0.05%	0.05%	0.03%
	(0.03%)	(0.02%)	(0.04%)	(0.04%)	(0.02%)
Industrial Organization	0.33%	0.19%	0.33%	0.35%	0.24%
	(0.15%)	(0.11%)	(0.16%)	(0.15%)	(0.12%)
Business Econ, Marketing	0.18%	0.10%	0.18%	0.19%	0.13%
	(0.07%)	(0.06%)	(0.08%)	(0.10%)	(0.06%)
Econ Development, Growth	0.24%	0.13%	0.24%	0.28%	0.17%
	(0.12%)	(0.07%)	(0.10%)	(0.25%)	(0.12%)
Econ Systems	0.21%	0.11%	0.22%	0.23%	0.15%
	(0.09%)	(0.07%)	(0.09%)	(0.10%)	(0.08%)
Agricultural, Environmental Econ	0.07%	0.03%	0.07%	0.07%	0.04%
	(0.07%)	(0.03%)	(0.08%)	(0.06%)	(0.04%)
Urban Econ	0.08%	0.05%	0.08%	0.09%	0.05%
	(0.06%)	(0.05%)	(0.06%)	(0.06%)	(0.04%)

Note: The measures of paper research fields are constructed by parsing papers' full texts.

Table C.4: Research field effect on paper citation paths

	Dependent variable:			
	Cumulative Citations: $C_{i,t}$			
	(1)	(2)	(3)	(4)
Math, Quant Methods×t	490.61	1,481.13**	1,457.14***	1,438.43**
	(492.18)	(582.91)	(563.35)	(561.92)
Micro×t	−1,903.75***	−638.29*	−5.35	−0.50
	(450.50)	(382.91)	(368.23)	(364.77)
Macro, Monetary Econ×t	−485.73	−883.85*	−1,005.47**	−1,022.29**
	(474.79)	(460.42)	(449.34)	(444.83)
International Econ×t	−1,558.10**	−1,117.04	−1,023.46	−1,017.59
	(791.18)	(709.26)	(638.52)	(638.68)
Financial Econ×t	1,862.29**	950.68	337.65	219.56
	(783.15)	(753.28)	(663.49)	(660.48)
Public Econ×t	−2,194.23***	−2,404.43***	−1,962.39***	−1,935.02***
	(819.65)	(774.58)	(670.04)	(655.95)
Health, Education, Welfare×t	701.35	1,433.11	1,364.02	1,374.77
	(1,199.16)	(1,168.41)	(1,128.90)	(1,124.99)
Labor Econ×t	−848.09	−1,797.47***	−1,423.48***	−1,429.61***
	(583.93)	(565.79)	(515.73)	(513.51)
Law, Econ×t	−754.59	−2,120.05	−2,400.78	−2,456.80
	(1,717.47)	(1,617.78)	(1,497.91)	(1,494.97)
Industrial Organization×t	−926.91**	−1,067.15**	−687.93*	−669.96*
	(467.46)	(427.86)	(389.16)	(386.55)
Business Econ, Marketing×t	−975.72	579.85	83.61	62.62
	(1,218.49)	(1,088.99)	(946.43)	(918.17)
Econ Development, Growth×t	4,757.84***	4,735.20***	3,863.91***	3,862.97***
	(1,451.15)	(1,190.83)	(906.89)	(841.11)
Econ Systems×t	4,620.56**	3,447.81**	3,388.94**	3,375.26**
	(1,980.17)	(1,701.31)	(1,541.77)	(1,535.50)
Agricultural, Environmental Econ×t	−131.16	−1,277.32	−413.13	−447.82
	(1,093.11)	(1,093.30)	(897.18)	(878.08)
Urban Econ×t	433.06	520.08	540.09	552.06
	(1,486.46)	(1,417.87)	(1,181.87)	(1,164.87)
FE: Paper ID	Yes	Yes	Yes	Yes
Control: Paper Topic and Presentation		Yes	Yes	Yes
Control: Author Information			Yes	Yes
Control: Co-author Information				Yes
Observations	73,706	73,706	73,706	73,706
R^2	0.27	0.29	0.38	0.38
Adjusted R^2	0.22	0.24	0.34	0.34

Note: *p<0.1; **p<0.05; ***p<0.01. The measures of paper information are constructed by parsing papers' full texts. Standard errors clustered by paper ID in parentheses.

Table C.5: Research topic effect on paper citation paths

	Dependent variable:				
	Cumulative Citations: $C_{i,t}$				
	(1)	(2)	(3)	(4)	(5)
Popular Topic Coverage×t	62.56***		43.69***	47.40***	20.04**
	(6.61)		(8.48)	(9.44)	(8.58)
Presentation: Descriptiveness×t		75.04	−1,106.64	551.20	20.09
		(2,607.64)	(2,622.63)	(2,244.25)	(2,035.85)
Presentation: Vocabulary Richness×t		27.87	−4.73	−327.09**	−429.72***
		(92.05)	(103.81)	(153.52)	(142.65)
Presentation: Complexity×t		−0.23	−0.07	−0.10	−0.05
		(0.16)	(0.16)	(0.15)	(0.14)
Num. Pages×t		0.47***	0.29***	0.18***	0.13**
		(0.05)	(0.07)	(0.06)	(0.06)
FE: Paper ID	Yes	Yes	Yes	Yes	Yes
Control: Field Information				Yes	Yes
Control: Au. and Co-au. Information					Yes
Observations	73,706	73,706	73,706	73,706	73,706
R^2	0.24	0.24	0.25	0.29	0.38
Adjusted R^2	0.19	0.19	0.20	0.24	0.34

Note: *p<0.1; **p<0.05; ***p<0.01. The measures of paper information are constructed by parsing papers' full texts. Standard errors clustered by paper ID in parentheses.

