A MODEL OF ONLINE PUBLISHING:
THE CASE OF arXiv.org

MEIXIA ZHANG RUDERMAN

A DISSERTATION
PRESENTED TO THE FACULTY
OF PRINCETON UNIVERSITY
IN CANDIDACY FOR THE DEGREE
OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE
BY THE DEPARTMENT OF ECONOMICS
ADVISOR: FARUK GUL

NOVEMBER 2013
ABSTRACT

arXiv.org is a publishing platform, without formal refereeing, that is mainly used in physics and math. Taking arXiv.org as an example, we show that authors can self-referee, i.e., among authors of the same quality, those with good papers are more likely to make papers more visible to readers to help readers evaluate papers. Therefore, authors play a role analogous to referees in the peer review process. Our work suggests that online publishing, like arXiv.org, provides an alternative to journals, where the traditional role of evaluating papers through peer review is replaced by authors’ self-refereeing. We also address the broader question: do producers have incentives to reveal the quality of their bad products? We propose that authors’ self-refereeing is driven by their reputation concerns. Given papers of the same quality, good authors make their papers less visible than bad authors in order to benefit from the reputation improvement in the future. Good authors are more likely to invest in their reputation because they are more likely, than bad authors, to have good papers in the future. In general, good papers are cited more than bad papers. This citation advantage is higher if a good paper’s author invested in their reputation. The more limited readers’ attentions are, the more likely authors are to self-referee, because authors with good papers incur a higher cost for their papers to be missed by readers. In Chapter 1, we identify properties of authors’ strategic submissions, which motivates the model that we present in Chapter 2. In Chapter 3, we collect data and test the theoretical model of Chapter 2.
I am grateful to the many people whose help and support have made this thesis possible. First and foremost, I would like to thank my advisor, Faruk Gul. His support for free thinking and independent work is behind my journey at Princeton to find out who I am and what type of research I should pursue in the future. His inspiration and advice have been tremendously helpful. I would also like to thank Alexandre Mas. He guided me into the empirical work. I really enjoy the numerous discussions with him. He contributed significantly to this thesis and provided me with strong encouragement especially during the difficult period of my life. I would also especially like to thank Wolfgang Pesendorfer. His support and encouragement for this thesis have been really important for me.

I am also especially grateful to other people that have helped me a lot. All through my graduate school, Roland Benabou has discussed with me many times and always read my drafts carefully. I benefited a lot from his advice and his critical questions. Stephen Morris, whom I worked with during the first summer at Princeton, has provided support during different phases of my graduate studies. I would also like to thank my good friend Ye Chen, who is currently at Cornerstone Research. As an empirical researcher and econometrician, she selflessly spared her time and helped me tremendously in many discussions. I also benefited greatly from the discussions with Yoon Soo Park, a labor economist, who spent time on me even during the busiest season of job application. I am looking forward to have more conversations with Sylvain Chassang who has shown
support and whose research is inspirational to me. I would also like to thank the audience in the microeconomic theory student lunch seminar at Princeton. I have always been inspired by their questions.

Finally, I would like to thank other people that are important to my life. My parents, Qin Han and Shumin Zhang, above anyone else, sacrificed, beyond what language can describe, to support my career even though I have never been able to visit home since I came to the US. Their love to me and their devotion to work have endlessly driven me to pursue something better. My husband, Joshua Ruderman, has consistently provided me love and support since we first met at Princeton. My grandfather, Jinsheng Han, and my grandmother, Zhang Han (Zhang is her last name. She does not have a formal name because she never went to school.), both passed away during my stay in the US. I know they have been always proud of me and looked forward to my coming back to China. I would also like to thank my brothers, Sen and Lin, and my sister, Xiumei, who provided love and comfort for my parents during the years I was away from home. My husband’s family, parents Maureen and Steve as well as sister Danielle, have treated me as a family member from the beginning.

Princeton was much more fun because of my classmates (Ye Chen, Dong Beom Choi, Hyun Soo Choi, Qi Ge, Peter Ho, Jia Li, Tao Wang, and others), my economist friends (Qingqing Cao, Yoon Soo Choi, Liang Dai, Cheng Chen, Xiaotong Niu, Yi Li, Weicheng Lian, Takuo Sugaya, Sergey Zhuk, and many others), my physicist friends (Alex, Cliff, Jared, and others), and many others!
To my parents Qin Han and Shumin Zhang.
Contents

Abstract iii

Acknowledgements iv

1 Introduction 1

2 Chapter 1: The Existence of Strategic Submissions 5

  2.1 Main Facts ......................................................... 8
  2.2 The Length of the List and Fact Three .......................... 14
  2.3 The Publishing on Journals and Fact Three ..................... 18
  I. Data .............................................................. 21
  II. Before-and-after Analysis ...................................... 22
  III. Difference-in-Difference Model .................................. 25
  2.4 Literature ......................................................... 28

3 Chapter 2: The Model of Self-refereeing 34

  3.1 The Model ........................................................ 37
3.2 Theoretical Interpretation of Empirical Results

From Chapter 3 ................................................................. 61
I. The Model With Multiple Groups of Authors ...................... 62
II. The Inference From Chapter 3 ......................................... 63

3.3 Welfare Analysis .......................................................... 65

3.4 Discussions About Assumptions ....................................... 67
I. Authors’ Information About The Quality of Their Paper ........ 67
II. Other Assumptions ....................................................... 69

4 Chapter 3: Testing The Model ......................................... 71

4.1 Data and Measurement ................................................ 73
I. Data ........................................................................... 73
II. Measurement of Main Variables ......................................... 76
III. Summary Statistics ...................................................... 80

4.2 Authors’ Submission Behavior:
Aggregate Data ............................................................. 82
I. The Graphic Evidence .................................................. 82
II. Empirical Modelling ..................................................... 86
III. Empirical Results ...................................................... 87

4.3 Authors’ Submission Behavior:
Different Ages ............................................................ 89

4.4 Discussions ............................................................... 94
I. Deadline Effect ......................................................... 94
II. Naive and Sophisticated Authors .................................. 97

4.5 Positional Effect .................................................. 98

I. Positional Effect Involving Top Positions .......................... 99

II. Positional Effect Involving Bottom Positions ..................... 101

III. Relation To The Literature ........................................ 102

5 Conclusion ......................................................... 109
Introduction

In recent years, there have been two trends of academic publishing. On one hand, the peer review process has been significantly slowed down\(^1\); on the other hand, online publishing has become increasingly popular. arXiv.org is a free website mainly used in physics and math for publishing papers. From August 1991 until September 2013, there has been over 877,000 papers stored. Taking arXiv.org as an example, we show that through one website without formal refereeing, authors can *self referee*, *i.e.*, among authors of the same quality, those with good papers are more likely to make their papers more visible to readers to help readers evaluate their papers. Therefore, authors play a role analogous to referees in the peer review process. Since online publishing expedites the speed of publishing, our work suggests that online publishing should also be considered as an alternative way to peer view in terms of journals’ traditional role of disseminating and evaluating papers for other disciplines. Referees are free of the burden of reading papers. Researchers can spend more time working on new ideas.

The idea of self refereeing gives a negative answer to the following general question: will producers make all their products more visible to consumers if there is no physical cost involved? arXiv.org is the market place for papers which are public goods. Authors are producers, always wanting their papers cited; readers are consumers, only benefiting from

\(^1\)Ellison [2001] provides evidence for the slow down of peer review for main journals in different fields between 1975 and 1999 (Table 2, p. 45). The slow down is more severe for social sciences, for example, economics.
citing good papers. Authors’ self-refereeing is driven by their reputation concerns. Good authors make papers less visible for the reputation gain the future. With papers of the same quality, good authors gain more from the investment in reputation because they are more likely to have good papers in the future and the citation advantage of the future good papers is higher if their author invested in reputation. If readers’ attention to papers at less visible spots is more limited, authors have an even stronger incentive to self referee because good papers have a higher cost of being missed by readers. The idea of self refereeing is not limited to this market. For example, some companies advertise on TV during shows of higher rating and some others choose shows of lower rating. One possible explanation is these companies are self refereeing. Good products are more likely to be shown during shows of higher rating and bad products are more likely to be shown during shows of lower rating. By less heavily advertising their bad products, good companies build up their brand.

The current submission system of arXiv.org provides a channel for authors to influence the visibility of their papers through strategically choosing their submission times. We choose data from high energy physics-theory because authors in this field have a strong incentive for strategic submissions. In Chapter 1, we identify the existence and properties of strategic submissions which are summarized in three facts. In Chapter 2, we build the game-theoretic model of self refereeing to accommodate these facts. In particular, we derive the necessary conditions for any equilibrium consistent with the facts for the empirical test in Chapter 3. The results from Chapter 3 show all types of authors self referee and bad authors are more likely to self referee. Both bad authors with bad papers and good authors with good papers invest in reputation. It implies readers’ attention to less visible spots is probably very limited. The reputation gain for authors with bad papers is reduced as more and more bad authors
with bad papers invest in reputation. Bad authors with bad papers still have an incentive to invest in reputation as long as good authors with good papers invest in reputation with a sufficiently high probability. Good authors with good papers have an incentive to invest in reputation as long as a lot of them invest in reputation such that the reputation gain is sufficiently big. With a big reputation gain, authors with good papers make papers more visible only when good papers’ cost of being missed by readers is sufficiently high.

Our work belongs to the literature on limited attention, reputation concerns and economics of publishing. First, we contribute to the rare study of strategic behavior under limited attention. The most related work is from DellaVigna and Pollet (2008). They show that investors are more distracted on Fridays than other weekdays. To them, bad news is worse than no news and therefore companies hide bad news on Fridays. For us, any news is better than no news because authors always want their papers cited. Authors make papers less visible because of their reputation concerns. Though our model has implications about readers’ citing behavior, we only focus on authors’ submission behavior for the empirical analysis. Their empirical analysis is only from investors’ rather than companies’ side. Second, our reputation mechanism is mostly related to Morris (2001) and we contribute to the limited literature on the testing of reputation concerns with real world data. In Morris’ repeated cheap talk model, reputation concerns are due to advisors’ different utility functions and their different way of discounting the future. The utility function is constructed such that good advisors with a bad message and bad advisors with a good message always have a strict incentive to tell the truth. Therefore, reputation effect as defined in our paper automatically holds. In our model, good and bad authors have the same utility function and discount the future in the same way. Reputation effect works through authors’ dif-
ferent prospect about the future and the information readers get from reading the paper. In terms of empirical testing, except for the line of literature in the study of ebay, Banerjee and Duflo (2000) and Crocker and Reynolds (1993) empirically examine the effect of reputation on the choice of contracts. Finally, this paper also contributes to economics of publishing. Berger (2009) examines the citation impact of papers’ order in journals. Edelman and Larkin (2009) studies self-downloading on SSRN. Ellison has a series of papers on publishing in economics. For example, Ellison (2002) and Ellison (2011) document the slow down, and decline, of peer review in economics publishing, respectively. A model of peer review with evolving standards is built by Ellison (2001). More recently, Ellison (2010) constructs variants of h-index to match the labor market outcomes in the field of economics\textsuperscript{2}.

\textsuperscript{2}Please check Chapter 3 for the discussion on h-index.
Chapter 1: The Existence of Strategic Submissions

The current submission system of arXiv.org provides a channel for authors to influence the location of their paper through strategically choosing their submission times. Authors can submit any time. Papers submitted between 4pm eastern time on any working day (day 1) and 4pm eastern time on the next working day (day 2) will be on the same list to be announced at around 8pm eastern time on day 2. The order is based on original submission times and first-come, first-served\(^1\). Authors can influence the positions of their papers on the list by strategically choosing the date of submission and specific submission time within any day\(^2\). For example, papers submitted very close to 4pm are more likely to get top or bottom positions. If top and bottom positions are more visible to readers, authors always want their papers to be cited more, and all readers only cite through the reading on arXiv.org, then authors have a strong incentive to strategically choosing submission times.

To show the existence of strategic submissions, we choose data from the field of high energy physics-theory (hep-th). arXiv.org originated from this field in 1992. The reason for choosing the field of high energy physics-theory is that authors have a strong incentive for

---

\(^{1}\) First-come, first-served is in terms of the availability of spots, which is roughly consistent with the original submission times. This definition takes into account the possibility that some author withdraws before 4pm. Then, one paper which is submitted away from 4pm may accidentally get this author’s spot. However, for the submitters in our sample, between 2002 and 2004, only two such papers got top 2 positions. Therefore, for the remaining part of the paper, we can just assume the order is based on original submission times.

\(^{2}\) We define one day in terms of the period of time during which all submissions are on the same list.
strategic submissions. More specifically, an overwhelming majority of papers are submitted to arXiv.org and all readers only cite through the reading on arXiv.org. That is to say, all information on all readers’ evaluation is through arXiv.org and incorporated in the number of citations, which has two implications. On one hand, authors have a strong incentive to influence readers’ citing behavior; on the other hand, we can identify properties of strategic submissions by analyzing the number of citations for each paper. We summarize the evidence in terms of the two traditional functions of academic publishing: the disseminating and evaluating of papers.

First, arXiv.org has been successful in disseminating papers. All papers related to the field of high energy physics-theory are accepted to arXiv.org\(^3\). Therefore, papers are disseminated without formal refereeing. To be consistent with other analysis, we choose the data between 2002 and 2004. During this time window, a vast majority of papers are disseminated through arXiv.org. Figure 5.1 gives an overview of publishing in this field between 2002 and 2004\(^4\). As you can see, papers on arXiv.org are divided into *arxivjournal* and *arxiv*. *arxivjournal* denotes papers on both arxiv and peer review journals, while *arxiv* refers to papers only on arXiv.org. In total, 80.5% of the total publications are on arXiv.org. Moreover, those papers also published on journals are submitted to arXiv.org before the acceptance to journals. For example, Almost all papers on the four main journals between 2002 and 2004 are submitted to arXiv.org before the submission (for the marjority) or acceptance to

---

\(^3\)There is an algorithm to automatically identify papers unrelated to this field. Staff also manually check each list.

\(^4\)The data are collected from inspirehep.net. We take all papers with subject theory-HEP as papers in the field of high energy physics-theory. Some papers have missing subject information. We divide these papers based on the ratio of papers with subjects in theory-HEP and other subjects.
journals\textsuperscript{5}. In addition, arXiv.org has also been widely accepted since 2005 as suggested by Figure 5.2. Figure 5.2 shows the average monthly submission rate over the past 20 years. We can see that, since 2002, the usage has been stable.

Second, the publishing on journals does not have a citation impact, which implies authors only cite from arXiv.org. The argument in the previous paragraph shows readers evaluate through either arXiv.org or journals or both. The following evidence further shows that arXiv.org is the only platform through which readers evaluate papers. More specifically, for any paper, almost all recent references come from arXiv.org. If one paper cited is also published in a peer review journal, the paper’s arXiv id is usually listed together with the journal in the citing paper. More formally, in section 2.3, we show that the publishing on journals does not have a significant citation impact.

With papers submitted between 2002 and 2004 from US or Canada, we find the evidence for strategic submissions. The outline of the following analysis in chapter 1 is as follows. In the next section, we will summarize the evidence and properties of strategic submissions in three main facts. Fact One shows both the evidence and one property. The other two facts show two additional properties. Section 2.2 shows the property as identified in Fact Three exists in lists with different number of papers. Section 2.3 shows the number of citations for each paper is not significantly influenced by journals, which implies the citation pattern in

\textsuperscript{5}The four main journals are Journal of High Energy Theory (JHEP), Nuclear Physics B (NPB), Physics Letters B (PLB) and Physics Review D (PRD). They include 63.1\% of all papers from hep-th during this time window. For JHEP, 99.27\% (96.4\%) were submitted to hep-th before the acceptance (submission) to journals. For NPB, this ratio is at least 88.8\% (76.6\%). It is 99.68\% (96.2\%) for PLB and 92.85\% (79.6\%) for PRD.
Fact Three results from the strategic interaction between authors and readers rather than the influence of journals. All directly related literature to Chapter 1 are from physicist Dietrich [2008, a,b] and Hague and Ginsparg [2009, 2010], which will be discussed in Section 2.4.

2.1 Main Facts

In this section, we will show the existence of strategic submissions and identify two main properties, which can be summarized in three facts. All papers in the sample are submitted from the US or Canada by their authors. The physical submission cost is small for our sample since 4pm eastern time is within submitters’ working time. The small physical submission cost for submitters from US or Canada highlights the importance of strategic interaction between authors and readers. Our findings in this section motivate us to build a game theoretical model without physical submission cost in Chapter 2.

Fact One: There are strategic submissions around the deadline (D submissions, within 30 minutes before and after the deadline) and away from the deadline (ND submissions, more than 30 minutes away from the deadline). We show Fact One with Figure 2.1 and 2.2. As shown in the top figure of Figure 2.1, there is a submission spike within 30 minutes before and after 4pm eastern time between 2002 and 2004 and a lot of other submissions are out of this time window. Moreover, the further away from 4pm, the less number of submissions. The bottom figure of Figure 2.1 and both figures in Figure 2.2 show that the

6Some papers are not submitted by their authors. The detailed description of the data is in Chapter 3 and Appendix C.
submission pattern is not due to 4pm as a specific deadline. As is shown in the bottom figure of Figure 2.1, between 2000 and September of 2001 when the deadline was 7pm eastern time, we find a similar submission pattern. Figure 2.2 shows that according to local times, the submission patterns for submitters from eastern time zone and pacific time zone are similar. For submitters from Pacific time zone, there is a submission spike around 1pm pacific time but not around 4pm pacific time.

Fact Two: Almost all D submissions appear in the top 2 or bottom 2 positions and a significant fraction of papers that appear in the top 2 or bottom 2 positions are ND submissions. As shown in Table 2.1, almost 90% of D submissions between 2002 and 2004 have top 2 or bottom 2 positions and, for papers submitted within 30 minutes after 4pm, this proportion is even higher at 93.5%. Moreover, almost all D submissions before (after) 4pm have bottom 2 (top 2) positions. Table 2.2 shows that there are not enough D submissions to fill the top 2 and bottom 2 positions. In total, 69.57% of papers at top 2 and bottom 2 positions are ND submissions. For bottom 2 positions, this proportion is even higher at 73.06%.

Fact Three: Given papers at the top 2 or bottom 2 positions, D submissions on average have more citations than ND submissions. Figure 2.3 visually shows that, at both top 2 and bottom 2 positions, the median number of citations for D submissions is higher than ND submissions. Wilcoxon-ranksum test shows the distribution of the number of citations for D submissions stochastically dominates that for ND submissions and this comparison is significant at 1% level. Since Table 2.2 shows that almost all D submissions at top 2 (bottom 2) positions are submitted after (before) 4pm, the citation advantage of D submissions is not caused by submissions right before (after) 4pm which by accident have top 2 (bottom
Figure 2.1: North American (NA) submissions by time of day, in 30 minutes bins, between 2000 and 2004. All holiday, Saturday and Sunday submissions are excluded. The top figure shows submissions between 2002 and 2004 when 4pm eastern time was the deadline (2242 papers); the bottom figure shows submissions between Jan. 2000 and Sep. 2001 when 7pm eastern time was the deadline (897 papers). The red vertical lines show the deadlines.
Figure 2.2: North American (NA) submissions from Eastern and Pacific timezones by time of day, in 30 minutes bins, according to local times between 2002 and 2004. The top figure shows submissions from the Eastern timezone (1299 papers); the bottom figure shows submissions from the Pacific timezone (519 papers). The red vertical lines show 4pm eastern time and 1pm pacific time, respectively. The green vertical line in the bottom figure shows 4pm pacific time. All other specifications are the same as in Figure 2.1.
2) positions. When we exclude D submissions before 4pm (after 4pm) at top 2 positions (bottom 2 positions), we get a similar citation pattern.