Table C.6: Author effect on paper citation paths

	Dependent variable:				
	Cumulative Citations: $C_{i,t}$				
	(1)	(2)	(3)	(4)	(5)
Num. Author×t	3.54***	3.70***	3.63***	3.37***	3.23***
	(0.80)	(0.85)	(0.81)	(0.81)	(0.81)
Au: Cumulative Citations(×1000)×t	1.31***		1.22**	1.35***	1.49***
	(0.50)		(0.53)	(0.51)	(0.51)
Au: Num. Pub.×t	0.10		0.11	0.07	0.11
	(0.09)		(0.09)	(0.09)	(0.09)
Au: Num. Top Field×t	−0.24**		−0.24**	−0.20*	−0.18
	(0.12)		(0.12)	(0.11)	(0.11)
Au: Num. Top 5×t	−0.50**		−0.51**	−0.42*	−0.48**
	(0.24)		(0.24)	(0.24)	(0.23)
Au: Num. Co-authored Pub.×t	−0.27**		−0.29**	−0.25**	−0.27**
	(0.12)		(0.12)	(0.12)	(0.11)
Au: Num. Co-authors×t		0.12*	0.06	0.05	0.05
		(0.07)	(0.06)	(0.06)	(0.06)
Co-au: Cumulative Citations(×1000)×t		0.15***	0.02	0.02	0.02
		(0.03)	(0.02)	(0.02)	(0.02)
Co-au: Num. Pub.×t		−0.005***	−0.002**	−0.002**	−0.001*
		(0.001)	(0.001)	(0.001)	(0.001)
Co-au: Num. Top Field×t		−0.002	0.002	0.001	0.001
		(0.004)	(0.004)	(0.004)	(0.004)
Co-au: Num. Top 5×t		0.01	−0.01*	−0.01	−0.01
		(0.01)	(0.01)	(0.01)	(0.01)
FE: Paper ID	Yes	Yes	Yes	Yes	Yes
Control: Field Information				Yes	Yes
Control: Paper Topic and Presentation					Yes
Observations	73,706	73,706	73,706	73,706	73,706
R^2	0.34	0.25	0.35	0.37	0.38
Adjusted R^2	0.30	0.20	0.30	0.33	0.34

Note: $^*p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$. "Au" is an abbreviation of "Author", and "Co-au" is an abbreviation of "Co-author". The measures of author information are constructed by "No-duplicated", and "Cumulative" ways described in Appendix B.2. Standard errors clustered by paper ID in parentheses.

Table C.7: Effects on paper citation paths by journal part I

	Dependent variable:					
	Cumulative Citations: $C_{i,t}$					
	All Papers	AER	ECMA	JPE	QJE	RES
Math, Quant Methods×t	1,438.43**	942.94**	1,032.38	−655.66	1,098.22	4,624.76
	(561.92)	(389.14)	(715.55)	(760.93)	(760.93)	(3,740.38)
Micro×t	−0.50	−894.85*	−1,254.66	1,225.40	−546.04	1,999.65
	(364.77)	(466.56)	(851.62)	(895.16)	(895.16)	(1,804.72)
Macro, Monetary Econ×t	−1,022.29**	−867.76*	872.81	−1,878.62***	359.70	−2,106.54*
	(444.83)	(497.19)	(1,400.71)	(728.41)	(728.41)	(1,204.11)
International Econ×t	−1,017.59	−476.35	−293.11	−1,512.29	−3,414.28***	−1,752.88
	(638.68)	(675.31)	(1,934.67)	(1,260.20)	(1,260.20)	(1,845.24)
Financial Econ×t	219.56	615.64	2,510.06	807.94	−1,220.20	1,738.08
	(660.48)	(1,008.63)	(4,221.18)	(1,177.61)	(1,177.61)	(1,742.59)
Public Econ×t	−1,935.02***	129.15	−3,324.18	−2,686.70*	−4,273.60***	−1,932.77
	(655.95)	(728.25)	(2,251.34)	(1,594.85)	(1,594.85)	(1,433.98)
Health, Education, Welfare×t	1,374.77	−280.70	5,908.52	−437.89	3,028.76**	−3,851.02
	(1,124.99)	(789.49)	(3,854.81)	(1,520.04)	(1,520.04)	(6,437.33)
Labor Econ×t	−1,429.61***	−1,499.48***	−444.63	12.85	−2,655.66***	4,497.87
	(513.51)	(559.22)	(1,478.55)	(906.64)	(906.64)	(3,858.73)
Law, Econ×t	−2,456.80	235.38	3,536.21	−7,456.79**	−2,433.76	−1,273.40
	(1,494.97)	(1,900.69)	(6,341.38)	(2,968.82)	(2,968.82)	(5,188.22)
Industrial Organization×t	−669.96*	−684.47	3,361.92	−259.37	−1,633.51***	−1,240.11
	(386.55)	(512.43)	(2,108.15)	(585.45)	(585.45)	(1,236.33)
Business Econ, Marketing×t	62.62	1,338.87	−2,856.37	925.18	−468.26	−5,980.06
	(918.17)	(1,288.53)	(2,872.51)	(1,742.13)	(1,742.13)	(7,007.99)
Econ Development, Growth×t	3,862.97***	3,182.21***	3,691.61	4,169.82***	6,485.30***	2,435.14
	(841.11)	(1,027.09)	(4,087.98)	(1,615.64)	(1,615.64)	(2,382.32)
Econ Systems×t	3,375.26**	3,529.96**	1,688.96	2,500.01	9,481.88***	4,684.73
	(1,535.50)	(1,630.83)	(3,041.96)	(1,922.31)	(1,922.31)	(5,261.63)
Agricultural, Environmental Econ×t	−447.82	−1,216.70	6,762.59*	−1,800.36*	1,987.03**	−1,241.55
	(878.08)	(895.75)	(4,059.34)	(954.73)	(954.73)	(2,766.14)
Urban Econ×t	552.06	−2,285.13*	5,829.24	3,518.88	−3,247.41	706.89
	(1,164.87)	(1,180.32)	(3,670.58)	(2,955.47)	(2,955.47)	(5,373.11)

Effects on paper citation paths by journal part II

	Dependent variable:					
	Cumulative Citations: $C_{i,t}$					
	All Papers	AER	ECMA	JPE	QJE	RES
Popular Topic Coverage×t	20.04**	13.03	−29.29	8.21	26.88	26.68
	(8.58)	(11.68)	(26.47)	(30.53)	(30.53)	(21.77)
Presentation: Descriptiveness×t	20.09	5,761.79	796.61	−4,748.21	841.22	−4,643.56
	(2,035.85)	(3,528.81)	(5,760.18)	(3,487.79)	(3,487.79)	(6,737.29)
Presentation: Vocabulary Richness×t	−429.72***	−478.16**	−977.34***	−40.37	−849.63***	−963.67
	(142.65)	(203.24)	(377.10)	(199.60)	(199.60)	(626.57)
Presentation: Complexity×t	−0.05	−0.43***	0.40	0.27	−0.79***	−0.47
	(0.14)	(0.16)	(0.31)	(0.30)	(0.30)	(0.47)
Num. Pages×t	0.13**	0.01	0.35**	0.48**	−0.02	−0.14
	(0.06)	(0.06)	(0.15)	(0.23)	(0.23)	(0.36)
Num. Authors×t	3.23***	1.97**	3.37**	4.38***	4.08***	3.41**
	(0.81)	(0.87)	(1.49)	(1.53)	(1.53)	(1.43)
Au: Cumulative Citations(×1000)×t	1.49***	−0.01	2.05*	3.73***	1.13	0.93
	(0.51)	(0.73)	(1.10)	(1.01)	(1.01)	(0.97)
Au: Num. Pub.×t	0.11	−0.09	−0.02	0.49	0.30	0.09
	(0.09)	(0.10)	(0.13)	(0.35)	(0.35)	(0.18)
Au: Num. Top Field×t	−0.18	−0.33	0.04	−0.48*	−0.18	−0.40
	(0.11)	(0.21)	(0.17)	(0.27)	(0.27)	(0.43)
Au: Num. Top 5×t	−0.48**	−0.21	−1.10*	−0.40	−0.68	−0.12
	(0.23)	(0.41)	(0.56)	(0.58)	(0.58)	(0.41)
Au: Num. Co-authored Pub.×t	−0.27**	0.06	−0.17	−0.64	−0.52	−0.17
	(0.11)	(0.12)	(0.14)	(0.42)	(0.42)	(0.18)
Au: Num. Co-authors×t	0.05	−0.05	0.03	0.08	0.07	−0.04
	(0.06)	(0.08)	(0.07)	(0.11)	(0.11)	(0.12)
Co-au: Cumulative Citations(×1000)×t	0.021	0.013	0.054	0.059	−0.004	−0.008
	(0.022)	(0.030)	(0.053)	(0.052)	(0.052)	(0.039)
Co-au: Num. Pub.×t	−0.001*	−0.003**	−0.002	0.0003	−0.001	−0.003
	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)
Co-au: Num. Top Field×t	0.001	−0.002	−0.003	0.01	−0.01	0.002
	(0.004)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Co-au: Num. Top 5×t	−0.01	−0.004	−0.02**	−0.01	0.01	−0.01
	(0.01)	(0.02)	(0.01)	(0.02)	(0.02)	(0.02)
FE: Paper ID	Yes	Yes	Yes	Yes	Yes	Yes
Observations	73,706	19,808	15,736	13,305	13,134	11,723
R^2	0.38	0.51	0.37	0.57	0.48	0.21
Adjusted R^2	0.34	0.48	0.33	0.54	0.44	0.16

Note: *p<0.1; **p<0.05; ***p<0.01. "Au" is an abbreviation of "Author", and "Co-au" is an abbreviation of "Co-author". The measures of paper information are constructed by parsing papers' full texts. The measures of author information are constructed by "No-duplicated", and "Cumulative" ways described in Appendix B.2. Standard errors clustered by paper ID in parentheses.