Table 2.1: D Submissions (NA) and Positions

<table>
<thead>
<tr>
<th>Time</th>
<th>Total</th>
<th>Top2</th>
<th>Bot2</th>
<th>Fraction*</th>
</tr>
</thead>
<tbody>
<tr>
<td>after 4pm</td>
<td>246</td>
<td>229</td>
<td>1</td>
<td>93.5%</td>
</tr>
<tr>
<td>before 4pm</td>
<td>181</td>
<td>6</td>
<td>148</td>
<td>85.1%</td>
</tr>
<tr>
<td>total</td>
<td>427</td>
<td>235</td>
<td>149</td>
<td>89.9%</td>
</tr>
</tbody>
</table>

*Note: The fraction of papers at top 2 or bottom 2 positions among all D submissions.

Table 2.2: Top/Bottom Positions and NA Submissions

<table>
<thead>
<tr>
<th>Positions</th>
<th>Submissions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Name</td>
</tr>
<tr>
<td></td>
<td>Number</td>
</tr>
<tr>
<td>Top2</td>
<td>709</td>
</tr>
<tr>
<td>Bot2</td>
<td>553</td>
</tr>
<tr>
<td>total</td>
<td>1262</td>
</tr>
</tbody>
</table>

*Note: The fraction of ND submissions among all papers at top 2 or bottom 2 positions.

Fact Two shows a significant fraction of submissions are away from 4pm, which raises the following question: given the small physical submission cost and the fact that readers
Figure 2.3: Median number of citations for D and ND submissions at the top 2 and bottom 2 positions between 2002 and 2004. The number of citations are collected at the end of September of 2012. All holiday submissions defined according to What’s New on arXiv.org are excluded. P-values for Wilcoxon rank sum test at top 2 positions and bottom 2 positions are 0.0004 and 0.0041 respectively.

Probably pay more attention to top and bottom positions, why do authors submit away from 4pm? Fact Three makes the question more specific: why do authors submit papers cited less away from 4pm? Our conjecture is authors self referee, i.e., among authors of the same quality, those with bad papers are more likely to submit away from 4pm. Therefore, papers submitted away from 4pm on average have lower paper quality. Since readers cite based on their estimation of paper quality which positively correlated to paper quality, papers submitted away from 4pm are cited less. Authors submit away from 4pm for the reputation improvement in the future. With papers of the same quality, good authors are more likely to submit away from 4pm because they are more likely to have good papers in the future.
2.2 The Length of the List and Fact Three

In this section, we will check whether the citation pattern as we find in Fact Three depends on the length of the list, that is, the number of papers on the same list. There are 741 lists with at least one paper submitted from US or Canada between 2002 and 2004. The mean number of papers on the lists is 13 with a standard deviation of 4.751102. Since we define top positions as top 2 positions and bottom positions as bottom 2 positions, each list in our sample should have at least 4 papers. Therefore, for the following analysis, we will only consider lists with at least 4 papers. Only 4 out of 741 lists have less than 4 papers.

Figure 2.4: The citation comparison of D and ND submissions at top 2 or bottom 2 positions for lists of different lengths. Short lists refer to lists which has 4 to 8 papers, medium lists refer to lists with 9 to 15 papers (25-75 percentile), and long lists have at least 16 papers.
Figure 2.4 shows the citation pattern for lists of different lengths\(^7\). Short lists include only lists with 4 to 8 papers; medium lists refer to lists with between 9 and 15 papers, which account for papers with length between 25 percentile and 75 percentile of the whole sample. We have the following observations. First, the citation pattern in Fact Three is mostly caused by long lists and bottom positions for short lists. For long lists at top positions, the citation distribution of D submissions stochastically dominates that of ND submissions within 1% significance level. For both short and long lists at bottom positions, the citation comparison is significant within 2% significance level. To check whether the citation pattern for bottom positions of short lists is caused by reasons other than the strategic interaction between authors and readers in response to readers’ limited attention, we first need to figure out the reasons for these lists to be shorter. Since we deleted submissions during holidays, we base the analysis on different weekdays. In Table 2.3, we number lists according to the weekday. List 1 refers to lists submitted between Monday 4pm and Tuesday 4pm, list 2 for Tuesday and Wednesday, and so on. Table 2.3 shows almost all short lists are lists other than list 5 and most lists are list 2, 3 or 4. Table 2.3 also shows list 1 and 5 have a higher number of papers than list 2, 3 and 4. Figure 2.5 shows even though list 2 and 3 are short, there are more D submissions before 4pm, especially for list 3. The reason that authors try to submit before 4pm is probably they want their paper to be announced earlier. For example, if they submit after 4pm on Thursday, the paper will be announced on Sunday. If they submit before 4pm on Thursday, the paper will be announced on the same day. Therefore, the citation advantage of D submissions before 4pm for short lists may be caused by the deadline effect,

\(^7\)The same pattern still holds if we only consider top 1/bottom 1 position for short lists, top 2/bottom 2 positions for medium lists, and top 3/bottom 3 positions for long lists.
that is, papers delayed until the last minutes are on average better. However, when we eliminate submissions on Thursday, a similar citation pattern for bottom positions still remains. We can conclude that deadline effect is not the driving force for the citation advantage of D submissions at bottom positions. Second, for other positions, by Ranksum test, the citation comparison is not significant within 5% significance level, even though for top 2 positions of medium lists, it is significant within 8% significance level. Third, the citation comparison is only significant at both top and bottom positions for long lists. Fourth, the longer the lists are, the more likely authors self referee. Finally, there exists non-linearity for bottom positions, that is, only for short and medium lists, authors self referee.

Figure 2.5: D submissions before and after 4pm per weekday. There are a highest mean number of D submissions before 4pm on Thursday.

We propose that authors’ different submission patterns for lists of different lengths are due to readers’ different visibility constraints. For long lists, readers pay more attention to both top and bottom positions. This argument for long lists is similar to what Hirshleifer,
Table 2.3: Table 3. Length and Announcement Lists

<table>
<thead>
<tr>
<th>List</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short Lists</td>
<td>20</td>
<td>31</td>
<td>24</td>
<td>47</td>
<td>3</td>
<td>125</td>
</tr>
<tr>
<td>Length (median)</td>
<td>13</td>
<td>11</td>
<td>12</td>
<td>10</td>
<td>17</td>
<td>13</td>
</tr>
</tbody>
</table>

Ranksum Test 1

<table>
<thead>
<tr>
<th>z-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2.41</td>
<td>0.0159</td>
</tr>
<tr>
<td>-1.243</td>
<td>0.2138</td>
</tr>
<tr>
<td>-4.805</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

Ranksum Test 2

<table>
<thead>
<tr>
<th>z-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>-10.292</td>
<td>0.0000</td>
</tr>
<tr>
<td>-9.636</td>
<td>0.0000</td>
</tr>
<tr>
<td>-11.54</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

Note: Short lists are lists with at least 4 and at most 8 papers. We number lists according to the submission weekday. List 1 refers to the list submitted between Monday and Tuesday, list 2 for Tuesday and Wednesday, and so on. This table shows most short lists are list 2, 3 and 4. List 1 and 5 on average have more papers than other lists. Ranksum test 1 takes List 1 as group 1 and other lists as group 0. Ranksum test 2 takes List 5 as group 1 and other lists as group 0.
Lim and Teoh [2006] find for the financial markets. They conclude that investors are most
distracted when there are a lot of earning news on any given day. The relative attention to
top positions (relative to positions in the middle of any list) increases with the length of the
lists. Readers’ relative attention to bottom positions is strong for both short and long lists,
weak for medium lists. Authors choose the submission date according to their expectation
about the length of the lists. For example, list 5 is much longer than other lists. in response
to readers’ limited attention for different lists, authors adjust their submission behavior.

2.3 The Publishing on Journals and Fact Three

In this section, we will show that the publishing on journals does not have a significant
citation impact, which implies that only the information on all readers’ evaluation of papers
through arXiv.org is incorporated in the number of citations. Therefore, the number of ci-
tations reveal information on the strategic interaction between authors and readers purely
through arXiv.org. If D submissions are more likely to be published on journals and journals
have a significant citation impact on readers, the citation pattern in Fact Three is mostly
caused by the influence of journals. When authors submit to arXiv.org, they know what
journal they will submit to. Their submission times are driven mostly by reenforcing the in-
fluence of journals on the citing behavior of readers. For simplicity, in our following analysis,
we choose only 4 journals, i.e., Journal of High Energy Physics (JHEP), Nuclear Physics B
(NPB), Physics Letters B (PLB) and Physics Review D (PRD). Figure 2.6 shows D submis-
sions are more likely to be published in journals, especially these four main journals. These
four journals are expected to have more influence on readers than other journals because they dominate peer review in the field of high energy physics-theory. Table 2.4 shows the peer review in this field during the given time window is very concentrated. These four journals published over 63% of all papers submitted to arXiv.org in the field of high energy physics-theory between 2002 and 2004. Only four other journals each have more than 200 papers from this field, but each of them has less than 350 papers, in comparison to at least 800 papers for each of the four main journals in concern. For over 87% of all journals, each have only less than 50 papers from this field.

Figure 2.6: D submissions are more likely to be published on journals, especially the four main journals: Journal of High Energy Physics (JHEP), Nuclear Physics B (NPB), Physics Letters B (PLB) and Physics Review D (PRD). The bars on the left show the fraction of D and ND submissions at Top 2 or Bottom 2 positions which are published in any journal. The bars on the right show the fraction of papers published in the four main journals among all papers published in journals.
Table 2.4: High Concentration in Peer Review

<table>
<thead>
<tr>
<th>No. of papers/journal</th>
<th>No. of journals</th>
<th>Total No. of papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;800</td>
<td>4*</td>
<td>4961 (63.1%)</td>
</tr>
<tr>
<td>(200,350)</td>
<td>4</td>
<td>1121 (14.3%)</td>
</tr>
<tr>
<td>(70,150)</td>
<td>10</td>
<td>928 (3.7%)</td>
</tr>
<tr>
<td>&lt;50</td>
<td>128</td>
<td>853 (10.8%)</td>
</tr>
</tbody>
</table>

Note: *the four journals are Journal of High Energy Physics (JHEP), Nuclear Physics B (NPB), Physics Letters B (PLB) and Physics Review D (PRD). All papers are submitted to arXiv.org in the field of hep-th (high energy physics-theory) between 2002 and 2004.

Table 2.5: Summary Statistics for Main Variables

<table>
<thead>
<tr>
<th>Variables</th>
<th>$Q_1$</th>
<th>$Q_2$</th>
<th>$Q_3$</th>
<th>max</th>
<th>No. of papers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$PubMon^*$</td>
</tr>
<tr>
<td>PM (overall)</td>
<td>3</td>
<td>4</td>
<td>6</td>
<td>103</td>
<td>4175</td>
</tr>
<tr>
<td>PM (JHEP)</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>103</td>
<td>1498</td>
</tr>
<tr>
<td>PM (NPB)</td>
<td>4</td>
<td>5</td>
<td>7</td>
<td>48</td>
<td>788</td>
</tr>
<tr>
<td>PM (PLB)</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>75</td>
<td>777</td>
</tr>
<tr>
<td>PM (PRD)</td>
<td>4</td>
<td>5</td>
<td>7</td>
<td>92</td>
<td>1112</td>
</tr>
<tr>
<td>Cites</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>19</td>
<td>1588</td>
</tr>
</tbody>
</table>

Note: $*PubMon$ is the month of publishing on journals after the publication on arXiv.org.
I. Data

To test the citation impact of publishing on journals, we choose papers submitted to arXiv.org (hep-th) between 2002 and 2004 which are also published in the four main journals. There are two variables: the monthly number of citations and the publication month on journals. Data come from 3 sources: arXiv.org, INSPIRE and journal websites. The monthly number of citations is collected from INSPIRE. The journal information for each paper is retrieved from arXiv.org. The publication months for papers on JHEP are directly constructed from the information on the volume and issues. For all other 3 journals, we manually collected the publication information from the journal websites. If one paper has information on the time of availability online, then we use that month as the publishing month. Only PRD announces forthcoming papers. Therefore, we subtract the publishing month by one for papers on PRD. We choose papers which got published on journals after arXiv and started accumulating citation after publishing on arXiv. If any paper only has information on the year of publishing, then we assume this paper is published in June of that year.

We denote $PubMon$ as the publication month on journals after the publication month on arXiv. Table 2.5 shows $PubMon$ for all journals. For the following analysis, we further choose papers published at least on the 5th month after the publication month on arXiv. Reasons for this restriction will be explained soon afterwards. Both for all journals and for each journal, between 50 percent and 75 percent papers are included in the sample. There are in total 1588 papers. Table 2.5 also shows the distribution of the monthly number of citations in the sample.
II. Before-and-after Analysis

Figure 2.7: The monthly mean number of citations before and after the publishing on journals. The left figure is for papers on all four journals. The right figure shows data for each journal separately. For both figures, there is not significant change of the shape of the citation path. Data source: inspirehep.net, arXiv.org and journal websites

In this section, we will estimate with panel negative binomial regression. The dependent variable is the monthly number of citations $cite_{it}$, for $t \in \{1..12\}$. We choose citations four months before and eight months after the publication on journals. In consideration of Table...
2.5, we choose four months because we need enough information on the citation path before the publication on journals and also we want to make sure the sample size for the number of papers is not too small. And we give journals more time to accumulate citations. Figure 2.7 shows the mean monthly number of citations 4 months before and 8 months after the publishing on journals for all journals and for each journal separately. We can see that, there is no significant change of the citation path.

The overdispersion parameter for paper $i$ follows

$$
\lambda_{it} = \alpha_0 + \alpha_1\text{pub}_{it} + \alpha_2\text{pub}_{it}d_i + \sum_{k=1}^{r} \alpha_{3k}t^k + \sum_{k=1}^{r} \alpha_{4k}t^kd_i + \alpha_i + \epsilon_{it}
$$

Where

- $\text{pub}_{it} = \begin{cases} 
0 & \text{if } t \leq 4 \\
1 & \text{if } t > 4 
\end{cases}$

- $d_i = \text{PubMon}_i - 5$ measures how early the paper get published (after arXiv);

- $\alpha_i$ captures paper fixed effect, the possibility that $\text{cov}(\alpha_i, d_i) \neq 0$

- $\text{cov}(\epsilon_{it}, d_i) = 0$ and $\text{cov}(\epsilon_{it}, \text{pub}_{it}) = 0$.

Our regression controls for the citation path. We assume the citation paths before and after the publishing on journals have the same shape. The citation path can be influenced by $d_i$. Therefore, we put the interaction term in the controls. If there is citation impact of publication on journals, we will expect $\alpha_1 + d_i\alpha_1 \neq 0$ for some $d_i$. The results in Table 2.6 shows that for all journals, the coefficient on publication time is not significantly different from zero, even having a negative sign. The coefficient on the interaction between publication
time and $d_i$ is also not significant. Therefore, there is no significant citation impact of publication on journals. Table 2.7 shows the results for each journal, controlling for a 3rd-order polynomial and the interaction of both publication time and the polynomial with $d_i$. As we can see, we get similar results as column (4) in Table 2.6.

<table>
<thead>
<tr>
<th>Regressors</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r = 1$</td>
<td>-0.031</td>
<td>-0.022</td>
<td>-0.032</td>
<td>-0.059</td>
</tr>
<tr>
<td>$r = 3$</td>
<td>(0.036)</td>
<td>(0.048)</td>
<td>(0.049)</td>
<td>(0.055)</td>
</tr>
<tr>
<td>$p_{ub_t} \times d_i$</td>
<td>0.004</td>
<td>0.013</td>
<td>(0.004)</td>
<td>(0.009)</td>
</tr>
<tr>
<td>Poly($r,t$)</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Poly($r,t \times d_i$)</td>
<td></td>
<td></td>
<td>×</td>
<td></td>
</tr>
<tr>
<td>P-value</td>
<td>0.419</td>
<td>0.164</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Chi-Square Statistic)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$N$</td>
<td>1588</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: (1) $t \in \{1, ..., 12\}$; (2) *P-value of Chi-Square Statistic is for coefficients of pub and pub interacting $d$ with the biggest $d(70)$. 
III. Difference-in-Difference Model

In this section, we will use difference-in-difference model to answer the hypothetical question: if a paper were not published on any journal, would this paper generate the same citation path? More specifically, we choose papers published on journals on the $d + 5$th month after

---

### Table 2.7: Citation Impact of Publication on Journals (per journal)

<table>
<thead>
<tr>
<th>Regressors</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{pub}_{it}$</td>
<td>$-0.035$</td>
<td>$-0.024$</td>
<td>$-0.148$</td>
<td>$-0.066$</td>
</tr>
<tr>
<td></td>
<td>$(0.103)$</td>
<td>$(0.116)$</td>
<td>$(0.164)$</td>
<td>$(0.1)$</td>
</tr>
<tr>
<td>$\text{pub}<em>{it} \times d</em>{i}$</td>
<td>$0.019^+$</td>
<td>$-0.036$</td>
<td>$0.025$</td>
<td>$0.01$</td>
</tr>
<tr>
<td></td>
<td>$(0.011)$</td>
<td>$(0.033)$</td>
<td>$(0.046)$</td>
<td>$(0.025)$</td>
</tr>
<tr>
<td>Poly(3,t)</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Poly(3,t) $\times d_{i}$</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>P-value</td>
<td>0.1605</td>
<td>0.2649</td>
<td>0.6103</td>
<td>0.7277</td>
</tr>
<tr>
<td>(Chi-Square Statistic)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$N$</td>
<td>4320</td>
<td>5544</td>
<td>2724</td>
<td>6468</td>
</tr>
</tbody>
</table>

Note: (1) $t \in \{1, \ldots, 12\}$; (2) Chi-Square Statistic (for JHEP) is for coefficients of $\text{pub}$ and $\text{pub}$ interacting $d$ and $\text{pub}$ interacting $d$ with the biggest $d(70)$; (3) Chi-Square Statistic (for all other journals) is for coefficients of $\text{pub}$ and $\text{pub}$ interacting $d$ with $d = 10$, 90 percentile of the distribution of $d$. 

---
their publishing on arXiv.org as the treatment group and papers published on journals at least \(d + 10\)th month after the publishing on arXiv.org as the control group, where \(d \in \{0, 1, 2, 3\}\). For the treatment group, we choose the citation path as the monthly number of citations for 4 months before and 4 months after the publishing on journals; for the control group, we choose the citation path during the same time frame as the citation path for the treatment group, that is \(t \in \{d + 1, d + 2, d + 3, d + 4\} \cup \{d + 6, d + 7, d + 8, d + 9\}\). We run the following regression.

\[
\log mcite_{it} = \alpha_0 + \alpha_1 d_i + \alpha_2 group_i + \alpha_3 pub_{it} + \alpha_4 pub_{it} \times group_i +
\log mc_i \times (\alpha_5 + \alpha_6 pub_{it} + \alpha_7 pub_{it} \times group_i) + \epsilon_{it}
\]

where

- \(mcite_{it} = cite_{it} + 1\) and \(mc_i = c_i + 1\), where \(c_i\) is the sum of numbers of citations for \(t \in (d + 1...d + 4)\);

- \(group_i = 1\) if paper \(i\) belongs to the treatment group and 0 otherwise;

- \(pub_{it} = 1\) if \(t > d_i + 5\) and 0 otherwise;

If there is a significant citation impact of publishing on journals, we should expect \(\alpha_4 + \alpha_7 \log mc_i > 0\) for some paper \(i\). Table 2.8 shows the results for papers on all all four journals and all \(d_s\), where \(d \in \{0, 1, 2, 3\}\). We do not find a significant citation impact of publishing on journals. The result in Column (1) is consistent with Figure 2.8. The negative coefficient of \(pub_{it}\) shows for the control group, the monthly number of citations decreases with time. The negative coefficient of \(pub_{it} \times group_i\) shows publishing on journals has a surprisingly
Figure 2.8: There is no positive citation impact of publishing on journals. The treatment group includes only papers published on journals on $d + 5$th month after the publishing on arXiv.org and the control group includes only papers published on journals at least $d + 10$th month after the publishing on arXiv.org, where $d \in \{0, 1, 2, 3\}$. $d$ measures how early papers gotten published on journals after their publishing on arXiv.org.

negative effect on the number of citations in comparison to the control group, even though it is not significant within 10% significance level. The positive coefficient of $group_i$ shows the control group in general is better cited. The negative coefficient of $d_i$ shows papers published earlier on journals are cited more. In column (2), when we control for the total number of citations for the 4 months before the publishing on journals of the control group, not surprisingly, the coefficients of $pub_{it}$ and $pub_{it} \times group_i$ are not changed. The coefficient of $group_i$ becomes much smaller and not significant, which shows the two groups of papers have no significant citation difference. In column (3), we check whether and how the citation path before the publishing on journals of the control group influences the result. We can conclude the negative coefficient of $pub_{it} \times group_i$ is mostly caused by papers which are not
cited much before the publishing on journals of the control group, which can be shown in the negative coefficient of $pub_{it} \times group_i$ and the positive coefficient of $logmc_i \times pub_{it} \times group_i$.