Table C.8: Effects on paper citation paths by author institution ranking part I

	Dependent variable:				
	Cumulative Citations: $C_{i,t}$				
	All Authors	Bottom 25%	Bottom 25%-50%	Top 50%-25%	Top 25%
Math, Quant Methods×t	1,438.43**	−1,584.05	−784.17	19,611.57*	1,187.48
	(561.92)	(1,402.51)	(804.78)	(11,304.30)	(2,295.02)
Micro×t	−0.50	−1,345.54	−987.65	2,561.25	1,367.85
	(364.77)	(930.39)	(1,087.25)	(3,819.16)	(3,732.98)
Macro, Monetary Econ×t	−1,022.29**	−1,299.78	−1,586.13	−458.99	6,166.12
	(444.83)	(1,796.06)	(1,660.00)	(4,029.64)	(4,548.20)
International Econ×t	−1,017.59	5,002.46*	445.92	−3,056.96	−8,555.33**
	(638.68)	(2,760.75)	(1,483.65)	(6,504.30)	(4,000.81)
Financial Econ×t	219.56	−4,161.97	324.84	59.96	−9,021.22
	(660.48)	(3,794.41)	(2,353.19)	(6,643.61)	(6,751.38)
Public Econ×t	−1,935.02***	−3,231.60	−2,272.10	−12,214.95	−7,307.19
	(655.95)	(2,052.51)	(1,913.56)	(9,698.25)	(4,928.71)
Health, Education, Welfare×t	1,374.77	2,539.51	−2,128.36	−28,586.25	30,986.40*
	(1,124.99)	(2,868.02)	(3,066.06)	(23,287.69)	(17,796.70)
Labor Econ×t	−1,429.61***	−4,602.27**	−334.13	10,448.19	−8,590.61
	(513.51)	(2,076.03)	(1,248.91)	(9,287.75)	(6,888.46)
Law, Econ×t	−2,456.80	1,519.05	−9,340.29**	−3,239.01	−28,335.03
	(1,494.97)	(5,637.31)	(4,501.60)	(14,672.95)	(20,604.39)
Industrial Organization×t	−669.96*	−654.99	474.81	−824.53	−690.28
	(386.55)	(1,653.13)	(1,206.66)	(5,132.80)	(2,447.25)
Business Econ, Marketing×t	62.62	−906.09	1,868.46	−11,862.93	22,354.55**
	(918.17)	(3,721.27)	(3,082.73)	(17,514.07)	(10,999.56)
Econ Development, Growth×t	3,862.97***	1,385.45	4,578.29*	2,124.42	13,417.02*
	(841.11)	(2,170.47)	(2,616.25)	(7,510.29)	(7,754.45)
Econ Systems×t	3,375.26**	4,297.17	4,793.70	20,079.63	1,600.16
	(1,535.50)	(4,602.62)	(4,408.76)	(16,827.50)	(8,754.70)
Agricultural, Environmental Econ×t	−447.82	−6,007.83	−3,685.53	6,234.40	3,396.70
	(878.08)	(3,710.40)	(4,215.52)	(12,893.47)	(3,827.76)
Urban Econ×t	552.06	27.55	−1,571.46	−10,483.08	9,119.27
	(1,164.87)	(5,249.81)	(2,476.89)	(11,604.05)	(12,005.66)

Effects on paper citation paths by author institution ranking part II

	Dependent variable:				
	Cumulative Citations: $C_{i,t}$				
	All Authors	Bottom 25%	Bottom 25%-50%	Top 50%-25%	Top 25%
Popular Topic Coverage×t	20.04**	42.72	−4.22	−45.26	−48.77
	(8.58)	(29.54)	(25.89)	(75.31)	(60.27)
Presentation: Descriptiveness×t	20.09	−6,839.54	5,935.79	−12,846.81	−7,305.42
	(2,035.85)	(6,685.48)	(6,507.05)	(24,875.84)	(18,078.70)
Presentation: Vocabulary Richness×t	−429.72***	441.93	79.23	−1,545.64	−2,888.37*
	(142.65)	(431.22)	(408.48)	(1,155.61)	(1,559.41)
Presentation: Complexity×t	−0.05	−0.94***	0.41	−2.25	−0.54
	(0.14)	(0.34)	(0.33)	(1.61)	(0.59)
Num. Pages×t	0.13**	0.43***	0.16	−0.22	−0.02
	(0.06)	(0.11)	(0.18)	(0.45)	(0.35)
Num. Authors×t	3.23***	−2.41	0.78	−10.98	16.78**
	(0.81)	(2.43)	(1.78)	(9.66)	(7.66)
Au: Cumulative Citations(×1000)×t	1.49***	1.62	−1.61	13.38	6.53***
	(0.51)	(2.61)	(2.42)	(9.64)	(2.18)
Au: Num. Pub.×t	0.11	−0.03	−0.38	−1.76	1.35*
	(0.09)	(0.28)	(0.29)	(1.31)	(0.69)
Au: Num. Top Field×t	−0.18	−0.51	0.17	−1.69	−0.42
	(0.11)	(0.32)	(0.36)	(1.23)	(0.98)
Au: Num. Top 5×t	−0.48**	−1.48*	−0.62	−0.54	0.02
	(0.23)	(0.82)	(0.89)	(1.68)	(1.10)
Au: Num. Co-authored Pub.×t	−0.27**	0.17	0.30	2.62	−1.89**
	(0.11)	(0.22)	(0.25)	(1.91)	(0.82)
Au: Num. Co-authors×t	0.05	−0.21	−0.10	−1.75	−0.46**
	(0.06)	(0.21)	(0.11)	(1.23)	(0.20)
Co-au: Cumulative Citations(×1000)×t	0.02	−0.08*	0.03	0.35	0.15
	(0.02)	(0.05)	(0.07)	(0.31)	(0.09)
Co-au: Num. Pub.×t	−0.001*	0.001	−0.004	−0.01	−0.01*
	(0.001)	(0.001)	(0.004)	(0.01)	(0.01)
Co-au: Num. Top Field×t	0.001	−0.01	0.003	0.04	−0.003
	(0.004)	(0.02)	(0.01)	(0.03)	(0.02)
Co-au: Num. Top 5×t	−0.01	−0.01	−0.04*	0.07	0.02
	(0.01)	(0.02)	(0.02)	(0.11)	(0.03)
FE: Paper ID	Yes	Yes	Yes	Yes	Yes
Observations	73,706	3,953	3,956	3,446	4,486
R^2	0.38	0.52	0.58	0.50	0.65
Adjusted R^2	0.34	0.46	0.54	0.44	0.62

Note: *p<0.1; **p<0.05; ***p<0.01. "Au" is an abbreviation of "Author", and "Co-au" is an abbreviation of "Co-author". The measures of paper information are constructed by parsing papers' full texts. The measures of author information are constructed by "No-duplicated", and "Cumulative" ways described in Appendix B.2. Standard errors clustered by paper ID in parentheses.

Table C.9: Effects on paper 10-year citation deciles by journal part I

	Dependent variable:					
	Cumulative Citations: $C_{i,t}$					
	All Papers	AER	ECMA	JPE	QJE	RES
Journal: ECMA	0.13					
	(0.14)					
Journal: JPE	0.13					
	(0.13)					
Journal: QJE	0.48***					
	(0.12)					
Journal: RES	−0.77***					
	(0.15)					
Math, Quant Methods	192.46***	182.40***	142.49**	172.65***	15.01	297.36***
	(28.06)	(51.48)	(60.22)	(59.63)	(75.65)	(90.99)
Micro	−89.80***	−290.80***	−274.12***	23.81	−110.62	−181.37*
	(29.59)	(57.12)	(76.40)	(55.99)	(76.78)	(99.98)
Macro, Monetary Econ	−19.00	29.78	7.99	−211.12**	126.83	66.54
	(43.29)	(74.57)	(151.31)	(86.77)	(93.45)	(116.19)
International Econ	32.32	−56.99	171.66	−55.94	76.79	10.13
	(47.58)	(83.85)	(149.09)	(118.04)	(94.16)	(141.05)
Financial Econ	70.29	−120.38	377.78*	179.12	56.59	85.14
	(62.09)	(113.83)	(215.36)	(126.75)	(113.58)	(220.37)
Public Econ	−165.18***	63.28	−292.26*	−144.05	−420.50***	290.63*
	(61.50)	(112.26)	(176.32)	(123.40)	(107.90)	(158.02)
Health, Education, Welfare	243.62***	156.17	304.38	174.22	418.77***	493.30*
	(72.64)	(132.09)	(242.24)	(138.43)	(157.22)	(271.56)
Labor Econ	−179.43***	−367.35***	−100.07	−142.67*	−158.54**	−49.88
	(40.03)	(74.69)	(134.05)	(86.14)	(73.32)	(175.21)
Law, Econ	−689.58***	−243.12	154.50	−1, 257.99***	−493.08*	−369.69
	(147.36)	(288.93)	(505.91)	(313.10)	(253.74)	(489.98)
Industrial Organization	−61.55	−141.89*	13.78	−51.42	−139.81*	146.81
	(44.31)	(79.61)	(163.83)	(91.25)	(80.13)	(149.87)
Business Econ, Marketing	149.29*	454.17**	122.22	176.15	207.10	34.22
	(89.25)	(179.06)	(258.52)	(176.75)	(170.68)	(280.53)
Econ Development, Growth	174.12**	430.08***	23.52	539.77***	149.46	175.73
	(72.80)	(123.93)	(247.13)	(167.48)	(119.00)	(204.85)
Econ Systems	88.55	161.80	264.17	44.15	281.04	−146.45
	(96.80)	(183.98)	(264.57)	(204.54)	(187.37)	(250.61)
Agricultural, Environmental Econ	−180.83**	−276.78*	−57.47	−267.91*	−140.20	−574.62*
	(90.34)	(159.07)	(340.25)	(140.86)	(165.26)	(330.67)
Urban Econ	155.08*	−128.92	550.57**	23.36	301.19	393.67
	(87.45)	(156.76)	(277.83)	(161.85)	(183.22)	(408.49)