In Table 2.9 and 2.10, we do separate analysis for different journals. Overall, we get the similar results. The citation impact may depend on how early papers are gotten published, that is, the distance ($d$). Therefore, we separate the case of $d = 0, 1$ and $d = 2, 3$. For the first case, we show both the results for all journals and for each journal. For all columns, neither the coefficient of $pub_{it} \times group_i$ nor the coefficient of $logmc_i \times pub_{it} \times group_i$ is significantly different from zero within 10% significance level except for column (4). In column (4), for $logmc$ below 75 percentile of all samples, there is no significant citation impact of publishing on PLB (the p-value is 0.176). In Table 2.10 when we do analysis with $d = 2$ and 3, both coefficients in column (4) are not significant within 10% significance level. In column (5) of Table 2.10, even though the coefficient of $pub_{it} \times group_i$ is significantly positive within 10% significance level, the column (5) of Table 2.9 shows no significant citation impact. Moreover, in Table 2.10, the ward test shows for the citation impact is not significant as long as some papers are cited during the first four months before the publishing on journals.

2.4 Literature

All previous literature are from physicist Dietrich [2008,a,b] and Hague and Ginsparg [2009, 2010]. Dietrich [2008, a] first noticed that, for astroph (a sub-archive on arXiv.org for

---

Paul Ginsparg is physicist who has a joint position in the department of information science and physics at Cornell University. Hague is his student from the department of information science.
Table 2.8: Citation Impact of Publication on Journals (all journals)

<table>
<thead>
<tr>
<th>Dependent Variable: logmcite</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Regressors</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>pubit</td>
<td>$-0.010416^{**}$</td>
<td>$-0.010416^{**}$</td>
<td>$0.05368^{**}$</td>
</tr>
<tr>
<td></td>
<td>$(0.0036036)$</td>
<td>$(0.0036037)$</td>
<td>$(0.0041349)$</td>
</tr>
<tr>
<td>pubit $\times$ groupi</td>
<td>$-0.0069811$</td>
<td>$-0.0069811$</td>
<td>$-0.001235$</td>
</tr>
<tr>
<td></td>
<td>$(0.0050422)$</td>
<td>$(0.0050423)$</td>
<td>$(0.006217)$</td>
</tr>
<tr>
<td>groupi</td>
<td>$0.0202558^{**}$</td>
<td>$0.001876$</td>
<td>$-0.0027139$</td>
</tr>
<tr>
<td></td>
<td>$(0.0069057)$</td>
<td>$(0.0023909)$</td>
<td>$(0.0020448)$</td>
</tr>
<tr>
<td>di</td>
<td>$-0.0051399^{+}$</td>
<td>$-0.0015743$</td>
<td>$-0.0015665$</td>
</tr>
<tr>
<td></td>
<td>$(0.0027938)$</td>
<td>$(0.001263)$</td>
<td>$(0.001262)$</td>
</tr>
<tr>
<td>logmc[i]</td>
<td>$0.3700865^{**}$</td>
<td>$0.4593985^{**}$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$(0.0062257)$</td>
<td>$(0.0059935)$</td>
<td></td>
</tr>
<tr>
<td>logmc[i] $\times$ pubit</td>
<td></td>
<td></td>
<td>$-0.1837325^{**}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$(0.0125908)$</td>
</tr>
<tr>
<td>logmc[i] $\times$ pubit $\times$ groupi</td>
<td></td>
<td></td>
<td>$0.0092421$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$(0.0176337)$</td>
</tr>
</tbody>
</table>

N (papers) 2433

Note: all standard errors are clustered at the paper level. $mcite_i = cite_i + 1$ and $mc_i = c_i + 1$, where $cite_i$ is the month number of citations and $c_i$ is papers’ total number of citations for the first 4 months after the publishing on arXiv.org.
### Table 2.9: Citation Impact of Publication on Journals (per journal)

<table>
<thead>
<tr>
<th>Regressors</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>all journals</td>
<td>JHEP</td>
<td>NPB</td>
<td>PLB</td>
<td>PRD</td>
</tr>
<tr>
<td>$pub_{it} \times group_{i}$</td>
<td>-0.0059046</td>
<td>-0.0207387</td>
<td>-0.0214448</td>
<td>-0.0253572^+</td>
<td>0.0123328</td>
</tr>
<tr>
<td></td>
<td>(0.0078347)</td>
<td>(0.0201357)</td>
<td>(0.0149177)</td>
<td>(0.0140649)</td>
<td>(0.0108386)</td>
</tr>
<tr>
<td>$logmc_{i} \times pub_{it} \times group_{i}$</td>
<td>0.0245348</td>
<td>0.0405789</td>
<td>0.0665162</td>
<td>0.1139114^*</td>
<td>0.0096206</td>
</tr>
<tr>
<td></td>
<td>(0.0218841)</td>
<td>(0.0470349)</td>
<td>(0.041607)</td>
<td>(0.0509678)</td>
<td>(0.0304247)</td>
</tr>
<tr>
<td>$N$ (papers)</td>
<td>1592</td>
<td>362</td>
<td>435</td>
<td>242</td>
<td>553</td>
</tr>
<tr>
<td>Ward Test (p-value)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.176</td>
</tr>
</tbody>
</table>

Note: $d_i = 0$ or 1. All standard errors are clustered at the paper level. The Ward test in the top table is for papers at 75 percentile of $logmc_i$ for all $i$ on PLB.
Table 2.10: Citation Impact of Publication on Journals (per journal)

<table>
<thead>
<tr>
<th>Dependent Variable: ( \log mcite )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regressors</td>
</tr>
<tr>
<td>all journals</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>( pubit \times group_i )</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>( \log mc_i \times pubit \times group_i )</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>( N ) (papers)</td>
</tr>
<tr>
<td>Ward Test (p-value)</td>
</tr>
</tbody>
</table>

Note: \( d_i = 2 \) or \( 3 \) per journal. All standard errors are clustered at the paper level. The Ward test in the bottom table is for papers at 50 percentile of \( \log mc_i \) for all \( i \) on PRD. 50 percent of papers have no citation at all.
astrophysics), some authors submitted seconds after 4pm. Dietrich [2008, b] further chose only papers with the first author’s affiliation located in North or South America. He showed that, for papers at the top 3 positions, papers submitted within 5 minutes after 4pm have more citations than those submitted more than 90 minutes away from 4pm. With submissions from all regions in high energy theory, Hague and Ginsparg [2009, 2010] show that from 2002 to 2004, at the first or last position, papers submitted within 20 minutes before or after 4pm have more citations.

Our work confirms and extends their work from a different perspective. We view high energy physicists as economic agents with incentives. We are interested in whether there exist strategic interactions between authors and readers, and more specifically, whether, without formal refereeing, authors automatically reveal information about the quality of their papers. The analysis in the literature is not enough to serve our interests for the following reasons.

First, we explicitly show there exists strategic submission behavior, while they only show the aggregate submission pattern under 4pm deadline. We cannot tell the existence of strategic submissions from the submission pattern in their analysis. It is possible the aggregate submission pattern follows from each author’s random schedule. The aggregate random submissions center around 4pm because most of the submissions from NA are from the eastern time zone and 4pm is close to the end of the working day. Authors want to submit the paper when they feel tired for the whole day and before they leave for dinner. Moreover, we study the submission incentives of the submitter. Dietrich [2008, b] chose papers based on the first author. But the first author is not necessarily the submitter. Her incentives are possibly not aligned with the submitter.
Second, we study the incentives not only in D submissions, but more importantly in ND submissions. We emphasize the role of strategic interactions between authors and readers in ND submissions. We use the data from the US or Canada to reduce the heterogeneity among authors and the influence of their physical submission cost. However, except for Dietrich [2008, b], all the literature use data from all regions. For a lot of submitters, there exists significant submission cost. For example, 4pm eastern time is night time at Asia. Asian submitters submit away from the deadline may be only because of the too high physical cost of submitting around 4pm.

Third, we choose the data in the field of high energy physics-theory so that the citation pattern in Fact Three is caused by the strategic interaction between authors and readers rather than by the influence of journals. Dietrich [2008, a, b] only uses the data from astro-ph. But, in this field, between 2002 and 2004, journals still play an important evaluation role. Many authors wait for the acceptance to, or publishing in, journals before their submission to arXiv.org. When they submit to arXiv.org, they usually put comments on arXiv.org to indicate the publishing status of their paper. We can infer readers’ citing decision is heavily influenced by the publishing status of papers on journals. Authors’ submission times may be driven by their desire to reinforce the influence of journals on the readers’ citing behavior.
Chapter 2: The Model of Self-refereeing

In this chapter, we will build a game-theoretical model to match the motivating facts as identified in Chapter 1. The three facts are as follows. First (Fact One), there are strategic submissions both around the deadline and away from the deadline\(^1\). Second (Fact Two), almost all submissions around the deadline get top 2 or bottom 2 positions, and at top 2 and bottom 2 positions, a high fraction of papers are submissions away from the deadline. Finally (Fact Three), at top 2 or bottom 2 positions, papers submitted around the deadline are cited more than papers submitted away from the deadline.

We show that authors self-referee, that is, among authors of the same quality, those with good papers are more likely to submit around the deadline. Reputation effect is satisfied, that is, among authors with papers of the same quality, good authors are more likely to submit away from the deadline. So many papers are submitted away from the deadline that Fact Two holds. So high fraction of good papers are submitted around the deadline that papers submitted around the deadline are on average better. Since readers cite based on their estimation of paper quality at any position, Fact Three holds.

Suppose readers pay more attention to papers at the top or bottom positions. Self-refereeing can only be sustained by reputation effect. Given reputation effect, the more limited readers’ attention to positions in the middle of the list, the stronger authors’ incentive

---

\(^1\)In terms of the data in Chapter 1, submissions around the deadline mean papers submitted within 30 minutes before or after 4pm. Submissions away from the deadline are other submissions. The deadline does not only mean 4pm, but more generally refers to any deadline.
to self-referee. If there is no reputation concern, everyone submits around the deadline to reduce the risk of being missed by readers. With reputation concerns, even though anyone submitting away from the deadline suffers now from the lower number of citations resulting from lower average paper quality as well as less visibility at positions in the middle of the list, she gain from the reputation improvement in the future. With papers of the same quality, good authors benefit more from reputation effect because they are more likely to have good papers in the future and the citation advantage of good papers relative to bad papers is higher if their authors invested in reputation. Given reputation effect, authors self-referee because good papers have a higher cost of being place in the middle of any list and/or the reputation gain is smaller because good papers are more likely to be from good authors. Since good papers have a higher cost of being missed by readers, the more limited readers’ attention is, the higher good papers’ relative cost of being placed in the middle of the list.

In Chapter 3, we find that, only for intermediate reputations, both self-refereeing and reputation effect exist. Moreover, all authors self-referee and bad authors are more likely to self-referee, which implies both bad authors with bad papers and good authors with good papers invest in reputation with a high probability. It suggests readers pay more attention to the top and bottom positions. The more bad authors with bad papers invest in reputation, the smaller reputation gain for authors with bad papers. But the more good authors with good papers invest in reputation, the bigger reputation gain for authors with good papers. Therefore, authors self-referee only when papers at positions in the middle of the lists are very likely to be missed by readers.

The intermediacy of reputations as implied by results from Chapter 3 is not surprising. As readers become increasingly sure the author is a good author, the reputation gain diminishes,
which implies everyone has a strict incentive to submit around the deadline. Moreover, bad authors have a strict incentive to invest in reputation, which destroys reputation effect. As readers are increasingly sure the author is a bad author, the reputation gain for authors with bad papers also diminishes and hence they all submit around the deadline.

We analyze this two-period model by backward induction. Proposition 1 shows the unique equilibrium in the second period, which implies reputation concern is the driven force underlying the motivating facts and we need to match the equilibrium in the first period to the motivating facts. Proposition 2 and 4 are implications about the observable variables, which will guide our empirical test in Chapter 3. Proposition 2 shows any equilibrium consistent with the motivating facts must satisfy both self-refereeing and reputation effect. Proposition 4 shows both good authors with good papers and bad authors with bad papers randomize only for intermediate reputations. To show the existence of an equilibrium consistent with the motivating facts and illustrate the intuition for both self-refereeing and reputation effect, in Proposition 3, we analyze the equilibrium incentives and conditions for the unique pure strategy equilibrium which shows the highest degree of self-refereeing (in terms of good authors) and the strongest reputation effect (in terms of bad papers). Based on the unique pure strategy equilibrium, we describe the incentives for the equilibria consistent with the results from Chapter 3. We further show the existence of such an equilibrium numerically.
3.1 The Model

There are an unit mass of authors and a unit mass of readers\(^2\). There are two periods, \(t = 1\) and \(t = 2\). In period 1, each author submits a paper and the paper gets a position. If readers see the paper, they decide whether to cite or not. At the beginning of period 2, readers get citation information from all readers for all papers in period 1 and each author writes a new paper. The rest of period 2 is the same as in period 1. Figure 3.1 illustrates the timeline.

Each author is either good (type \(G\)) or bad (type \(B\)). Each paper is either good with quality \(\theta_H\) or bad with quality \(\theta_L\), where \(\theta_H > \theta_L\). At the beginning of period 1, the proportion of good authors is \(q \in (0, 1)\). In each period, a good author (bad author) writes a high quality paper with probability \(p\) \((1 - p)\), and a low quality paper with probability \(1 - p\).

\(^2\)We make this assumption without loss of generality. In the real world, there is an unit mass of researchers. Each researcher is both an author and a reader. For his paper, he is the author and all other researchers are the readers. For other researchers' paper, he is the reader and all researchers other than the paper's author are all readers. We separate authors and readers because the role played by each researcher as an author or a reader is independent.
We assume $p > 1/2$, which means good authors are more likely to have good papers and bad authors are more likely to have bad papers. The relationship between author quality and paper quality is shown in Figure 3.2.

![Figure 3.2: The probabilities of good (bad) authors to have good or bad papers. We assume $p > 1/2$, which implies good authors are more likely to have good papers and bad authors are more likely to have bad papers.](image)

Authors can submit around the deadline (D) or away from the deadline (ND) at no physical cost. Without loss of generality, we assume D and ND submissions are papers on the same list. If the deadline is 4pm eastern time, D means submitting within 30 minutes after 4pm on one working day or within 30 minutes before 4pm on the next working day. Submissions between 4pm on the first working day and 4pm on the next working days are ND submissions. Each paper submitted will get either higher visibility ($HV$) or lower visibility ($LV$) position. $HV$ positions correspond to top or bottom positions and $LV$ positions correspond to positions in the middle of the list. Following the well-accepted visibility effect in psychology as proposed by Ebbinghaus (1913), we assume readers exogenously pay more attention to top/bottom positions and less attention to positions in the middle of the list.
That is to say, each paper at any LV position has probability $\epsilon$ of being missed by each reader independently. Exactly $\eta \in (0, 1)$ fraction of all papers submitted will have HV positions. Accordingly, LV positions hold the rest, $1 - \eta$ fraction of papers. D submissions have the priority for HV positions. More specifically, if less than $\eta$ fraction of papers are submitted around the deadline (D), all of these papers are assigned HV; the rest of the HV positions are filled randomly with papers submitted away from 4pm (ND). Similarly, if the fraction of D submissions is greater than $\eta$, some of these papers are randomly assigned position LV (after all HV positions are filled with papers from D submission). The rest of the LV positions are filled with papers submitted away from the deadline (ND). Let $p_d$ the probability that a D submission will have position LV, and $p_{nd}$ the probability that a ND submission having position HV. Then, $p_dp_{nd} = 0$. Figure 3.3 shows the relationship between submission and position.

There exists asymmetric information between authors and readers. We assume authors know their own author quality and paper quality. Readers do not know authors’ quality,
but each readers independently gets noisy signals about the quality of the papers which are distributed according to
\[ \hat{\theta} = \theta_i + \sigma \xi \]
where \( i \in \{H, L\} \), \( \sigma > 0 \) and \( \xi \sim N(0, 1) \). Denote \( f(g) \) as the density function of \( \hat{\theta} \) with mean \( \theta_H \) (\( \theta_L \)). Readers cannot see the submission time of any paper, but if they do not miss the paper, they can see the position directly. Even if they missed the paper in period 1, they can infer the paper had LV position in period 2.

Authors and readers have different incentives. Authors maximize expected number of citations which is measured by the fraction of readers authors expect to cite their paper. More specifically, if his paper is cited by \( \alpha \) proportion of readers, the author’s utility is \( \alpha \). Readers, however, only want to cite good papers. A reader’s utility is \( \theta_H \) if he cites a good paper and \( \theta_L \) if he cites a bad paper. Since his reservation value is \( \theta^* \) and satisfies \( \theta_L < \theta^* < \theta_H \), there is a conflict of interests between the readers and authors when both know that the paper is bad. Define \( R = \frac{\theta^* - \theta_L}{\theta_H - \theta_L} \). \( R \) measures the reader’s riskiness of citing a paper. Assume \( 1 - p < R < p \). That is, if readers only know the authors’ quality, they only cite good authors’ paper. Moreover, even though readers do not have any long term concerns, authors have and discount the future payoff at a rate of \( \beta \). Hence, the parameters of this strategic interaction are \( (\epsilon, \eta, \sigma, \beta, p, \theta_i, \theta^*, i \in \{H, L\}) \).

Given the setup of the model, we can describe the strategies for both authors and readers. Each author’s strategies are \((s^1_a, s^2_a)\) such that \( s^1_a : \{\theta_H, \theta_L\} \to [0, 1] \) and \( s^2_a : [0, 1] \times \{\theta_H, \theta_L\} \to [0, 1] \), where \( a = G, B \) is the author’s type, \( s^1_a(\theta^1) \) is the probability that a type-\( a \) author will submit a quality-\( \theta^1 \) paper around the deadline in period 1 while \( s^2_a(z, \theta^2) \) is
the probability that a type-$a$ author will submit a quality-$\theta^2$ paper around the deadline in period 2 given that the public assigns probability $z$ to him being a good author at the beginning of period 2. Each reader’s strategy in the first period is $s_r^1 : \{HV, LV\} \times (-\infty, +\infty) \to [0, 1]$. Given the position and the noisy signal from the paper, readers decide whether to cite or not. In the second period, $s_r^2 : \{\{HV, LV\} \times (-\infty, +\infty), \emptyset\} \times [0, 1] \times \{HV, LV\} \times (-\infty, +\infty) \to [0, 1]$. Any reader, if he sees the paper, knows the paper’s position and gets the noisy signals at $t = 2$, but his decision to cite the paper or not also depends on the author’s publication history. At $t = 1$, if he didn’t miss this author’s paper, he saw the position and got noisy signals about the quality of the paper. At the beginning of $t = 2$, he also knows the proportion of readers citing this paper from the first period.

Denote reader’s period $t$ information other than noisy signals as $I_t$ for $t \in \{1, 2\}$. Let $\mu_{I_t} = p(\theta_H \mid I_t)$. So, $\mu_{I_t}$ is a reader’s belief about the paper’s quality (i.e., the probability that he assigns to the paper being good) before observing the noisy signal.

**Lemma 1** In each period, any reader’s optimal strategy is cutoff strategy. That is, there is $\bar{\theta}(\mu_{I_t})$ such that $s_r^1(\mu_{I_t}, \hat{\theta}) = 1$ for any $\hat{\theta} > \bar{\theta}(\mu_{I_t})$ and $s_r^1(\mu_{I_t}, \hat{\theta}) = 0$ for any $\hat{\theta} < \bar{\theta}(\mu_{I_t})$.

**Proof:** The expected quality of the paper given $\mu_{I_t}$ and signal $\hat{\theta}$ is

$$E(\theta \mid \mu_{I_t}, \hat{\theta}) = \frac{\theta_H \mu_{I_t} f(\hat{\theta}) + \theta_L (1 - \mu_{I_t}) g(\hat{\theta})}{\mu_{I_t} f(\hat{\theta}) + (1 - \mu_{I_t}) g(\hat{\theta})}.$$ 

By MLRP and $\theta_H > \theta_L$, $E(\theta \mid \mu_{I_t}, \hat{\theta})$ strictly increases with $\hat{\theta}$. So, readers cite if $E(\theta_H \mid \mu_{I_t}, \hat{\theta}) > \theta^*$, and don’t cite if $E(\theta_H \mid \mu_{I_t}, \hat{\theta}) < \theta^*$.

□
Ceteris paribus, high quality papers are more likely to have good signals than low quality papers. So, given good signal, it is more likely to come from a high quality paper. Hence, papers with good signals are more likely to be cited.

By law of large numbers, the proportion of readers citing any paper with quality $\theta_i$ and position $HV$ is $Pr(\hat{\theta} > \bar{\theta}(\mu_i) \mid \theta_i)^3$. Similarly, for each paper with position $LV$, $1 - \epsilon$ fraction of readers read it. So, the proportion of readers citing a paper with position $LV$ is $(1 - \epsilon)Pr(\hat{\theta} > \bar{\theta}(\mu_i) \mid \theta_i)$. At $t = 2$, if any reader didn’t see some author’s paper at $t = 1$, he can infer that paper had position $LV$ in the first period and he missed it. By MLRP, $Pr(\hat{\theta} > \bar{\theta}(\mu_i) \mid \theta_H) > Pr(\hat{\theta} > \bar{\theta}(\mu_i) \mid \theta_L)$. So, given $\mu_I$, each paper’s quality is revealed. We can state this result formally below.

**Lemma 2** The quality of any first period paper is revealed at the beginning of the second period.

Therefore, in period 2, a reader, who sees the current paper, has the following information: he knows the position of the paper that the author submitted in the current period, he sees a signal about this paper, he knows the position of the paper that the author submitted in the previous period and he knows the quality of that paper. Below, we study the optimal strategy of a reader with this information. We will start from backward induction.

---

3More specifically, for papers with position $HV$, all readers can see it. Given each paper with quality $\theta_i$, every reader’s ex-ante probability of citing it is $Pr(\hat{\theta} > \bar{\theta}(\mu_i) \mid \theta_i)$, for all $i \in \{\theta_H, \theta_L\}$. We have one sequence of binary variables $(x_{im}^b)_{m=1}^{+\infty}$ distributed according to Bernoulli distribution with probability $Pr(\hat{\theta} > \bar{\theta}(\mu_i) \mid \theta_i)$. For $n$ readers, the proportion of readers citing it is $\frac{1}{n} \sum_{m=1}^{n} x_{im}^b$. By law of large numbers, as $n \to +\infty$, $\frac{1}{n} \sum_{m=1}^{n} x_{im}^b \to E(x^i) = Pr(\hat{\theta} > \bar{\theta}(\mu_i) \mid \theta_i)$. 

42
In this section, we will establish that in any equilibrium all papers will be submitted around the deadline in the second period. D (ND) submissions always have a positive probability of getting HV (LV) positions. Since \( \epsilon > 0 \) (i.e., since some fraction of papers at position LV get lost), all else being equal, D yields a higher payoff than ND. Since good papers are more likely to have good signals and hence more likely to be cited, they suffer more from the risk of being lost. Hence, authors with good paper have a strictly stronger incentive to choose D\(^4\). Now consider any group of authors with the same reputation in the second period. If any author submits a good paper away from the deadline, all bad papers must also be submitted away from the deadline. If some good papers are submitted around the deadline, papers from the same group at HV positions on average have better quality. Then, everyone has incentive to deviate to avoid the risk of being lost and also to get higher chance of being cited if not lost. If all good papers are submitted away from the deadline, no matter whether any paper from this group will get HV position or not, all authors have a strict incentive to deviate. For the first case, deviating means avoiding the risk of being lost at LV positions. For the second case, readers believe anyone deviating has a good paper. If all good papers are submitted around the deadline, papers at position HV are on average

\(^4\)We call this relationship as Crossing-Order Condition, which can be formally stated below and proved in Appendix A. Given all other information, denote readers’ strategy when he reads a paper with position HV as \( \bar{b}_{HV} \) and \( \bar{b}_{LV} \) when he reads a paper with position LV. Crossing-Order Condition is satisfied if \( 1 - F(\bar{b}_{HV}) > (1 - \epsilon)(1 - F(\bar{b}_{LV})) \) whenever \( 1 - G(\bar{b}_{HV}) \geq (1 - \epsilon)(1 - G(\bar{b}_{LV})) \), where \( F \) and \( G \) are the cumulative distribution functions for density \( f \) and \( g \) respectively.
better, so nobody wants to submit away from the deadline. Hence, all authors submit around the deadline. This observation is stated formally below.

**Proposition 1** *In the second period, all papers are submitted around the deadline.*

Proposition 1 establishes that, as long as there exists higher visibility, the motivating facts in Chapter 2 are driven by authors’ reputation concerns. If there is no reputation concern, all authors will choose higher visibility. Moreover, since in the second period, papers from authors with higher reputation are cited more, every author has a strict incentive to improve their reputation in the first period.

*The First Period (Authors Have Reputation Concerns)*

Given the observations made above, we can simplify our description of (candidate) equilibrium strategy profiles. First, since we know that all authors will submit all papers around the deadline in period 2, we suppress period 2 behavior in our description. Hence, a strategy for authors is four numbers \((x_{GH}, x_{GL}, x_{BH}, x_{BL})\), where \(x_{GH}\) (\(x_{BH}\)) is the probability with which a good (bad) author submits a high quality paper around the deadline in period 1 and \(x_{GL}\) (\(x_{BL}\)) are the corresponding probabilities for low quality papers.

In period 2, if the reader didn’t see the author’s paper in the first period, he infers that the paper had position \(LV\) and he missed it. The complete citation information informs him of this paper’s quality. So, for all cases, the reader knows the position of the first period paper \((s_1)\), the quality of the first period paper \((\theta_i)\) and based on this information forms his citation.
cutoffs $\bar{\theta}_{s_i}$ for $s_i = HV, LV$ and $i = H, L$. Hence, an equilibrium is described by ten numbers, $(x^G_H, x^G_L, x^B_H, x^B_L, \bar{\theta}^1_{HV}, \bar{\theta}^1_{LV}, \bar{\theta}^{2H}_{HV}, \bar{\theta}^{2H}_{LV}, \bar{\theta}^{2L}_{HV}, \bar{\theta}^{2L}_{LV})$. Denote $\bar{\theta}_I$ generally as readers’ cutoff under information $I_t$. With $\mu_{I_t} = p(\theta_H | I_t)$ and $R = \frac{\theta_H - \theta_L}{\theta_H - \theta_L}$, $\bar{\theta}_I$, satisfies $E(\theta_H | \bar{\theta}_I, \mu_{I_t}) = \theta^*$, which means

$$\bar{\theta}_I = \frac{\theta_H + \theta_L}{2} + \frac{\sigma^2}{\theta_H - \theta_L} \ln \left[ \frac{R}{1 - R} \left( \frac{1}{\mu_{I_t}} - 1 \right) \right]$$

where $I_1$ includes only information on papers’ positions and $I_2$ also incorporates complete citation information from the readers. Therefore, readers’ citing behavior in each period completely follows from authors’ strategy profile. Since, for each strategy profile of authors, each reader’s response is unique, we will suppress the reader behavior from our description of strategies and generally identify each strategy profile as authors’ strategy profile $(x^G_H, x^G_L, x^B_H, x^B_L)$.

Now we will interpret the motivating facts in terms of the observable variables and unobservable parameters for any strategy profile of this model. We will first do it generally and then illustrate with two examples.

First, Fact One means $\max \{ x^G_H, x^G_L, x^B_H, x^B_L \} > 0$ (some authors submit around the deadline) and $\min \{ x^G_H, x^G_L, x^B_H, x^B_L \} = 0$ (there exist authors submitting away from the deadline); second, Fact Two means $q(px^G_H + (1 - p)x^G_L) + (1 - q)((1 - p)x^B_H + px^B_L) < \eta$, that is, papers submitted around the deadline $(q(px^G_H + (1 - p)x^G_L) + (1 - q)((1 - p)x^B_H + px^B_L))$ are not enough to fill the top 2 and bottom 2 positions; finally, Fact Three is equivalent to $p(\theta_H | LV) < p(\theta_H | HV)$, which means papers at HV positions are on average better.
papers. This follows from the following argument. *Fact Three* means, at *HV* positions, D submissions have on average higher quality. Since ND submissions at *HV* positions are randomly chosen from all papers submitted away from the deadline, all ND submissions on average have the same quality no matter what positions they have. Therefore, papers at *LV* positions are on average worse than those at *HV* positions.

<table>
<thead>
<tr>
<th>author</th>
<th>G</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_H$</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$\theta_L$</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>author</th>
<th>G</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_H$</td>
<td>$x_H^G$</td>
<td>1</td>
</tr>
<tr>
<td>$\theta_L$</td>
<td>0</td>
<td>$1/2$</td>
</tr>
</tbody>
</table>

Figure 3.4: The top table shows the strategy profile $(1, 0, 1, 1)$ and the bottom table shows $(x_H^G, 0, 1, 1/2)$. Let $\eta = 0.7$ and $p = 0.65$. $(1, 0, 1, 1)$ is consistent with the motivating facts for any reputation satisfying $q > 6/7$. $(x_H^G, 0, 1, 1/2)$ with $x_H^G \leq 0.1$ is consistent with the motivating facts for any reputation $q$.

As a simple example, consider the strategy profile with good authors submitting around the deadline if and only if they have good papers, and bad authors submit all papers around the deadline. That is, $(x_H^G, x_L^G, x_H^B, x_L^B) = (1, 0, 1, 1)$ as shown in Figure 3.4. This strategy profile can be consistent with the facts: (1) Good authors with bad papers submit away
from the deadline and all others submit around the deadline (Fact One); (2) some papers submitted away from the deadline get HV positions if and only if \( q > \frac{1-p}{1-\eta} \) with \( \eta > p \) (Fact Two); and (3) Papers submitted away from the deadline are bad papers (Fact Three).

Let \( p = 0.65 \) and \( \eta = 0.7 \). If \( q > \frac{6}{7} \), Fact Two is satisfied and hence this strategy profile is consistent with the motivating facts. Now consider the strategy profile \((x_H^G, 0, 1, x_L^B)\) with \( x_H^G, x_L^B \in (0, 1) \). It is consistent with Fact One. It is also more likely to be consistent with Fact Two than \((1, 0, 1, 1)\) in the sense that whenever \((1, 0, 1, 1)\) satisfies Fact Two, \((x_H^G, 0, 1, x_L^B)\) with \( x_H^G, x_L^B \in (0, 1) \) also satisfies Fact Two. If \( x_L^B \) is sufficiently large relative to 1/2, it is also consistent with Fact Three. Under the same values of \( \eta \) and \( p \), any strategy profile satisfying \( x_L^B \) and \( x_H^G \leq 0.1 \) is consistent with Fact Three for all reputation \( q \). In particular, it shows a strategy profile satisfying \( x_H^G + x_L^B < 1 \) as we will find in Chapter 3 can be consistent with the motivating facts.

**Definition 1** The strategy profile \((x_H^G, x_L^G, x_H^B, x_L^B)\) involves self-refereeing if \( x_H^G > x_L^G \) or \( x_H^B > x_L^B \); it displays a reputation effect if \( x_i^G \leq x_i^B \) for all \( i \in \{H, L\} \) and \( \exists i \) such that \( x_i^G < x_i^B \).

We can interpret these two key concepts in the following way. If, among authors of the same quality, those with good papers are (strictly) more likely to submit around the deadline, we say authors of this quality (strictly) self referee their papers. A strategy profile displays self-refereeing if at least one type of authors self referee. The definition of self-refereeing is analogous to the formal refereeing for peer review journals. Both authors on arXiv.org and referees for journals are helping readers to evaluate papers. For each journal, given all
submissions, refereeing means publishing good papers and rejecting bad papers. Therefore, it is a 0-1 game, either readers can see the paper or not. For arXiv.org, submitting around the deadline or away from the deadline means papers are more likely or less likely to be seen by readers. The key difference between journals and arXiv.org is, submission and refereeing are through different people for journals, while for arXiv.org, both are through authors. We should separate our definition from authors’ selective submissions to different journals. Authors submit their best papers to good journals probably because peer review is very time-consuming rather than because of authors’ reputation concerns. If they submit a bad paper to a good journal, they know it will be very likely to be rejected. Since the refereeing is very slow and they have to wait to resubmit to a worse journal, submitting to a good journal only wastes their time. If, among authors with the same quality of papers, bad authors are always more likely to submit around the deadline, a reputation effect is displayed. By Lemma 2, at the beginning of the second period, the quality of the first period paper is revealed. Given the quality of first period papers, in the second period, readers infer papers at LV positions are more likely to be from good authors than those at HV positions. Therefore, given papers of the same quality in the first period, authors submitting away from the deadline gain from the reputation effect relative to those submitting around the deadline. It is easy to see any strategy profile \((x_H^G, 0, 1, x_L^P)\) with \(x_H^G > 0\) and \(x_L^P > 0\) satisfies both self-refereeing and reputation effect. In particular, for \((x_H^G, 0, 1/2)\) with \(x_H^G \leq 0.1\), both types of authors strictly self referee and, given any paper quality, there is a strict reputation effect.

As shown in the following result, these two properties show the key properties of all equilibria consistent with the motivating facts.
**Proposition 2** Any equilibrium consistent with the motivating facts must involve self-refereeing and display a reputation effect.

This result shows that the motivating facts can be explained by self-refereeing and self-refereeing can only be sustained by reputation effect. Enough good papers are self-refereed such that papers at HV positions have on average higher quality. Self-refereeing must be satisfied because, otherwise, at HV positions, papers submitted around the deadline are worse papers and hence cited less, which is contradictory to *Fact Three*. However, self-refereeing can only be sustained through reputation effect. Self-refereeing means authors submit their bad papers away from the deadline. Since papers submitted away from the deadline are on average worse papers, any author submitting away from the deadline gains less in the first period from both the risk of being lost and a less number of citations at LV positions even if not lost. We can infer that they sacrifice because of the reputation effect. They can only gain from the reputation effect because, in the second period, all authors submit around the deadline. Therefore, there is a trade-off between submitting around the deadline for more citations from other authors’ self-refereeing as well as papers’ smaller risk of being lost, and submitting away from the deadline to benefit from the reputation effect.

But how can the reputation effect be sustained? Given the reputation effect, when will authors self referee? The answer will show the basic intuition for all equilibria consistent with the motivating facts. To answer these questions, in the following analysis, we will analyze the equilibrium incentives and equilibrium conditions for the strategy profile \((1,0,1,1)\). This choice is due to the two implications of Proposition 2. First, some good authors with bad papers must submit away from the deadline. Otherwise, all bad papers are submitted...
around the deadline, which violates self-refereeing. (1, 0, 1, 1) shows good authors with bad papers are the only group of authors submitting away from the deadline and good authors submit all their bad papers away from the deadline. Good authors with bad papers have the strongest incentive to submit away from the deadline than other authors. Second, (1, 0, 1, 1) is the only possible pure strategy equilibrium consistent with the facts. It shows the strongest degree of self refereeing and reputation effect.

Let’s first consider good authors with bad papers. They submit away from the deadline because now they have bad papers. Readers are very likely to get bad signals. Hence, even if they submit around the deadline, their papers can only generate a small number of citations in the first period. In the second period, readers will believe they are very likely to be from bad authors. If they submit away from the deadline, even if they are not cited now, in the future, readers will believe they are good authors and their future papers will be cited more.\(^5\)

Bad authors with bad papers have a strictly weaker incentive to invest in reputation because they are less likely to have good papers in the future. The citation advantage for good papers relative to bad papers in the second period (as defined below) is higher if their authors invested in reputation. Suppose some authors’ first period papers with quality \(\theta_i\) have \(s_1\) positions, where \(i \in \{H, L\}\) and \(s_1 \in \{HV, LV\}\). In the second period, their good papers’ citation advantage is

\[
CA(\bar{\theta}_{s_1}^{2i}) = G(\bar{\theta}_{s_1}^{2i}) - F(\bar{\theta}_{s_1}^{2i})
\]

\(^5\)Without loss of generality, to explain things in a simple way, we ignore the fraction of papers submitted away from the deadline but at HV positions due to \(p_{nd} > 0\)

50
Good papers have citation advantage because, all else being equal, they are more likely to have good signals. When inequality (2), \(i.e.\)

\[ CA(\bar{\theta}^{2L}_{LV}) > CA(\bar{\theta}^{2L}_{HV}) \]  \hspace{1cm} (3.1.2)

holds, bad authors with bad papers benefit strictly less from the investment in reputation. Since their future is less promising, they prefer their current papers to have HV positions for a bigger number of citations.

Given the reputation effect, any author always has a stronger incentive to submit their good papers around the deadline. In the first period, good papers at HV positions have a higher citation advantage than those good papers at LV positions, which means

\[ CA(\bar{\theta}^{1}_{HV}) > (1 - \epsilon)CA(\bar{\theta}^{1}_{LV}) \]  \hspace{1cm} (3.1.3)

Inequality (3) holds for all \(\epsilon > 0\) because papers at LV positions are not cited and hence good papers have a higher cost of being placed at LV positions. We can decompose the intuition into two parts. First, good papers have a higher cost of being missed by readers because good papers are more likely to have good signals. Second, good papers have a higher cost of being placed at LV positions if the ex ante probability of being lost to readers is zero. In the second period, the reputation gain is smaller. Readers believe whoever with good papers deviating in the first period is at most a good author. Readers also believe authors with good papers submitting around the deadline are more likely to be good authors than those with bad papers submitting around the deadline because good authors are more likely to have good papers and good authors self referee. Generally, inequality (3) shows, the higher \(\epsilon\) is, the stronger authors’ incentive to self-referee.
As shown in the following result, given parameters $\beta$ and $p$, the sustainability of these equilibrium incentives depends on three unobservable parameters (the size of HV positions ($\eta$), the noisiness of papers’ signals ($\sigma$) and the citing criterion (equivalently, $R$)) as well as one observable variable (the reputation $q$). Such an equilibrium exists for all $\epsilon > 0$.