Effects on paper 10-year citation deciles by journal part II

	Dependent variable:					
	Cumulative Citations: $C_{i,t}$					
	All Papers	AER	ECMA	JPE	QJE	RES
Popular Topic Coverage	9.21***	6.73***	6.83***	7.04***	7.28***	5.84**
	(0.80)	(1.46)	(2.13)	(2.21)	(2.08)	(2.67)
Presentation: Descriptiveness	125.30	558.84	821.98	−1,398.64***	−183.43	894.95
	(200.13)	(360.85)	(534.02)	(446.05)	(415.23)	(593.08)
Presentation: Vocabulary Richness	−33.54***	−26.20	−12.85	−18.00	−58.09**	−60.41*
	(11.05)	(20.47)	(29.93)	(22.45)	(29.63)	(34.54)
Presentation: Complexity	−0.03**	−0.04**	−0.02	−0.01	−0.04	−0.04
	(0.01)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)
Num. Pages	0.01***	0.002	0.05***	0.04*	0.02	0.01
	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.03)
Num. Authors	0.39***	0.16	0.27**	0.47***	0.64***	0.51***
	(0.06)	(0.11)	(0.13)	(0.15)	(0.13)	(0.16)
Au: Cumulative Citations(×1000)	0.23***	0.21*	0.41*	−0.02	0.07	1.02***
	(0.08)	(0.13)	(0.22)	(0.20)	(0.14)	(0.39)
Au: Num. Pub.	0.01*	0.01	−0.002	0.04	0.04*	0.01
	(0.01)	(0.01)	(0.01)	(0.03)	(0.03)	(0.03)
Au: Num. Top Field	−0.03**	−0.10***	−0.02	−0.03	0.03	−0.003
	(0.02)	(0.03)	(0.03)	(0.04)	(0.04)	(0.05)
Au: Num. Top 5	0.08***	0.14***	0.02	0.08	−0.04	0.11*
	(0.03)	(0.05)	(0.04)	(0.06)	(0.06)	(0.07)
Au: Num. Co-authored Pub.	−0.02*	−0.002	−0.01	−0.05*	−0.03	−0.06*
	(0.01)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)
Au: Experience	−0.02***	−0.02	0.004	−0.02	−0.07***	−0.04**
	(0.01)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Au: Num. Co-authors	0.01	0.05*	0.06	0.03	−0.03	−0.08*
	(0.01)	(0.03)	(0.04)	(0.02)	(0.03)	(0.04)
Co-au: Cumulative Citations(×1000)	0.006**	−0.001	0.008	0.026**	0.004	0.009
	(0.003)	(0.005)	(0.007)	(0.010)	(0.005)	(0.011)
Co-au: Num. Pub.(×1000)	−0.07	0.05	−0.21	0.09	−0.04	−0.03
	(0.08)	(0.13)	(0.21)	(0.29)	(0.12)	(0.15)
Co-au: Num. Top Field	0.001	0.001	0.001	−0.001	0.001	0.0001
	(0.0005)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Co-au: Num. Top 5	0.003***	0.002	0.002	0.003**	0.001	0.003*
	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Control: Pub. Year	Yes	Yes	Yes	Yes	Yes	Yes
Observations	3,472	923	759	636	609	545
R^2	0.36	0.35	0.35	0.37	0.48	0.37
Adjusted R^2	0.35	0.32	0.30	0.31	0.44	0.31

Note: *p<0.1; **p<0.05; ***p<0.01. Papers in AER at the year of 1990 are used as the baseline. "Au" is an abbreviation of "Author", and "Co-au" is an abbreviation of "Co-author". The measures of paper information are constructed by parsing papers' full texts. The measures of author information are constructed by "No-duplicated", and "Cumulative" ways described in Appendix B.2. Robust standard errors in parentheses.

Table C.10: Vectors of variables in prediction models

Model	**Vectors of Variables**	**Number of Variables**
Model (1)	journal information $\boldsymbol{J_i}$	2
Model (2)	field information $\boldsymbol{F_i}$ and journal information $\boldsymbol{J_i}$	107
Model (3)	popular topic coverage variables of $\boldsymbol{W_i}$, field information $\boldsymbol{F_i}$, and journal information $\boldsymbol{J_i}$	119
Model (4)	popular topic coverage variables of $\boldsymbol{W_i}$, presentation style $\boldsymbol{P_i}$, field information $\boldsymbol{F_i}$, and journal information $\boldsymbol{J_i}$	136
Model (5)	popular topic coverage variables of $\boldsymbol{W_i}$, presentation style $\boldsymbol{P_i}$, author information as of the year of publication $\boldsymbol{A_{i,0}}$, field information $\boldsymbol{F_i}$, and journal information $\boldsymbol{J_i}$	165
Model (6)	popular topic coverage variables of $\boldsymbol{W_i}$, presentation style $\boldsymbol{P_i}$, author information as of the year of publication and two previous years $\boldsymbol{A_{i,-2}, A_{i,-1}, A_{i,0}}$, field information $\boldsymbol{F_i}$, and journal information $\boldsymbol{J_i}$	221
Model (7)	popular topic coverage variables and topic word dummies of $\boldsymbol{W_i}$, presentation style $\boldsymbol{P_i}$, author information as of the year of publication and two previous years $\boldsymbol{A_{i,-2}, A_{i,-1}, A_{i,0}}$, field information $\boldsymbol{F_i}$, and journal information $\boldsymbol{J_i}$	1,836
Model (8)	popular topic coverage variables, topic word dummies, and 2-word pairs of $\boldsymbol{W_i}$, presentation style $\boldsymbol{P_i}$, author information as of the year of publication and two previous years $\boldsymbol{A_{i,-2}, A_{i,-1}, A_{i,0}}$, field information $\boldsymbol{F_i}$, and journal information $\boldsymbol{J_i}$	4,050
Model (9)	popular topic coverage variables, topic word dummies, 2-word pairs, and 3-word pairs of $\boldsymbol{W_i}$, presentation style $\boldsymbol{P_i}$, author information as of the year of publication and two previous years $\boldsymbol{A_{i,-2}, A_{i,-1}, A_{i,0}}$, field information $\boldsymbol{F_i}$, and journal information $\boldsymbol{J_i}$	6,234

Note: The explanation for these vectors is provided in Section 1.2. The variables constructed from author's institution information are not used as predictors, because institution information of more than half of the observations is not available. The measures of paper information are constructed by parsing papers' first 10 sentences, first 100 sentences, first 200 sentences, and full texts. The measures of author information are constructed by "No-duplicated", "Duplicated", "Cumulative", and "Point" ways described in Appendix B.2.