**Proposition 3** (*existence*) Suppose $\beta \in (0, 1)^6$. $\forall \epsilon > 0$, if there are a large fraction of HV positions (big $\eta$), papers’ signals are sufficiently noisy (big $\sigma$), and readers’ citing standards are sufficiently high (big $R$), there exist $q(\eta, \sigma, R)$ and $\overline{q}(\eta, \sigma, R)$ such that, $(1, 0, 1, 1)$ is an equilibrium consistent with the motivating facts if and only if the reputation satisfies $q \in [\underline{q}, \overline{q}] \subset (\frac{1-\eta}{1-p}, 1)$.

The proof is in Appendix A. This result can be decomposed into the following three parts.

First, if the fraction of HV positions is sufficiently large, papers’ signals are sufficiently noisy, and readers’ citing criterion is sufficiently high, good authors with bad papers have a strict incentive to invest in reputation for sufficiently low reputations, i.e., $q \in (\frac{1-\eta}{1-p}, \overline{q})$. Good authors with bad papers at reputation $\overline{q}$ are indifferent between submitting around 4pm and away from 4pm.

If the fraction of HV positions is sufficiently big, even for the group of authors with a very low reputation $q$, some of good authors’ bad papers will get HV positions, which

---

6$(1, 0, 1, 1)$ is also an equilibrium for some $\beta$ when $\beta \geq 1$. However, we cannot give a conclusive answer for any equilibrium consistent with the motivating facts. In the following example, we will show that a strategy profile $(x^G_H, 0, 1, x^B_L)$ consistent with the motivating facts and satisfying $x^G_H < 1$ and $x^B_L < 1$ can be an equilibrium when $\beta > 1$. 

52
means Fact Two is satisfied ($p_{nd} > 0$). With such a low reputation, good authors with bad papers have a strict incentive to invest in reputation if papers’ signals are sufficiently noisy and/or readers’ citing criterion is sufficiently high. The argument follows from two observations. First, whenever Fact Two is satisfied, the incentive of authors with bad papers to invest in reputation strictly decreases with the reputation $q$. With a lower reputation, papers at HV positions in the first period are cited less and hence the sacrifice for the investment in reputation is smaller. Moreover, the reputation gain in the second period is higher. The higher reputation gain has two driving forces. On one hand, with $p_{nd}$ fixed, a lower reputation implies the first period papers at HV positions are less likely to be from good authors: among bad papers at HV positions, a less number of papers are from good authors and more from bad authors. On the other hand, $p_{nd}$, $1 - \frac{1-\eta}{q(1-p)}$, decreases with the reputation $q$. Second, the strong incentive of good authors with bad papers to invest in reputation is reenforced by papers’ noisier signals and readers’ higher citing criterion. When readers make decisions, they depend on two pieces of information: papers’ signals and the inference from papers’ positions as well as their authors’ history of publication information. The history of publication information means the positions and the total number of citations of papers in the first period. The noisier papers’ signals are, the less readers weigh papers’ signals in the decision making. Readers’ higher citing criterion further reduces the sacrifice for the investment in reputation. Therefore, when the reputation is sufficiently low, good authors with bad papers have a strict incentive to invest in reputation. Since the reputation gain diminishes with the increase of reputation, good authors with bad papers have a strict incentive to submit around the deadline if the reputation is sufficiently high.
Second, if readers’ citing criterion is sufficiently high, there exist \( \bar{q} > \bar{q} \) such that good authors with bad papers have a strictly stronger incentive to invest in reputation for reputations satisfying \( q \in (\frac{1-n}{1-p}, \bar{q}) \) than bad authors with bad papers.

The intuition follows from the observation that the citing criterion \( R \) and the reputation \( q \) are substitutes in the following sense. The bigger the citing criterion or the lower the reputation \( q \), the stronger the relative incentive of good authors with bad papers to invest in reputation. With a higher citing criterion, papers are always cited less in the second period. It implies that good papers’ citation advantage is higher if their authors had papers at HV positions and lower if not. The intuition is, the stronger readers’ incentive to cite or not to cite before observing any noisy signal, the lower good papers’ citation advantage is. Before observing noisy signals, readers cite papers whose authors had papers at LV positions and do not cite otherwise. A higher citing criterion weakens the incentive to cite and strengthens the incentive not to cite. The role of a lower reputation \( q \) is similar to a higher citing criterion in the sense that readers are less likely to cite papers if their authors did not have papers at LV positions. Since good papers’ citation advantage if their authors had bad papers at LV positions does not change, the relative incentive of good authors with bad papers to invest in reputation is strictly stronger. If the citing criterion is sufficiently high, there exists a \( \bar{q} > \frac{1-n}{1-p} \) such that good authors with bad papers still have a strictly stronger incentive to invest in reputation than bad authors with bad papers when the reputation \( q \) satisfies \( q \in (\frac{1-n}{1-p}, \bar{q}) \).

Finally, there exists \( q > \frac{1-n}{1-p} \) such that \((1,0,1,1)\) is an equilibrium if and only if the reputation \( q \) satisfies \( q \in [q, \bar{q}] \). With reputation \( q \), either bad authors with bad papers or good authors with good papers are indifferent between submitting around the deadline and
away from the deadline. Given papers of the same quality, good authors have a strictly stronger incentive to submit around the deadline because good papers have a higher cost of being placed at LV positions and the reputation gain is strictly smaller. \((1, 0, 1, 1)\) can only be sustained as an equilibrium for \(q \in [\underline{q}, \overline{q}]\). With a too low reputation, all authors with bad papers invest in reputation; with a too high reputation, good authors with bad papers have a strict incentive to submit around the deadline and/or bad authors with bad papers have a strictly stronger incentive to invest in reputation than good authors with bad papers.

Figure 3.5 illustrates the payoff advantage of HV positions (relative to LV positions) for good authors with bad papers \((V^G_L)\), bad authors with bad papers \((V^B_L)\) and good authors with good papers \((V^G_H)\). As is shown, if \(q > \frac{1-\eta}{1-p}\), the following three properties are satisfied. First, \(V^G_L\) and \(V^B_L\) both strictly increase with \(q\). Second, the relative incentive of good authors to invest in reputation, \(V^B_L - V^G_L\), strictly decreases with \(q\). At \(\overline{q}\), good authors and bad authors have the same incentive. Finally, given good papers, good authors have a strictly stronger incentive to submit around the deadline \((V^G_H > V^G_L)\). For any \(q \in [\underline{q}, \overline{q}]\), \((1, 0, 1, 1)\) is an equilibrium.

More generally, suppose, for some \(\beta > 0\) and \(q \in [\underline{q}, \overline{q}]\), \((1, 0, 1, 1)\) is an equilibrium, where both \(q\) and \(\overline{q}\) depend on \(\beta\). Under the same parameters, we will show that both good authors...
Figure 3.5: The strategy profile $(1, 0, 1, 1)$ is an equilibrium for any $q \in [q, \bar{q}]$. $V^a_i$ denotes the payoff advantage of higher visibility positions for any author who is of quality $a$ and whose paper is of quality $\theta_i$, where $a \in \{G, B\}$ and $i \in \{H, L\}$.

with good papers and bad authors with bad papers can have equilibrium incentives to invest in reputation. For all reputation $q \in [q, \bar{q}]$ and $x^G_L < 1$ such that $(1, x^G_L, 1, 1)$ is consistent with Fact Two, $(1, x^G_L, 1, 1)$ is not an equilibrium because good authors with bad papers have a strict incentive to submit away from the deadline. However, given $x^G_L < 1$, a strategy profile where bad authors with bad papers invest in reputation can be an equilibrium. Bad authors with bad papers invest in reputation only if some good authors with good papers also invest in reputation. If some bad authors with bad papers invest in reputation, the reputation gain for bad authors with bad papers is reduced, which implies bad authors with bad papers have a strict stronger incentive to submit around the deadline than in $(1, x^G_L, 1, 1)$. However, if a sufficient fraction of good authors with good papers invest in reputation, the sacrifice for the investment in reputation will be smaller, which reduces the incentive of bad authors with bad papers to submit around the deadline. Good authors with good papers
have an incentive to invest in reputation as long as a lot of good authors with good papers
(relative to $x_B^H$) invest in reputation such that the sacrifice for the investment in reputation
is sufficiently small and the reputation gain is sufficiently big. Moreover, given $x_L^G < 1$, the
more bad authors with bad papers invest in reputation, the more good authors with good
papers invest in reputation, i.e., $x_H^G$ strictly increases with $x_L^B$. If sufficiently many bad
authors with bad papers invest in reputation, we must have $x_H^G + x_L^B < x_L^G + x_H^B$.

Suppose $q < q$. For strategy profile $(1, x_L^G, 1, 1)$ consistent with Fact Two, bad authors
with bad papers have a strictly stronger incentive to invest in reputation. There exists
$x_L^B \in (x_L^G, 1)$ such that bad authors with bad papers are indifferent between submitting
around the deadline and submitting away from the deadline. Bad authors with bad papers
are more likely to invest in reputation only if good authors with good papers invest in
reputation, i.e., given $x_L^G$, if $x_L^B' (> x_L^B)$ is part of an equilibrium, $x_H^G < 1$ must be satisfied.
We can numerically show such an equilibrium satisfying $x_L^G = 0$, $x_H^B = 1$ and $x_H^G + x_L^B < 1$
exists when $q < q$. Let $x_L^B = 1/2$ (as shown in Figure 3.4). Assume $p = 0.65$, $\eta = 0.7$, $\epsilon = 0.4$,
$\theta_H = 2$, $\theta_L = 1$, $R = 0.6$ ($\theta^* = 1.6$), $\beta = 2$ and $\sigma = 0.76$. This strategy profile is consistent
with the motivating facts for all reputations. As shown in Figure 3.6, if $x_H^G = 0.227$, there
exists $q \in (0.35, 0.45)$ such that $V_H^G = V_L^B = 0$, $V_H^B > 0$, $V_L^G < 0$ and $p_{nd} > 0$. Since
$0.45 < 6/7$, this strategy profile is an equilibrium consistent with the motivating facts and
satisfies $q < q$.

With the parameter vector $(\beta, p)(= (2, 0.65))$ and $(\eta, R, \sigma, \epsilon)(= (0.7, 0.6, 0.76, 0.4))$, au-
thors’ incentives as a function of reputation $q$ follow similar properties as those in $(1, 0, 1, 1)$,
\footnote{If $x_H^G$ and $x_L^B$ are both sufficiently small, good authors may not self referee and only bad authors self
referee to be consistent with Fact Three. But $x_H^G < x_L^B$ has to be satisfied for the reputation effect to hold.}
Figure 3.6: \((0.227, 0, 1, 1/2)\) is an equilibrium at around \(q = 0.4\).

but also show additional characteristics. Since \(x_H^G < 1\) and \(x_L^B < 1\), this strategy profile is more likely to be consistent with Fact Two than \((1, 0, 1, 1)\). Under the given parameters, Fact Two is satisfied for all reputations. We still observe that if authors’ reputation is sufficiently high, their incentive to submit around the deadline strictly increases with the reputation and authors self referee. Moreover, given paper quality, good authors have a strictly stronger incentive to invest in reputation only when the reputation is sufficiently low. However, as reputation \(q\) becomes increasingly low, the reputation gain diminishes for authors with bad papers and increases for authors with good papers. With a sufficiently low \(x_L^B\), the sacrifice for the investment in reputation becomes bigger and bigger for all authors with the decrease of their reputation \(q\). Therefore, we have the following two additional observations. First, authors’ incentive to submit around the deadline decreases for sufficiently low reputations. For authors with good papers, as their reputation becomes increasingly low, the decrease in the sacrifice for the investment in reputation overweighs the increase in the reputation gain. Second, authors self-referee only when their reputation is sufficiently high, \(i.e.,\) when their
reputation gain is sufficiently small. Therefore, \((0.227, 0, 1, 1/2)\) can and only can be sustained as an equilibrium for the intermediate reputation, that is, the reputation \(q\) is around 0.4. These two observations are essential for the sustainability of \((0.227, 0, 1, 1/2)\) as an equilibrium. As we can see from Figure 14, bad authors with bad papers are indifferent between submitting around the deadline and away from the deadline at two reputations, which is due to the first additional observation. However, due to the second additional observation, the equilibrium can only be sustained at the lower reputation. At the higher reputation, all authors with good papers have a strict incentive to submit around the deadline, which means \((0.227, 0, 1, 1/2)\) can not be sustained as an equilibrium.

The intermediacy of the equilibrium reputation as we have shown in the previous two examples is a general property in the sense of the following result.

**Proposition 4** \(\forall \epsilon > 0\), there exist \(\sigma^* > 0\), \(q_B\) and \(q_T\) \((0 < q_B < q_T < 1)\) such that, for any equilibrium consistent with the motivating facts, all bad authors with bad papers submit around the deadline or good authors with good papers randomize only if \(\sigma > \sigma^*\) and their reputation satisfies \(q \in (q_B, q_T)\).

The proof is in Appendix A.

This result first shows, for strategy profile \((0.227, 0, 1, 1/2)\) to be an equilibrium, we must have \(\sigma > \sigma^*\), which implies \(0.76 > \sigma^*\). Otherwise, all good papers will be submitted around the deadline to avoid the risk of being missed by readers\(^9\). The reason is, if signals

\(^9\)Authors with bad papers may be indifferent between submitting around the deadline and away from the deadline. Authors with good papers and authors with bad papers have different incentives because good
are very precise, readers are almost sure about papers’ quality and hence authors’ strategic submission behavior can not influence readers’ citing behavior.

This result also shows that, for \((0.227, 0, 1, 1/2)\), since the equilibrium reputation is around 0.4, \(q_B < 0.4 < q_T\) must be satisfied. If the reputation is very high, no equilibrium is consistent with the motivating facts. Since the reputation gain is very small and the sacrifice is very big, all authors have a strict incentive to submit around the deadline. Moreover, for a sufficiently high reputation, given paper quality, bad authors have a strictly stronger incentive to invest in reputation than good authors, which destroys the reputation effect. If authors’ reputation is very low, we will differentiate two cases, depending on whether bad authors with bad papers invest in reputation or not. First, suppose at equilibrium all bad authors with bad papers submit around the deadline, that is, \(x_B^L = 1\). With sufficiently low reputations, given paper quality, good authors always have a strictly stronger incentive to submit away from the deadline than bad authors. Since authors self referee, all bad authors must invest in reputation. Their reputation must be sufficiently high for Fact Two to be satisfied. Second, suppose at equilibrium, good authors with good papers randomize and bad authors with bad papers invest in reputation. All bad authors with good papers submit around the deadline. We are considering strategy profiles \((x^G_H, 0, 1, x^B_L)\) with \(x^G_H \in (0, 1)\) and \(x^B_L \in (0, 1)\). \((0.227, 0, 1, 1/2)\) is one example. No matter Fact Two is satisfied or not, the reputation gain for authors with bad papers is very small, which implies that the sacrifice is also very small for bad authors with bad papers to be indifferent between submitting away papers have a higher cost of being missed by readers. That is to say, a strategy profile with all good papers submitted around the deadline can still be sustained as an equilibrium consistent with the motivating facts if papers’ signals are very precise.
from the deadline and submitting around the deadline. The sacrifice is very small only if papers’ signals are very noisy such that neither papers at HV positions nor papers at LV positions have a lot of citations. However, if papers’ signals are very noisy, for authors with good papers, the sacrifice for the investment in reputation is also very small. Then, the reputation gain for authors with good papers will outweighs the small sacrifice and hence all authors with good papers will submit away from the deadline.

3.2 Theoretical Interpretation of Empirical Results

From Chapter 3

In the model, we assume at the beginning of period 1, there is a prior $q$ that a fraction $q$ of authors are good authors. That is to say, we assume there is only one group of authors sharing the same reputation at the beginning of period 1. To test the model, we need to extend the model to the case with multiple groups of authors and each group shares the same reputation. The reason for the extension is, in the real data, there are multiple groups of authors in the population and readers can identify authors’ reputation through different measures, for example, h-index which will be introduced and constructed in Chapter 3. With the extension of the model, authors’ reputation can directly enter the empirical model and, based on Proposition 4, we need to estimate with authors of intermediate reputations in the sample. And the interpretation of the empirical results becomes more straightforward.
The Model With Multiple Groups of Authors

The generalization is straightforward. First, in the second period, all authors submitting around the deadline is still the unique equilibrium. The result from the model with one group of authors at the beginning of the first period can be generalized. The reason is, for the simplified model, in the second period, there are multiple groups of authors for the equilibrium where some authors submit around the deadline and some other authors submit away from the deadline in the first period. Second, any equilibrium profile can be analyzed under the simple framework with a redefined size of HV positions, $\eta_q$, for each group of authors with reputation $q$. Suppose they have in total $n_q$ papers and they submit $D_q$ papers around the deadline. Suppose there are in total $D$ papers submitted around the deadline among all authors. Then, the fraction of HV positions for this group of authors is

$$\eta_q = \frac{D_q + (n_q - D_q) \frac{\eta - D}{1 - D}}{n_q}$$

where $\eta - D$ is the space at HV positions left for papers submitted away from the deadline and $1 - D$ is the total fraction of papers submitted away from the deadline. Therefore, $\frac{\eta - D}{1 - D}$ is the probability of papers submitted away from the deadline to have HV positions. In particular, for the group of authors all submitting around the deadline, $\eta_q = 1$; if any group submitting all papers away from the deadline, then $\eta_q = \frac{\eta - D}{1 - D}$. Suppose there are in total $Q$ groups of authors and the reputation for group $i$ is $q_i$, then the total size of HV positions satisfies
II. The Inference From Chapter 3

In Chapter 3, we find the existence of reputation effect and self-refereeing, which is consistent with the prediction of Proposition 2. Moreover, for average intermediate reputations, all authors self-referee and in particular, bad authors are more likely to self-referee, which implies only equilibria satisfying the following relationship (relationship (4)) can be consistent with the empirical results from Chapter 3. Relationship (4) shows that $0 < x_H^G < 1$ and $0 < x_L^B < 1$ must be satisfied\footnote{By $x_L^G < x_L^B$ and relationship (4), we can infer $x_H^G < x_H^B$. Since all authors self-referee, $0 < x_H^G < 1$ must be satisfied. If $x_L^B = 1$, for relationship (4) to be satisfied, we must have $x_H^G < x_L^G$. However, authors’ self-refereeing will be violated.}, which confirms the prediction of Proposition 4 in that equilibria satisfying $x_H^G \in (0, 1)$ can only be sustained for intermediate reputations. If the intermediate reputations we choose for the testing are sufficiently low, $x_L^G = 0$, $x_H^B = 1$ and relationship (4) is equivalent to $x_H^G + x_B^B < 1$.

$$x_H^B - x_L^B > x_H^G - x_L^G > 0 \quad (3.2.4)$$

Now we can make inferences about the two key parameters. First, papers’ signals must be sufficiently noisy, that is, $\sigma > \sigma^*$. This result is significant in that, even though we do not know which equilibrium plays in the data, we are confident that the noisiness of papers’ signals has a lower boundary. When we talked to some physicists, they believed they knew papers’ quality pretty well. Our result shows that these physicists are over-confident. Our
Figure 3.7: When $\epsilon = 0$, $(0.227, 0, 1, 1/2)$ is not equilibrium for any reputation $q$.

model has the prediction that the number of citations of any paper should be higher if it is located at a position with many papers of higher quality. Second, $\epsilon$ is probably very high. Rewriting relationship (4), we have $x^G_H + x^B_L < x^B_H + x^G_L$. Therefore, relationship (4) holds for all sufficiently small $x^G_H$ and $x^B_L$. Higher $\epsilon$ reenforces authors’ incentive to self-referee. With sufficiently small $x^G_H$ and $x^B_L$, the reputation gain for authors with bad papers is sufficiently low and the reputation gain for authors with good papers is sufficiently high, authors self-referee only if $\epsilon$ is sufficiently big. If instead $\epsilon$ is very small, authors have no incentive to self-referee. Figure 3.7 shows $(0.227, 0, 1, 1/2)$ is not an equilibrium if $\epsilon = 0$. Authors with good papers have a strictly stronger incentive to submit around the deadline only when the reputation is sufficiently high. However, with high reputations, bad authors with good papers have a strictly stronger incentive to invest in reputation than good authors with good papers. When the reputation decreases, the relative reputation gain for authors with good papers is so big such that even for as high a reputation as 0.6, no author self-referees.
3.3 Welfare Analysis

For the welfare analysis, it is sufficient to only consider the model with only one group of authors sharing the same reputation at the beginning of the first period. The result applies to the model with multiple groups of authors as defined in Section 3.2.1. We find the social payoff of any equilibrium consistent with the motivating facts is higher than that of the unique uninformative equilibrium where all authors submit around the deadline. It suggests that we should not randomize the order of the lists. If the order is randomized, authors lose the incentive to strategically submit their papers. The social payoff is equivalent to the equilibrium with all authors submitting around the deadline. Some people suggested to randomize the order of the lists because they believe that the fact that some authors manipulate the positions of their papers is against the idea of research and is not fair to other authors who choose not to do so. We find the current submission system is beneficial to both authors and readers. Since in the second period all papers are submitted around the deadline, the welfare implication of any informative equilibrium can only come from the first period. The first period submission is informative of papers’ quality in period 1 and authors’ quality in period 2, all of which has welfare implications. First, we suppose the social planner puts all weight on readers. The result is formally stated as below.