Table C.11: Machine learning methods in prediction models

Machine Learning Method	**Implementation of Algorithm**	**Key Parameters**
Plug-in Post-Lasso, Lasso	Chernozhukov et al. [16]	λ: Theoretically grounded plug-in method.
Cross-Validation Lasso, Ridge, Elastic Net	Friedman et al. [27]	λ: Cross-validation method. Number of folds=10.
Neural Network	Tensorflow system [1]	Number of hidden layers: 2, Number of units in the first hidden layer: 32, Number of units in the second hidden layer: 200, Activation function: Rectified Linear Unit (ReLU), Learning rate: 0.001, Dropout rate: 0.9, Number of epochs: 50, Batch size: 10.
Random Forest	Liaw and Wiener [50]	Number of trees: 5,000, Number of candidates at each split: 20, Size of bootstrap sample: Max.
Gradient Boosted Trees	Ridgeway [58]	Number of trees: 5,000, Interaction depth: 5, Shrinkage parameter (learning rate): 0.001, Bag fraction: 0.5.
Shrinkage-Random Forest Hybrid	Regression Shrinkage and Random Forest	Same as Regression Shrinkage and Random Forest above
Shrinkage-Gradient Boosted Trees Hybrid	Regression Shrinkage and Gradient Boosted Trees	Same as Regression Shrinkage and Gradient Boosted Trees above

Note: The parameters of these machine learning methods are selected after testing a bunch of parameter combinations.

Bibliography

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016.

[2] Victoria Anauati, Sebastian Galiani, and Ramiro Gálvez. Quantifying the life cycle of scholarly articles across fields of economic research. *Economic Inquiry*, 54(2):1339–1355, 2016.

[3] Joshua Angrist, Pierre Azoulay, Glenn Ellison, Ryan Hill, and Susan Lu. Economic research evolves: Fields and styles. *American Economic Review*, 107(5): 293–297, 2017.

[4] Ofer H Azar. The slowdown in first-response times of economics journals: Can it be beneficial? *Economic Inquiry*, 45(1):179–187, 2007.

[5] Ofer H Azar. Evolution of social norms with heterogeneous preferences: A general model and an application to the academic review process. *Journal of Economic Behavior & Organization*, 65(3):420–435, 2008.

[6] Scott Baker, Nicholas Bloom, and Steven Davis. Measuring economic policy uncertainty. *The Quarterly Journal of Economics*, 131(4):1593–1636, 2016.

[7] Ayeh Bandeh-Ahmadi. *Accounting for information: Case studies in editorial decisions and mortgage markets*. PhD thesis, University of Maryland, 2014.

[8] Onur Bayar and Thomas J Chemmanur. A model of the editorial process in scientific journals. Working paper, 2013.

[9] Alexandre Belloni, Daniel Chen, Victor Chernozhukov, and Christian Hansen. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429, 2012.

[10] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit.* O'Reilly Media, Inc., 2009.

[11] Christopher Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, 1995.

[12] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[13] David Card and Stefano DellaVigna. Nine facts about top journals in economics. *Journal of Economic Literature*, 51(1):144–161, 2013.

[14] David Card and Stefano DellaVigna. What do editors maximize? evidence from four leading economics journals. Working Paper 23282, National Bureau of Economic Research, March 2017.

[15] Dana Chandler, Steven Levitt, and John List. Predicting and preventing shootings among at-risk youth. *The American Economic Review*, 101(3):288–292, 2011.

[16] Victor Chernozhukov, Chris Hansen, and Martin Spindler. Hdm: High-dimensional metrics. *The R Journal*, 8(2):185–199, 2016.

[17] Aaron Clauset, Cosma Rohilla Shalizi, and Mark Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

[18] Christopher Cotton. Submission fees and response times in academic publishing. *The American Economic Review*, 103(1):501–509, 2013.

[19] Tom Coupé, Victor Ginsburgh, and Abdul Noury. Are leading papers of better quality? evidence from a natural experiment. *Oxford Economic Papers*, 62(1): 1–11, 2010.

[20] Liran Einav and Jonathan Levin. Economics in the age of big data. *Science*, 346 (6210):1243089, 2014.

[21] Glenn Ellison. Evolving standards for academic publishing: A q-r theory. *Journal of Political Economy*, 110(5):994–1034, 2002.

[22] Glenn Ellison. The slowdown of the economics publishing process. *Journal of political Economy*, 110(5):947–993, 2002.

[23] Glenn Ellison. How does the market use citation data? the hirsch index in economics. *American Economic Journal: Applied Economics*, 5(3):63–90, 2013.