**Proposition 5** Suppose the social planner puts all weight on readers. Any equilibrium consistent with the facts generates higher social payoff than the unique uninformative equilibrium.

The proof is in Appendix A.
The intuition can be sketched as follows. In period 2, readers always benefit from informative submissions in period 1 because information about authors’ quality are revealed. In the first period, readers benefit from authors’ self-refereeing, i.e., they not only gain from the separation of good papers from bad papers, but the loss of bad papers at LV positions. In comparison to the case with $\epsilon = 0$, the condition that $\epsilon > 0$ increases the social payoff at least in reducing readers’ loss in reading bad papers. A higher $\epsilon$ is welfare-improving in the sense of increasing authors’ incentive to self-referee.

Now suppose the social planner also puts positive weight on authors\textsuperscript{11}. Though readers only benefit from citing good papers, authors want all their papers cited, which implies the social planner is “risk averse” with high beliefs about paper quality and “risk loving” with low beliefs about paper quality. When the belief is high, papers are cited a lot, authors also gain a lot from the citations and hence there is too much to lose. When the belief is low, papers are not cited much, there is little to lose. Noisier signals amplify this argument. Therefore, authors gain from informative submissions in each period only when the prior reputation $q$ is sufficiently low and papers’ signals are sufficiently noisy. Hence, Proposition 5 still holds for sufficiently low prior reputations ($q$) when papers’ signals are sufficiently noisy. However, Proposition 5 may still hold for higher reputations. With high reputations, papers at HV positions in the first period have higher expected quality and hence are cited more. Authors gain from more citations, which overweighs the social planner’s “risk averseness” when papers’ signals are sufficiently noisy.

\textsuperscript{11}The details are in Appendix B.
3.4 Discussions About Assumptions

In this section, we will address questions about assumptions of our model often raised by other researchers. The first is about authors’ private information. We assume authors know both the quality of themselves and their papers. For reasons to be explained below, we will only relax the strict assumption on authors’ belief about the quality of their papers. Second, we will highlight the role played by $\epsilon$ and the endogeneity of the probability that papers submitted away from the deadline get HV positions. Finally, we will also discuss limitations of the theoretical model in matching the real world.

I. Authors’ Information About The Quality of Their Paper

In the model, we assume authors know both the quality of their papers and themselves. We will only relax the assumption on authors’ belief about the quality of their papers. We believe that, relative to authors’ belief about the quality of themselves, authors’ belief about the quality of their papers has bigger variance, which is due to the random arrival time of good ideas and the fact that there are more papers than authors. Even bad authors sometimes write a good paper, and sometimes a good author write bad papers.

In this section, we will show that even if authors only have noisy information about the quality of their paper, any equilibrium consistent with the motivating facts still satisfies both self-refereeing and reputation effect. However, authors’ signals have to be sufficiently precise. Otherwise, no equilibrium satisfying both reputation effect and self-refereeing exists.
More specifically, suppose their signals are distributed according to \( \hat{\theta}_a = \theta_i + \sigma_a \xi_a \), where \( i \in \{H, L\} \) and \( \xi_a \sim N(0, 1) \). When signals are very informative, the previous results can still be generalized. As signals become increasingly noisy, researchers’ payoffs depend more on authors with all signals. This helps generating a richer set of submission behavior, but also imposes some constraints. On one hand, whenever one type of papers are submitted around the deadline with positive probability, the other type is also submitted around the deadline with positive probability; on the other hand, the submission behavior of different types of authors are related through signals. Given the same signal, good authors believe their paper is more likely to be good paper than bad authors. All equilibria consistent with the facts result from these conflicting driving forces.

As we formally show in Appendix B, some key results generally hold for all signals. First, in the second period, all papers are submitted around the deadline. Authors with sufficiently high signals submit around the deadline because they believe their paper is almost good. Given the same signal, good authors are always more likely to submit around the deadline than bad authors. Then, papers at lower visibility positions are more likely to be bad papers. Therefore, everyone submits around the deadline. Second, any equilibrium consistent with the facts must still display reputation effect and self-refereeing. If good papers from good authors are more likely to be submitted around the deadline than those good papers from bad authors, then by Fact Three, authors with higher signals submit around the deadline. Bad papers from bad authors are also more likely to be submitted around the deadline. Then, away from deadline submission does not bring about reputation gain in the future. Therefore, everyone has an incentive to submit around the deadline. Similarly, if bad papers from good authors are more likely to be submitted around the deadline, papers with lower
signals are all submitted around the deadline. Papers at higher visibility positions are worse. Again, we reaches a contradiction.

However, if authors’ signals are too noisy, no equilibrium can be consistent with the motivating facts. As signals become increasingly noisy, authors’ submission behavior is getting more and more uninformative of paper’s quality. Readers rely more on their information on authors’ quality to judge paper’s quality. Since good authors are more likely to have good papers, if papers at lower visibility positions are worse, they are more likely to come from bad authors. Then, no one has any incentive for away from deadline submission.

II. Other Assumptions

Now we will highlight the role played by $\epsilon$ and the endogeneity of the probability that papers submitted away from the deadline get HV positions. If $\epsilon = 0$, some strategy profiles consistent with the motivating facts, for example, $(1, 0, 1, 1)$, can still be equilibrium. However, $\epsilon > 0$ results in the uniqueness of equilibria in the second period: when there is no reputation concern, all authors submit around the deadline. Hence, authors’ reputation concern is the only driving force underlying the motivating facts. Moreover, as we have shown before, the higher $\epsilon$ is, the stronger authors’ incentive to self referee. A high $\epsilon$ is probably essential for the existence of an equilibrium consistent with the empirical results from Chapter 3. Some people suggest that for simplicity, the probability of papers submitted away from the deadline should be assumed to be fixed. We choose to make this probability endogenous due to the following three reasons. First, it is more realistic. The number of papers on each
list and to each author, the specific time for other authors’ submissions, are both random and therefore, the probability of papers submitted away from the deadline to have HV positions differs each day. Second, it is easier to do welfare analysis because the fraction of HV positions and LV positions are fixed at \( \eta \).

Our model has the following limitations in terms of matching the real world. Relaxing these assumptions are potentially good research directions to pursue in the future. First, we only consider submission times for one list. Why authors choose to be on different lists is out of the scope of this model. Second, we assume D submissions right after 4pm and right before 4pm are symmetric, driven by the same incentives. However, the incentives may differ. For example, as we will discuss in detail in Chapter 3, authors submitting right before the deadline may be due to the deadline concern, that is, they want their paper to be announced on that list, but they have worked until the last minute. Their papers are better than other papers. In Chapter 3, we find that, after eliminating all papers submitted with deadline concerns, we still find evidence for the prediction of our model. Finally, we match the motivating facts with the first period submission behavior of our model. Alternative models can be built to take into consideration factors not included in our model, for example, authors’ decision to enter or exit the market and the reasons for choosing one particular author as the submitter for coauthored papers.
Chapter 3: Testing The Model

In this Chapter, we will test the implications of the model of self-refereeing from Chapter 2. Our model has implications both in terms of authors’ submission behavior and readers’ citing behavior. Authors’ submission behavior satisfies both self-refereeing and reputation effect (Proposition 2). Self-refereeing means, among authors of the same quality, those with good papers are more likely to submit around the deadline. Reputation effect means, among authors with papers of the same quality, good authors are more likely to submit away from the deadline. Moreover, if all bad authors with bad papers submit around the deadline or good authors with good papers randomize, the equilibrium reputation must be intermediate (Proposition 4). In terms of readers’ citing behavior, readers’ citation decision is based on their estimation of papers’ average quality at any given position and papers’ signals they get from reading the paper. Therefore, there exists positional effect, i.e., papers of the same quality are cited differently at different positions, which are due to the following two possible reasons. First, papers are cited more if, all else being equal, papers at the same position have higher average quality (paper quality effect). Second, all else being equal, papers at higher visibility positions are cited more than those at lower visibility positions (visibility effect).

Our empirical results show that both self-refereeing and reputation effect exist among authors of intermediate academic age with intermediate reputations. Moreover, all these authors self-referee and bad authors are more likely to self-referee. We identified positional effect, visibility effect for papers submitted hours after 4pm and paper quality effect. The
empirical tests have salient features. First, we constructed an algorithm to resolve the long-standing author ambiguity problem, i.e., the data source does not have an id for each author. This algorithm is necessary because, for the measurement of author quality and reputation, we need to get the publication list for each author. Second, we use the number of citations of the most cited papers citing any paper as the measurement of paper quality, which resolves the simultaneity problem if we use the number of citations instead.

The outline of this Chapter is as follows. We will first introduce the data source and the measurement of the three main variables (author quality, paper quality and reputation). Then, we will focus on testing my model’s implications for authors’ submission behavior, which is followed by testing the positional effect. For authors’ submission behavior, we will first use the sample with authors of all ages, then we differentiate authors of different ages. At the end, we discuss alternative explanations, that is, the existence of deadline effect as well as the coexistence of naive and sophisticated authors.
4.1 Data and Measurement

In this section, we will introduce the data source and show the measurement of main variables (author quality, paper quality and reputation).

I. Data

We choose data only for submitters who submitted at least one paper from the US or Canada between 2002 and 2004. There are two main reasons for choosing these submitters. First, we want the analysis to be consistent with Chapter 1 and Chapter 2. In Chapter 1, we present the motivating facts mainly with data between 2002 and 2004. Our theoretical model in Chapter 2 accommodates these facts. Though we believe our theoretical model has general explaining power for other submitters, we still choose these submitters in order to be consistent. Second, we didn’t choose authors who submitted only after 2004 because we intend to reduce the measurement error of author quality. As we will see later, the measurement of author quality is based on the publications until September 2012. The longer the publication history, the smaller the measurement error. Therefore, in choosing submitters starting their academic life earlier, we give these submitters more time to accumulate publications and hence reduce the measurement error. Our testing of the theoretical model requires data on these submitters’ publication history (publication date, number of authors and citation path (the year and month of each cite)), which includes the submission history (publication date, number of authors and citation path), and the paper-level information for each paper in the submission history (original submission time, final submission time (to be introduced
shortly later in this section), timezone and the citation path of each paper citing this paper (the year and month of each cite)).

The data come from various sources. Original submission times and the submitters’ email addresses for all papers submitted between Jan 2000 and Mar 2007 are provided by Paul Ginsparg. Final submission times and each submitter’s name are retrieved from arXiv.org. The names of authors and authors’ affiliations are mainly collected from SPIRES and INSPIRE, supplemented by arXiv.org. SPIRES and INSPIRE are both online libraries storing all papers related to high energy theory\footnote{As of 2012, INSPIRE replaced SPIRES for most of its functions. But some data are still provided through SPIRES. For example, the database on institutions still stores on SPIRES, linked through INSPIRE.}. The US timezone for each affiliation from US or Canada and last names for authors with Chinese or Korean descent are gathered from various sources on internet. We retrieve each submitter’s publication records until September 2011 from SPIRES and INSPIRE, but each paper’s citation path that we get from INSPIRE lasts until September 2012. Therefore, each paper in the publication history has at least one year to accumulate citations. More detailed information about the criterion of retrieving publication records are in Appendix C.

The main difficulty in collecting and cleaning the data is there is no author ID. It means we have to rely on names. However, different authors may have the same name, and even the same author often list her name as the submitter or the author in multiple ways. In Appendix C, we describe the algorithm used to identify all relevant submitters and their publication history. In particular, we eliminate all papers submitted during holidays because the incentives for these submissions are out of interests of this paper\footnote{Holidays are defined according to What’s New on arXiv.org.}. We assume
a submission is from the US or Canada if his/her first affiliation is located in the US or Canada. We exclude authors with Korean and Chinese descent because there is a large degree of overlap in their names, making it more difficult to clean the data.

With papers submitted from authors of Chinese or Korean descent excluded, between 2002 and 2004, the motivating facts as in Chapter 1 still hold. Since papers submitted from these authors only account for less than 10 percent of the total submissions, we still find similar patterns as in Fact One and Fact Two. Fact Three also holds. Given top 2 (bottom 2) positions, papers from D submission get higher citations than papers from ND submission (P-value from Wilcoxon ranksum test is 0.000 for top 2 positions (and 0.0313 for bottom 2 positions)). Our theoretical model can accommodate these facts. In the next section, we will identify the strategic submission pattern as suggested by my model from these subset of data.

---

3If a submitter has only one affiliation located in the US or Canada, the US timezone for this submission is based on the timezone of this affiliation. Suppose a submitter has multiple affiliations from the US or Canada. If one of these affiliations matches the email address, we use the timezone of the affiliation in the email address. If the email address is not affiliated or if the affiliation in the email address doesn’t match the submitter’s affiliations from the US or Canada, we use the timezone for the submitter’s first affiliation from the US or Canada.

4Between 2002 and 2004, there are 2569 papers submitted from US or Canada. 228 are submitted from authors of Chinese or Korean descent.
II. Measurement of Main Variables

Our model in Chapter 2 has implications from both readers and authors’ side. From the readers’ side, there exists positional effect which means papers of any given quality are cited more due to one or both of the following two reasons. First, papers having the same position have on average higher paper quality and hence also cited more (paper quality effect); second, papers in concern are at more visible positions (visibility effect). From the authors’ side, only authors with intermediate reputations submit according to equilibria consistent with the motivating facts. These equilibria satisfy two properties. First, among authors of the same quality, those with good papers are more likely to submit around the deadline (self-refereeing). Second, given paper quality, good authors are more likely to submit away from the deadline (reputation effect). To test these two implications, we need to measure author quality, paper quality and reputation.

We measure author quality by calculating the co-author adjusted h-index accounting for the author’s age. h-index was proposed by physicist Jorge Hirsch [2005] to measure the quality of theoretical physicists. Since then, it has been widely used to measure scientific contribution. An author’s h-index is $h$ if he has at most $h$ papers each with at least $h$ citations. A paper accounts for $\frac{1}{n}$ papers for an author if the paper has $n$ authors [Ellison, 2010]. Age is the number of months since one author’s first paper published until September 1, 2012. Since h-index weakly increases with an author’s age, to measure $q_i$ with h-index, we need to eliminate the influence of age. We measure $q_i$ by scaling each author’s coauthor-adjusted h-index with age, as shown in the top figure of Figure 5.3. That is, given each age, we scale each author’s coauthor-adjusted h-index with the value on the scaling line.
Therefore, each author’s quality is measured by how good he is relative to his peers at the same age\textsuperscript{5}. Due to the discreteness of the data, for simplicity, we choose a linear scaling line.

Similarly, we measure reputation by constructing submission h-index accounting for the submission age. Submission h-index is the coauthor-adjusted h-index at the time of submission. Submission age is the number of months from the author’s first paper until the submission in question. The bottom figure of Figure 5.3 shows the scaling of submission h-index with submission age.

To measure paper quality, we cannot use this paper’s total number of citations. Otherwise, we will have a simultaneity problem: readers’ citing behavior is influenced by authors’ strategic submissions. To deal with this problem, we measure paper quality \( q_p \) by the number of citations of the most cited papers citing paper \( p \). Highly cited papers are likely to be good papers. Their authors’ citing decision is less influenced by other people’s citing behavior, but better reflective of paper \( p \)’s quality. More specifically, the measurement involves the following steps. First, we calculate the mean of the total number of citations of the five most cited papers citing paper \( p \) until September 2012 one year after paper \( p \) is published.

\textsuperscript{5}There are two potential problems here. First, authors at different ages may have different quality. Second, even if authors of different ages have the same quality, older authors survive the selection process. Therefore, the worst older authors should have higher quality than the worse young authors. We deal with these problems by controlling for author age in the main regression. Another way to measure the author quality is to first regress author h-index on a polynomial of author age and use the residual as the measurement of author quality. We choose not to use this method because there is a bigger variance as author age increases. We will overestimate the quality of younger and bad authors. We will underestimate the quality of younger and good authors.
The one year time window is chosen to reduce the impact of the immediate publishing position over arXiv.org on the citing decision of these papers’ authors. Then, we regress this mean number of citations on a second order polynomial of the mean age of the five papers involved. The age of a paper is measured by the number of months since this paper’s publication until September 1, 2012. Finally, we regress the estimated error from the first regression on a second order polynomial of the age of paper \( p \). We need to control the age of paper \( p \) because, the older paper \( p \) is, the more likely a good paper “finds” it and cites it. We use the estimated error from the second regression to measure paper quality.

Figure 4.1 shows the measurement of paper quality as a function of submission time and positions. The top figure shows the measurement of paper quality as a function of submission time. The x-axis shows the minutes away from the deadline, y-axis on the left shows the median paper quality for submissions in each 30 minutes or 60 minutes time interval. Overall, the curve of paper quality exhibits a U-shape, \( i.e. \), papers submitted around the deadline are better papers\(^7\). The bottom figure shows that the measurement of paper

\(^6\)If paper \( p \) is cited less than five times after one year on the arXiv, then we add dummy papers as zero citation paper to bring the number of papers to five. Assume these dummy papers’ age starts from one year after paper \( p \) gets published on arXiv. If paper \( p \) was withdrawn, we take the citation as 0 and the age starts from the date on arXiv.

\(^7\)The curve of paper quality also shows that papers submitted between 2.5 hours and 12 hours after 4pm have on average higher paper quality than those submitted between 30 minutes and 150 minutes after 4pm, but with big variance. The big variance is due to the much smaller number of submissions. We can infer that authors’ submission behavior is also influenced by their random schedule, \( i.e. \), the schedule without the influence of the submission deadline. Either good authors like working and submitting in the evening until late night or papers finished and submitted at night are more likely to be good papers. We will take this into consideration when we run OLS for testing authors’ submission behavior by clustering on authors.
Figure 4.1: The measurement of paper quality. The top figure shows D submissions have higher quality than ND submissions. Within 180 minutes after 4pm and 360 minutes before 4pm, the data are shown in 30 minutes time interval; otherwise, the data are shown in 60 minutes time interval. The bottom figure shows the relationship between paper quality and the number of citations. The scaled number of citations means the estimated error from the regression of the number of citations on paper age.
quality is consistent with the number of citations. To eliminate the influence of paper age, we regress the number of citations on paper age and use the estimated error as the measurement of citations, which we name as the scaled number of citations. Our result has two implications. Under the null hypothesis that our measure of paper quality is correct, we confirm our model’s prediction that, for any given position, papers’ number of citations is determined by paper quality. Under the null hypothesis that our model is true, our measurement of paper quality is reasonable. Moreover, the shape of paper quality as a function of position shows that papers at top 2 or bottom 2 positions have higher paper quality, which confirms our idea that papers submitted around the deadline have on average higher quality.

III. Summary Statistics

The information is in Table 4.1.

---

It is noteworthy that papers at position L-1 have lower paper quality than position L-2. This observation is not contradictory to the top figure where papers submitted closer to the deadline have higher paper quality. The probable reason is submissions are clustered on some days. For many lists, no papers are submitted close to 4pm (before 4pm), therefore, a paper submitted away from 4pm get the bottom position, and/or on some days, there are multiple papers submitted right before 4pm and hence even papers at L-2 have high paper quality.
<table>
<thead>
<tr>
<th>Variable</th>
<th>Min</th>
<th>Median</th>
<th>Max</th>
<th>Mean</th>
<th>Stan. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Paper level information (4191 authors)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of authors</td>
<td>1</td>
<td>2</td>
<td>8</td>
<td>2.104748</td>
<td>1.074661</td>
</tr>
<tr>
<td>Submission time (min.)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(within 720 min. after 4pm)</td>
<td>0</td>
<td>110</td>
<td>715</td>
<td>164.6922</td>
<td>169.7195</td>
</tr>
<tr>
<td>(within 720 min. before 4pm)</td>
<td>0</td>
<td>146</td>
<td>718</td>
<td>186.6381</td>
<td>154.7935</td>
</tr>
<tr>
<td>Length of the list</td>
<td>2</td>
<td>13</td>
<td>35</td>
<td>13.11586</td>
<td>4.61146</td>
</tr>
<tr>
<td>(the number of papers)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Author level information (745 authors)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of papers</td>
<td>1</td>
<td>4</td>
<td>34</td>
<td>5.633065</td>
<td>5.432422</td>
</tr>
<tr>
<td>Author Age (mon.)*</td>
<td>93</td>
<td>208.5</td>
<td>761</td>
<td>258.6976</td>
<td>134.97</td>
</tr>
<tr>
<td>Submission Age (mon.)*</td>
<td>0</td>
<td>123</td>
<td>693</td>
<td>167.7738</td>
<td>128.5182</td>
</tr>
</tbody>
</table>

Note: Author age is defined as the number of months since the publishing of the author’s first paper; submission age is defined as the number of months from the publishing of the author’s first paper until the publishing of the paper in concern.
4.2 Authors’ Submission Behavior:

Aggregate Data

In this section, we will test the model in terms of authors’ submission behavior. We will show the existence of reputation effect and self-refereeing with the aggregate data. We will first illustrate with graphics before we introduce the empirical model.

I. The Graphic Evidence

(1) Self-Refereeing

Figure 4.2 shows authors self-referee and bad authors are more likely to self-referee. We define good papers to be those with quality above 75th percentile among all papers and bad papers to be those with quality below 25th percentile. As we can see, for authors of different range of quality, good papers are always more likely to be submitted around the deadline (self-refereeing). Both curves show a general decreasing trend and the speed of decreasing is higher for good papers, which means bad authors are more likely to self-referee.

(2) Reputation Effect

We will illustrate the implications of reputation effect. The reputation effect implies, for any paper quality, authors submitting away from the deadline are more likely to be good
Figure 4.2: The evidence for self-refereeing. All authors self referee and bad authors are more likely to self referee. The data points on each curve shows the fraction of D submissions for authors with quality in $[0, 10\%)$, $[10\%, 25\%)$, $[25\%, 50\%)$, $[50\%, 75\%)$, $[75\%, 90\%)$ and $[90\%, 1]$ among all authors.

Authors. We find initial evidence in Figure 4.3. More specifically, Figure 4.3 shows the medians of the coauthor-adjusted h-index for authors who submitted away from the deadline and around the deadline respectively as a function of the paper quality. It shows authors submitting away from the deadline have a higher coauthor-adjusted h-index and hence are better authors. Table 5.2 shows, by ranksum test, this comparison is significant at 1\% level for the best papers, significant at 2\% level for papers with quality between 25th percentile and 50th percentile, 75th percentile and 90th percentile among all papers. It is still significant at 10\% for papers with quality between 50th percentile and 75th percentile. For other paper quality, even though the comparison is not significant within 10 percentile significance level, the z-value is still positive.
Figure 4.3: The evidence for reputation effect. The data points on each curve shows the median coauthor-adjusted h-index for all authors who submitted papers with quality in [0, 10%), [10%, 25%), [25%, 50%), [50%, 75%), [75%, 90%) or [90%, 1] among all papers.

(3) Reputation Effect By Reputation

The graphic evidence for self-refereeing shows that good authors self-referee, which implies authors’ reputation must be intermediate for their submission behavior to show both self-refereeing and reputation effect according to Proposition 4. We confirm this prediction by graphically checking the existence of reputation effect for authors of different reputations. We define authors with intermediate reputation as those with reputation between 25 percentile and 75 percentile of all authors. Figure 4.4 shows, for each group of authors, those submitted away from the deadline still have higher quality. However, as shown in the bottom figure of Table 5.2, this comparison is only significant for authors with intermediate reputations. Moreover, in comparison to the top figure of Table 5.2, for authors with intermediate reputation, even though the sample size is smaller, the significance level is generally
increased. More specifically, except for papers with quality in $[25\%, 50\%]$ and $[90\%, 1]$, the p-value for all other groups all become smaller. Moreover, among these four groups, except for those with paper quality in $[0, 25\%)$, the h-index comparisons are all significant within 1% significance level. For papers with the highest quality ($[90\%, 1]$), the h-index comparison is still significant within 5% percent significance level.

Figure 4.4: Reputation effect by reputation. A significant reputation effect only exists among authors with intermediate reputations. Low reputations refer to $[0, 25\%]$ of all reputations, intermediate reputations refer to $(25\%, 75\%]$ of all reputations, and high reputations refer to $(75\%, 1]$ of all reputations.
II. Empirical Modelling

Based on the idea of reputation effect and self refereeing, we can set up a static empirical model as follows.

\[
subtime_{ip} = \alpha_0 + \alpha_1 q_i + \alpha_2 q_p + \alpha_3 q_i q_p + \alpha_4 q_{ip} + \alpha_5 nau_p + \mu_{tz} + \mu_{weekday} + \mu_{season} + \mu_{list} + \epsilon_{ip}
\]  

(4.2.1)

where

- \( subtime_{ip} = 1 \) if paper \( p \) is submitted by author \( i \) within 30 minutes before the deadline or within 30 minutes after the deadline (D submission), and \( subtime_{ip} = 0 \) (ND submission) if otherwise;
- \( q_i, q_p \) and \( q_{ip} \) are author \( i \)'s quality, paper \( p \)'s quality and author \( i \)'s reputation when she submits paper \( p \);
- \( nau_p \) is the number of authors for paper \( p \);
- \( \mu_{tz} \) and \( \mu_{weekday} \) are timezone and weekday fixed effects. \( \mu_{season} \) is the season fixed effect, where December and January are taken to be winter, February through May to be spring, June through August to be summer, and September through November to be fall; \( \mu_{list} \) is the announcement fixed effect and shows what day the paper is announced;
- \( \epsilon_{ip} \) incorporates all other variables influencing \( subtime_{ip} \).
Using these notations, we make the following predictions for authors with intermediate reputations. Let $X$ be the vector of all regressors. Reputation effect is satisfied if, for all $q_p$,

$$\frac{\partial \text{prob}(D \mid X)}{\partial q_i} = \alpha_1 + \alpha_3 q_p < 0$$

and self-refereeing is satisfied if, for some $q_i$,

$$\frac{\partial \text{prob}(ND \mid X)}{\partial q_p} = \alpha_2 + \alpha_3 q_i > 0$$

III. Empirical Results

Table 4.2 shows the existence of both reputation effect and self-refereeing. In column (1), authors of all reputations are included in the sample. We assume the error term in each period and/or the unobserved author heterogeneity are uncorrelated to independent variables. The negative coefficients of both author quality and the interaction between author quality and paper quality show the existence of reputation effect. The positive coefficient of paper quality implies authors self-referee at least for some papers. Since the p-value for Wald test for 99 percentile of all paper quality is over 0.1, at least 99 percent of all authors in the sample (weakly) self-referee within 10% significance level. The negative coefficient of the interaction between author quality and paper quality means bad authors are more likely to self-referee, which confirms our graphic illustration. The negative coefficient of Author Age shows, giving the same submission age, older authors are less likely to submit around the deadline. The positive coefficient of Submission Age shows, among authors of the same
author age, those with longer submission age are more likely to submit around the deadline, which implies the possible existence of learning. By proposition 4, the reputation effect and self-refereeing should only exist among authors with intermediate reputations. In column (2), we do analysis for authors of intermediate reputations. More specifically, we include only authors with reputation between 10 percentile and 90 percentile among all authors. In the OLS analysis of sub-column (i), the bigger absolute values of the coefficients of author quality and the interaction between author quality and paper quality shows reputation effect becomes stronger. The bigger coefficient of paper quality shows low quality authors are more likely to self-referee than those in the sample of column (1). The p-value for Wald test is still over 0.1. In sub-column (ii), we run author fixed effect regression to test the assumption that the unobserved heterogeneity is uncorrelated to dependent variables, assuming the error term in each period is uncorrelated to the dependent variables. The coefficients on paper quality and the interaction between paper quality and author quality are similar to those in sub-column (i), but the coefficient on submission age differs significantly. One possible reason is the unobserved author heterogeneity is correlated to the dependent variables, especially the submission age. Moreover, the positive coefficient on submission age in column (i) suggests learning. Since the lagged submission time depends on author quality and the interaction between author quality and paper quality, the estimation in (i) and (ii) are inconsistent if the lagged dependent variable is the omitted variable. In sub-column (iii), we include the lagged submission time in the pooled OLS regression. The coefficients are similar to those in sub-column (i). The positive coefficient on the lagged submission time shows the last period submission behavior really have a positive influence on the current submission. In sub-column (iv), we show the Arellano-Bond estimation, which confirms the estimation in
(iii) for reputation effect and self-refereeing, though the estimation is only significant within 10% significance level.

4.3 Authors’ Submission Behavior:

Different Ages

Even though we identify both self-refereeing and reputation effect in the last section, it is possible not all authors follow the same model. For example, very young authors have the strongest incentive to submit around the deadline to draw readers’ attention to their name; the oldest authors have no incentive to submit around the deadline because their reputation has already been established. In this section, we will analyze the submission behavior for authors of different ages. We differentiate authors based on author age and the number of papers submitted. Each group of authors submitted around 25 percent of all papers. Table 4.3 gives a detailed information about the division of authors. As is shown in Figure 4.5, reputation effect only holds for young authors. Table 5.3 shows, for papers with the quality below 25 percentile of all paper quality, authors submitting away from the deadline have higher quality, which is significant within 10 percent significance level. For paper quality between 75 percentile and 90 percentile, this comparison is significant within 6 percent significance level. However, for other groups of authors, either authors submitting around the deadline have significantly higher author quality or there is no significant difference. It suggests we should do analysis for each group separately.
<table>
<thead>
<tr>
<th>Regressors</th>
<th>(1)</th>
<th>(2)</th>
<th>(iii)</th>
<th>(iv)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>OLS</td>
<td>All</td>
<td>All</td>
<td>All</td>
</tr>
<tr>
<td>OLS</td>
<td>Inter. Rep.</td>
<td>(10% - 90%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OLS</td>
<td>OLS</td>
<td>FE</td>
<td>Lagged OLS</td>
<td>Arellano-Bond</td>
</tr>
<tr>
<td>Author Quality (AQ)</td>
<td>0.0212201</td>
<td>0.0327735</td>
<td>0.0357626</td>
<td></td>
</tr>
<tr>
<td>Paper Quality (PQ)</td>
<td>0.0003997</td>
<td>−0.000601</td>
<td>0.00005657</td>
<td></td>
</tr>
<tr>
<td>PQ × AQ</td>
<td>0.00004234</td>
<td>−0.0007711</td>
<td>0.00000829</td>
<td>−0.0008480</td>
</tr>
</tbody>
</table>

Table 4.2: Estimations with authors of all ages. Other controls are reputation, the number of authors, announcement fixed effect, timezone fixed effect, weekday fixed effect and season fixed effect. All submissions on Saturday and Sunday are excluded. Wald tests are for 99 percentile of author quality among each sample. The standard errors in the parenthesis are clustered at author level.
<table>
<thead>
<tr>
<th>Sub. Age</th>
<th>0.0011901**</th>
<th>0.0009686**</th>
<th>0.0005144</th>
<th>0.00089**</th>
<th>0.0010267</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(0.0002842)</td>
<td>(0.000313)</td>
<td>(0.0003504)</td>
<td>(0.0003266)</td>
<td>(0.0008573)</td>
</tr>
<tr>
<td>Author Age</td>
<td>-0.0014818**</td>
<td>-0.0012485**</td>
<td></td>
<td>-0.0011423**</td>
<td>-0.0011708</td>
</tr>
<tr>
<td></td>
<td>(0.0002836)</td>
<td>(0.0003103)</td>
<td></td>
<td>(0.0003248)</td>
<td>(0.0008187)</td>
</tr>
<tr>
<td>Sub. Time(_{-1})</td>
<td></td>
<td></td>
<td>0.1375279**</td>
<td>-0.0212652</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.0252897)</td>
<td>(0.0308199)</td>
<td></td>
</tr>
<tr>
<td>Wald Test (P-value)</td>
<td>0.121</td>
<td>0.186</td>
<td>0.173</td>
<td>0.187</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>4191</td>
<td>3350</td>
<td>3350</td>
<td>2903</td>
<td>2473</td>
</tr>
</tbody>
</table>
Figure 4.5: Young authors show the strongest reputation effect. We define youngest authors as those with author age at most 177 months, young authors with author age in $[177, 240)$ months, old authors with author age in $[240, 373)$ months and oldest authors with author age at least 373 months. Each group of authors submitted similar numbers of papers.
Table 4.3: The Division of Authors*

<table>
<thead>
<tr>
<th>Category</th>
<th>Author Age</th>
<th>Number of Papers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Range</td>
<td>Number</td>
</tr>
<tr>
<td>Youngest Authors</td>
<td>[0, 177]</td>
<td>272</td>
</tr>
<tr>
<td>Young Authors</td>
<td>[177, 240]</td>
<td>157</td>
</tr>
<tr>
<td>Old Authors</td>
<td>[240, 373]</td>
<td>159</td>
</tr>
<tr>
<td>Oldest Authors</td>
<td>[373, 761]</td>
<td>156</td>
</tr>
</tbody>
</table>

Table 4.4 confirms the graphic prediction. We find reputation effect and self-refereeing only exist among young authors with intermediate reputations. Both the degree of reputation effect and self-refereeing is strictly stronger than that in Table 4.2 when we do analysis for aggregate data. The intermediate reputations are defined to be between 10 percent and 75 percent of all young authors. Column (1) shows the estimation from pooled OLS. The higher absolute values of the coefficient on author quality and the interaction between author quality and paper quality than Table 4.2 means reputation effect becomes stronger. The bigger coefficient on paper quality implies bad authors are more likely to self-referee than what is shown in Table 4.2. In column (2), we run pooled OLS with lagged submission time. The coefficient on the lagged submission time is positive, which implies the submission behavior shows the pattern of inertia. In column (3), we show Arellano-Bond estimation. The two columns show similar estimation results on the coefficients of paper quality and the interaction between paper quality and author quality. Table 4.5 shows OLS estimation with lagged submission time for all other groups of authors with intermediate reputations. The
absolute values of the key variables are much smaller than Table 4.4. Moreover, there is neither significant reputation effect nor significant self-refereeing.

4.4 Discussions

I. Deadline Effect

In this section, we will show that self refereeing we identified is not caused by the deadline effect as we pointed out in the end of section 3.4.2. The deadline effect means submitters choose D submissions (before the deadline) only because they want to finish their paper before the deadline and these submissions gather more citations than ND submissions. The idea is related to the procrastination effect as pointed out by Ginsparg and Hague [2010]. Authors take 4pm as a deadline but procrastinate and keep working on their paper until the last minute. These papers get more citations probably because they address “hot” topics. The reason can be purely psychological or these authors made an agreement with other researchers for their paper to be announced on a given day. When we assume all rules of the game are common knowledge among researchers, this possibility is ruled out by the option of “booking” a position. More specifically, after submitting the paper and before the deadline, submitters can resubmit and the position is based on the original submission time. Amusingly, we have been told that some authors would submit a blank paper, simply to reserve a submission [15]. We call the last submission time before the deadline as the final submission time. Anyone trying to catch the deadline should have the incentive to book a position by submitting far away from the bottom deadline and resubmit right before the
deadline. Therefore, papers originally submitted right before the deadline are not subject to the deadline effect. For the rest of this section, we mainly discuss the case when the policy of resubmitting is not common knowledge.

Authors might not realize the existence of such a policy or might only gradually learn it because there are many rules. For example, now the option to edit and replace a submission is only 1 out of 11 instructions. When authors submit right before the deadline without knowing this policy, it is possible this submission is driven by the deadline effect. We identify authors who know this policy by checking their submission history. If they ever booked a position before the current submission, we assume they know this policy. Their current submission has strategic considerations. In this section, for D submissions right before the deadline, we will only keep these submissions with strategic concerns. More specifically, we say one author books a position if the final submission time is more than 3 minutes after the original submission time. We give 3 minutes for the system’s possible mistakes. We divide D submissions before the deadline into two parts: strategic papers (SP) and other papers (OP). For strategic papers, their submitters booked positions for the previous submissions.

Figure 4.6 shows authors’ self-refereeing before 4pm is due to SP. Among 268 D submissions at bottom positions, 158 are SP. Paper quality of D submissions stochastically dominates ND submissions within 91 significance level. SP shows much higher median paper quality than ND. And, paper quality of D submissions stochastically dominates SP within 99 significance level. After removing SP, paper quality of OP is lower than ND, through this comparison is not significant within 10 percent significance level. Therefore, we should expect a higher degree of self refereeing.
Note: by ranksum test, D stochastically dominates ND within 9% significance level;

SP dominates ND within 1% percent significance level.

Table 4.6 confirms there exist a stronger self-refereeing and reputation effect if we remove OP from the sample. Moreover, it shows D submissions before 4pm play an important role in the finding of self-refereeing and reputation effect. We use pooled OLS with lagged submission time. In the first sub-column of column (1), OP are removed from the sample. If we compare sub-column (iii) under column (2) in Table 4.4, the bigger coefficient of paper quality suggests a stronger degree of self-refereeing at least for bad authors. The bigger absolute value of the coefficient on the interaction between paper quality and author quality implies a stronger reputation effect. The second sub-column of column (1) shows the regression with all D submissions at bottom positions removed. As we can see, both reputation effect and self refereeing are significantly weakened. In column (2), we do analysis for young authors. In comparison to column (3) of Table 4.4, both reputation effect and self-refereeing are stronger. Moreover, it confirms our findings before that young authors have a higher
degree of self refereeing and reputation effect. Column (2) also shows that D submissions before 4pm play a role in reputation effect and self refereeing, though the effect is weaker than column (1). The weaker result is possibly due to the much smaller sample size. More specifically, the absolute value of the coefficient on the interaction between paper quality and author quality is bigger in the first sub-column than that in the second sub-column, which shows D submissions before 4pm plays a role in reputation effect. D submissions before 4pm also play a role. In the second sub-column of column (2), the smaller coefficient of paper quality than the first sub-column of column (2) shows that at least for bad papers, the degree of self refereeing is stronger. Moreover, for both column (1) and (2), the regressions in the second sub-columns are always less significant.