[24] Maxim Engers and Joshua S Gans. Why referees are not paid (enough). *The American Economic Review*, 88(5):1341–1349, 1998.

[25] Jerome Friedman. Greedy function approximation: A gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[26] Jerome Friedman. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378, 2002.

[27] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

[28] Xavier Gabaix. Zipf's law for cities: An explanation. *The Quarterly journal of economics*, 114(3):739–767, 1999.

[29] Xavier Gabaix. Power laws in economics and finance. *Annual Review of Economics*, 1(1):255–294, 2009.

[30] Matthew Gentzkow and Jesse Shapiro. What drives media slant? evidence from us daily newspapers. *Econometrica*, 78(1):35–71, 2010.

[31] Matthew Gentzkow, Bryan Kelly, and Matt Taddy. Text as data. Working Paper 23276, National Bureau of Economic Research, March 2017.

[32] Colin Gillespie. Fitting heavy tailed distributions: The powerlaw package. *Journal of Statistical Software*, 64(2), 2015.

[33] Rachel Griffith, Narayana Kocherlakota, and Aviv Nevo. Review of the review: A comparison of the review of economic studies with its peers. Working paper, 2009.

[34] Daniel Hamermesh. Citations in economics: Measurement, uses and impacts. Working Paper 21754, National Bureau of Economic Research, November 2015.

[35] Daniel Hamermesh and Gerard Pfann. Reputation and earnings: the roles of quality and quantity in academe. *Economic Inquiry*, 50(1):1–16, 2012.

[36] Daniel Hamermesh and Peter Schmidt. The determinants of econometric society fellows elections. *Econometrica*, 71(1):399–407, 2003.

[37] Daniel S Hamermesh. Six decades of top economics publishing: Who and how? *Journal of Economic Literature*, 51(1):162–172, 2013.

[38] Lee Hansen, Burton Weisbrod, and Robert Strauss. Modeling the earnings and research productivity of academic economists. *Journal of Political Economy*, 86 (4):729–741, 1978.

[39] Michael Hilmer, Michael Ransom, and Christiana Hilmer. Fame and the fortune of academic economists: How the market rewards influential research in economics. *Southern Economic Journal*, 82(2):430–452, 2015.

[40] Jorge Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences of the United States of America*, pages 16569–16572, 2005.

[41] Marek Hlavac. Stargazer: Well-formatted regression and summary statistics tables (r package version 5.2), 2015.

[42] Mitchell Hoffman, Lisa Kahn, and Danielle Li. Discretion in hiring. Working Paper 21709, National Bureau of Economic Research, November 2015.

[43] 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.

[44] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. Working Paper 23180, National Bureau of Economic Research, February 2017.

[45] David N Laband. Is there value-added from the review process in economics?: Preliminary evidence from authors. *The Quarterly Journal of Economics*, 105 (2):341–352, 1990.

[46] Hannes Leeb and Benedikt Pötscher. Can one estimate the conditional distribution of post-model-selection estimators? *The Annals of Statistics*, pages 2554–2591, 2006.

[47] Hannes Leeb and Benedikt Pötscher. Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory*, 24(2):338–376, 2008.

[48] Derek Leslie. Are delays in academic publishing necessary? *The American economic review*, 95(1):407–413, 2005.

[49] Danielle Li and Leila Agha. Big names or big ideas: Do peer-review panels select the best science proposals? *Science*, 348(6233):434–438, 2015.

[50] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[51] Laurent Linnemer and Michael Visser. The most cited articles from the top-5 journals (1991-2015). Working Paper 5999, CESifo, July 2016.

[52] Tim Loughran and Bill McDonald. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65, 2011.

[53] Sendhil Mullainathan and Jann Spiess. Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106, 2017.

[54] Motty Perry and Philip Reny. How to count citations if you must. *The American Economic Review*, 106(9):2722–2741, 2016.

[55] Martin Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[56] Derek De Solla Price. Networks of scientific papers. *Science*, pages 510–515, 1965.

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

[58] Greg Ridgeway. Gbm: Generalized boosted regression models. *R package version*, 1(3):55, 2006.

[59] Klaus Ritzberger. A ranking of journals in economics and related fields. *German Economic Review*, 9(4):402–430, 2008.

[60] Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-june Paul Hsu, and Kuansan Wang. An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th international conference on world wide web*, pages 243–246. ACM, 2015.

[61] Scott Smart and Joel Waldfogel. A citation-based test for discrimination at economics and finance journals. Working Paper 5460, National Bureau of Economic Research, February 1996.

[62] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of machine learning research*, 15(1):1929–1958, 2014.

[63] Paul Tetlock. Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3):1139–1168, 2007.

[64] Howard Tuckman and Jack Leahey. What is an article worth? *Journal of Political Economy*, 83(5):951–967, 1975.

[65] Brian Uzzi, Satyam Mukherjee, Michael Stringer, and Ben Jones. Atypical combinations and scientific impact. *Science*, 342(6157):468–472, 2013.

[66] Hal Varian. Big data: New tricks for econometrics. *The Journal of Economic Perspectives*, 28(2):3–27, 2014.

[67] Dashun Wang, Chaoming Song, and Albert-László Barabási. Quantifying long-term scientific impact. *Science*, 342(6154):127–132, 2013.

[68] Ivo Welch. Referee recommendations. *Review of Financial Studies*, 27(9):2773–2804, 2014.

[69] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.