II. Naive and Sophisticated Authors

So far, after including the lagged submission time as one independent variable, we have assume that the error term $\epsilon_{ip}$, which might include the unobserved author heterogeneity $\mu_i$, is not correlated with all controls. But this assumption may not be true. Some authors may choose to submit randomly due to philosophical reasons: they care about citations, but they are against the idea of strategically submitting to get more citations. We call these authors as naive submitters and the rest as sophisticated submitters. This characteristic $\mu_i$ is potentially correlated to author quality. Our estimation without taking this endogeneity into consideration will yield an inconsistent estimation. Since this characteristic is not observable, in the following analysis, we only include sophisticated submitters in the sample.
We identify sophisticated submitters in two ways. First, we assume authors who ever booked a position are more likely to be sophisticated submitters. By booking positions, these authors show that they play around the rules. Since these rules are common knowledge among these authors, there is no deadline effect. We choose authors who ever book a position. The estimation results are in Table 4.7 where we can clearly identify both reputation effect and self refereeing. In addition, the degree of reputation effect is strictly stronger than sub-column (iii) under column (2) in Table 4.2 and column (3) of Table 4.4. Second, we eliminate authors who are more likely to submit randomly. Suppose with probability $P$, an author randomly submits her paper around the deadline. For any such author with $n$ papers, since each paper is independent, the probability of submitting $m$ papers around the deadline is $Q = C_n^m P^m (1 - P)^{n-m}$. For any author with $n$ papers and $m$ D submissions, if $Q(n, m) < 10\%$, this author is strategic authors. Table 4.8 shows the results from the pooled OLS with lagged submission time for $P = 1/8$ and $p = 1/12$, that is, we assume an author works for 8 hours and 12 hours a day. Both the degrees of self refereeing and reputation effect are much stronger than sub-column (iii) under column (2) in Table 4.2.

### 4.5 Positional Effect

In this section, we will test the model from the readers’ side. Our model predicts the existence of positional effect, i.e., papers of the same quality will be cited more if they are at a position with papers of higher average quality (paper quality effect) and/or higher visibility (visibility effect). We find the existence of positional effect for submissions a few hours before
and after 4pm. Moreover, we identify visibility effect for top positions and paper quality effect through papers submitted hours before 4pm.

I. Positional Effect Involving Top Positions

For the positional effect of the top positions, we put the following restrictions. First, to reduce the influence of the positional effect involving bottom positions, we select samples according to the following criteria. (1) papers are submitted after 4pm but before 12am EST. Other papers have a smaller probability of getting top positions, but higher probability of getting bottom positions. (2) the number of papers on the same list is at least 7. Otherwise, if some lists are very short, the positional effect involving bottom positions can dominate. (3) papers are on the first half of the list. That is to say, if we number papers according to their order on the list from the top to the bottom, papers in our sample have positions less than or equal to half of the total number of papers on the same list. We put the third restriction because papers submitted within a few hours after 4pm may get bottom positions if there are not many authors submitting on the same day. Second, we choose papers of low quality because if papers at top positions on average have a higher number of citations, the positional effect is stronger for low quality paper. Given the restrictions, we run the following regression.

\[
scitation_i = \alpha_0 + \alpha_1 p_i + \alpha_2 top_i + \alpha_3 length + \mu_{weekday} + \epsilon_i
\]
where $scitation_i$ is the scaled number of citations as defined in the last section, $p_i$ is the paper quality, $top_i$ equals 1 if paper $i$ is at top positions and 0, and otherwise, and $\text{length}$ is the number of papers on the same list as paper $i$. A positive $a_2$ shows the positional effect at top positions. We find a significant positional effect. As shown in column (1) of Table 4.9, $\hat{a}_2 > 0$ for both papers with quality below 10 percentile and 25 percentile among all papers at top positions.

Given the current data structure, we can also infer the existence of visibility effect from the citation pattern in the bottom figure of Figure 4.1. If there does not exist visibility effect for top positions, since positional effect exists, paper quality effect must exist. From the bottom figure of Figure 4.1, we can observe that the scaled number of citations at position 1 or 2 is not significantly higher than that at position 3. If we take papers submitted within 12 hours after 4pm at position 1 (2) as group 1 and those at position 3 as group 0, the z-value from ranksum test is 0.095 (1.753) and the p-value is 0.9244 (0.0797). Papers at position 3 are cited more than papers at position 2. The higher number of citations at position 3 is due to the higher quality of papers submitted between 150 and 720 minutes after 4pm than those between 30 and 150 minutes (significant within 8 percent significance level). If we restrict the sample to papers submitted within 150 minutes after 4pm, the citation advantage of position 3 disappears. The z-value for the comparison between position 1 (2) and 3 becomes $-0.802$ (0.48) and it is not significant within 10 percent significance level (the p-value is 0.4227 (0.6315)). Therefore, papers at position 3 submitted during the same time window as those at position 1 and 2 should be cited more, which contradicts the positional effect for papers of low quality.
II. Positional Effect Involving Bottom Positions

Similar to the last section, we run the following regression.

\[ scitation_i = \alpha_0 + \alpha_1 p_i + \alpha_2 \text{bottom}_i + \alpha_3 \text{length} + \mu_{\text{weekday}} + \epsilon_i \]

We find a significant positional effect, \( \hat{\alpha}_2 > 0 \), by making the following restrictions. We first let \( \text{bottom}_i \) equals 1 if paper \( i \) has bottom positions and equals 0 if paper \( i \) has position L-3. We restrict to L-3 because both the paper quality and the number of citations are strikingly lower for this position than bottom positions. Moreover, similar to the last section, we also restrict the sample to lists with at least 7 papers, papers at the second half of the list, and papers of low quality. Moreover, we restrict the sample to be submitted within 180 minutes before 4pm. The results are summarized in column (2) of Table 4.9. The bottom position for the second and third sub-column is L-2. We choose L-2 for a stronger positional effect. L-1 has a lower number of citations than L-2 as visually shown in the bottom figure of Figure 3.1, though the citation comparison is not significant within 10 percent significance level from Ranksum test. \( \hat{\alpha}_2 > 0 \) shows the existence of positional effect for papers with quality below 10th percentile or 25th percentile among all papers at bottom positions.

We can show the existence of paper quality effect from the observation that papers at L-1 position is cited less than those at L-2 positions in the bottom figure of Figure 4.1. The result is shown in column (3). \( \text{bottom}_i \) equals 1 if paper \( i \)'s position is L-1 and 0 if the position is L-2 and L-4. Papers at L-2 and L-4 have on average a higher number of citations than papers at L-1, though the comparison is not significant within 10 percent significance level. Moreover, we restrict the sample to papers submitted within 90 minutes before 4pm.
The negative coefficient on $bottom_i$ shows the existence of paper quality effect.

III. Relation To The Literature

Our analysis resembles Dietrich [2008, a, b] and Ginsparg and Haque [2009, 2010], but differs significantly. First, our definition of positional effect means papers of the same quality are cited differently at different positions, while the positional citation effect in the literature means papers at different positions are cited differently. Second, the visibility effect as defined in the literature is misleading. In the literature, it is claimed visibility effect exists if the number of citations for papers submitted away from 4pm is on average higher for those at top/bottom positions than positions in the middle of the list. This definition is misleading for the following two reasons. On one hand, they assume papers submitted away from 4pm have the same quality, but papers submitted away from 4pm at different positions may have different quality. Maybe papers submitted closer to 4pm have higher quality than papers submitted further away from 4pm and hence the visibility effect they defined is caused by the higher quality of papers submitted closer to 4pm. On the other hand, the visibility effect they defined is at least partly due to the paper quality effect because papers at top or bottom positions are on average better.
<table>
<thead>
<tr>
<th>Dependent Variable: Submission Time</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regressors</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OLS Lagged OLS Arellano-Bond</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Author Quality (AQ)</td>
<td>$-0.1146233$</td>
<td>$-0.109759$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$(0.0996714)$</td>
<td>$(0.1023467)$</td>
<td></td>
</tr>
<tr>
<td>Paper Quality (PQ)</td>
<td>$0.0013258^{**}$</td>
<td>$0.0020008^{**}$</td>
<td>$0.0018819^{*}$</td>
</tr>
<tr>
<td></td>
<td>$(0.0004571)$</td>
<td>$(0.0005869)$</td>
<td>$(0.0007321)$</td>
</tr>
<tr>
<td>PQ × AQ</td>
<td>$-0.0018708^{**}$</td>
<td>$-0.0026345^{**}$</td>
<td>$-0.002795^{*}$</td>
</tr>
<tr>
<td></td>
<td>$(0.0007048)$</td>
<td>$(0.0009938)$</td>
<td>$(0.0014027)$</td>
</tr>
<tr>
<td>Sub. Age</td>
<td>$0.0003091$</td>
<td>$-0.0000603$</td>
<td>$-0.0004096$</td>
</tr>
<tr>
<td></td>
<td>$(0.0006699)$</td>
<td>$(0.0007891)$</td>
<td>$(0.0019275)$</td>
</tr>
<tr>
<td>Author Age</td>
<td>$-0.0019694$</td>
<td>$-0.0013081$</td>
<td>$-0.0034311$</td>
</tr>
<tr>
<td></td>
<td>$(0.0013643)$</td>
<td>$(0.0013782)$</td>
<td>$(0.0022478)$</td>
</tr>
<tr>
<td>Sub. Time$_{-1}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$0.1387973^{**}$</td>
<td>$-0.0314282$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$(0.0503587)$</td>
<td>$(0.0551711)$</td>
<td></td>
</tr>
<tr>
<td>Wald Test (P-value)</td>
<td>0.108</td>
<td>0.237</td>
<td>0.235</td>
</tr>
<tr>
<td>N</td>
<td>685</td>
<td>606</td>
<td>524</td>
</tr>
</tbody>
</table>

Table 4.4: Estimation With Young Authors of Intermediate Reputations. Other controls are reputation, the number of authors, announcement fixed effect, timezone fixed effect, weekday fixed effect and season fixed effect. All submissions on Saturday and Sunday are excluded. Wald tests are for author quality at 99 percentile among all author in each regression. For column (4), p-value for Wald test is based on all authors in Column (3). All standard errors are clustered at author level.
### Table 4.5: Estimation With Other Authors At Intermediate Reputations

Other authors refer to authors other than young authors. The estimation results for young authors are in Table 6. Other controls are reputation, the number of authors, announcement fixed effect, timezone fixed effect, weekday fixed effect and season fixed effect. All submissions on Saturday and Sunday are excluded. All estimations are for authors with reputation between 10 percentile and 75 percentile among all authors in each group. All standard errors are clustered at the author level.

<table>
<thead>
<tr>
<th>Regressors</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Youngest</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Author Quality (AQ)</td>
<td>0.0774156</td>
<td>-0.1035753</td>
<td>0.4669632</td>
</tr>
<tr>
<td></td>
<td>(0.1106162)</td>
<td>(0.2916462)</td>
<td>(0.3618825)</td>
</tr>
<tr>
<td>Old</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paper Quality (PQ)</td>
<td>0.0004493</td>
<td>0.0005013</td>
<td>0.0009961</td>
</tr>
<tr>
<td></td>
<td>(0.0003079)</td>
<td>(0.0006481)</td>
<td>(0.0007052)</td>
</tr>
<tr>
<td>Oldest</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PQ × AQ</td>
<td>-0.0007238</td>
<td>-0.0004275</td>
<td>-0.0021584</td>
</tr>
<tr>
<td></td>
<td>(0.000442)</td>
<td>(0.0008946)</td>
<td>(0.0013393)</td>
</tr>
</tbody>
</table>

| N                  | 540         | 577         | 600         |
Table 4.6: Estimation without deadline effect

(Pooled OLS With Lagged Submission Time)

<table>
<thead>
<tr>
<th>Dependent Variable: Submission Time</th>
<th>(1)</th>
<th>(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regressors</td>
<td></td>
<td></td>
</tr>
<tr>
<td>with SP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>without SP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>with OP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>without OP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>inter. rep</td>
<td></td>
<td></td>
</tr>
<tr>
<td>((10%, 90%))</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Young</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Author Quality (AQ)</td>
<td>-0.0210323</td>
<td>0.0150873</td>
</tr>
<tr>
<td></td>
<td>(0.0620967)</td>
<td>(0.0769105)</td>
</tr>
<tr>
<td>Paper Quality (PQ)</td>
<td>0.000728**</td>
<td>0.0005448*</td>
</tr>
<tr>
<td></td>
<td>(0.0002362)</td>
<td>(0.0002508)</td>
</tr>
<tr>
<td>PQ × AQ</td>
<td>-0.0009785**</td>
<td>-0.0006925†</td>
</tr>
<tr>
<td></td>
<td>(0.0003403)</td>
<td>(0.0003563)</td>
</tr>
<tr>
<td>Author Quality (AQ)</td>
<td>-0.0689589</td>
<td>-0.0809855</td>
</tr>
<tr>
<td></td>
<td>(0.0769105)</td>
<td>(0.1526219)</td>
</tr>
<tr>
<td>Paper Quality (PQ)</td>
<td>0.0024325**</td>
<td>0.0022333*</td>
</tr>
<tr>
<td></td>
<td>(0.000561)</td>
<td>(0.000858)</td>
</tr>
<tr>
<td>PQ × AQ</td>
<td>-0.0034114**</td>
<td>-0.002945*</td>
</tr>
<tr>
<td></td>
<td>(0.0009181)</td>
<td>(0.001666)</td>
</tr>
<tr>
<td>N</td>
<td>2834</td>
<td>2108</td>
</tr>
<tr>
<td>Wald Test (P-value)</td>
<td>0.535</td>
<td>0.56</td>
</tr>
</tbody>
</table>

Note: SP are D submissions whose submitters “know” the existence of a policy for resubmitting. OP are the rest of D submissions. Other controls are reputation, lagged submission time, author age, author age upon submission, the number of authors, announcement fixed effect, timezone fixed effect, weekday fixed effect and season fixed effect. All submissions on Saturday and Sunday are excluded. Standard errors are clustered for each author. P-values for Wald tests are for all authors with quality at 95 percentile of all authors in each regression.
### Table 4.7: Estimation With Only Authors Ever Booking A Position

<table>
<thead>
<tr>
<th>Regressors</th>
<th>(1)</th>
<th>(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>inter. rep</td>
<td>inter. rep</td>
<td></td>
</tr>
<tr>
<td>(10%, 90%)</td>
<td>(10%, 75%)</td>
<td></td>
</tr>
<tr>
<td>Young</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Author Quality (AQ)</td>
<td>-0.1775366*</td>
<td>-0.3099829*</td>
</tr>
<tr>
<td></td>
<td>(0.0762475)</td>
<td>(0.1336008)</td>
</tr>
<tr>
<td>Paper Quality (PQ)</td>
<td>0.0006475*</td>
<td>0.0019882**</td>
</tr>
<tr>
<td></td>
<td>(0.0002611)</td>
<td>(0.0006285)</td>
</tr>
<tr>
<td>PQ × AQ</td>
<td>-0.0008907*</td>
<td>-0.002808*</td>
</tr>
<tr>
<td></td>
<td>(0.0003475)</td>
<td>(0.0010694)</td>
</tr>
<tr>
<td>N</td>
<td>2092</td>
<td>484</td>
</tr>
<tr>
<td>Wald Test (p-value)</td>
<td>0.428</td>
<td>0.147</td>
</tr>
</tbody>
</table>

Other controls are reputation, lagged submission time, author age, author age upon submission, the number of authors, announcement fixed effect, timezone fixed effect, weekday fixed effect and season fixed effect. All submissions on Saturday and Sunday are excluded. All standard errors are clustered at the author level. The Wald test for column (1) is for paper quality at 90 percentile among all authors in the regression. For column (2), Wald test is for paper quality at 99 percentile among all authors in this regression.
(Pooled OLS With Lagged Submission Time)

<table>
<thead>
<tr>
<th>Dependent Variable: Submission Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regressors</td>
</tr>
<tr>
<td>$P = 1/8$</td>
</tr>
<tr>
<td>Author Quality (AQ)</td>
</tr>
<tr>
<td>(0.1644857)</td>
</tr>
<tr>
<td>Paper Quality (PQ)</td>
</tr>
<tr>
<td>(0.0004894)</td>
</tr>
<tr>
<td>PQ $\times$ AQ</td>
</tr>
<tr>
<td>(0.0006026)</td>
</tr>
</tbody>
</table>

| $N$ | 706 | 928 |
| P-value (Wald test) | 0.927 | 0.153 |

Table 4.8: Estimation With Only Sophisticated Submitters. Other controls are reputation, lagged submission time, author age, author age upon submission, the number of authors, announcement fixed effect, timezone fixed effect, weekday fixed effect and season fixed effect. All submissions on Saturday and Sunday are excluded. The sample follows the following constraints. First, all reputations are between 10 percentile and 90 percentile among all authors; second, the probability of any author to follow a random schedule is at most 10%. Wald tests are for papers with paper quality of 90 percentile among all authors.
### Table 4.9: Positional Effect

Other controls are paper quality, the number of papers on the same list, and weekday fixed effect. Column (1) ((2)) shows the estimation of positional effect. The different sub-columns show estimations with papers of different low quality. Papers are chosen to have quality less than 10th or 25th percentile among all papers at top/bottom positions. All estimations under Column (2) take L-3 as the non-bottom position. The second and the third sub-columns take L-2 as the bottom position. Column (3) shows the paper quality effect with papers submitted 1.5 hours before 4pm. L-1 is the bottom position; L-2 and L-3 are non-bottom positions.

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regressors</td>
<td>10th</td>
<td>25th</td>
<td>25th</td>
</tr>
<tr>
<td></td>
<td>10th</td>
<td>25th</td>
<td>25th</td>
</tr>
<tr>
<td>top</td>
<td>1.790632(^+)</td>
<td>1.664491(^*)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1.069985)</td>
<td>(0.7485947)</td>
<td></td>
</tr>
<tr>
<td>bottom</td>
<td>2.439694(^+)</td>
<td>6.086363(^**)</td>
<td>2.711317(^+)</td>
</tr>
<tr>
<td></td>
<td>(1.44529)</td>
<td>(2.115134)</td>
<td>(1.60449)</td>
</tr>
<tr>
<td>time constraint</td>
<td>6hr.</td>
<td>6hr.</td>
<td>3hr.</td>
</tr>
<tr>
<td>N</td>
<td>133</td>
<td>355</td>
<td>258</td>
</tr>
</tbody>
</table>
Conclusion

In this thesis, we show that, in the field of high energy physics-theory, all authors self-referee, *i.e.*, among authors of the same quality, those with good papers are more likely to submit around the deadline. And authors’ self-refereeing is driven by their reputation concerns. Authors with bad papers submitting away from the deadline now to gain from the reputation improvement in the future. Among authors with papers of the same quality, good authors are more likely to submit away from the deadline. Therefore, even though suffering now from the smaller number of citations due to authors’ self-refereeing and the lower visibility for positions in the middle of the list, authors submitting away from the deadline gain from the reputation improvement in the future. Authors with good papers submit around the deadline because either good papers have a higher cost of being placed in the middle of the list or the reputation gain for authors with good papers is smaller because good papers are more likely to be from good papers or both. The incentive for authors with good papers to submit around the deadline is strictly stronger if readers’ attention to positions in the middle of the list is more limited. We also find bad authors are more likely to self-referee, which implies authors with good papers must also invest in reputation. With big reputation gain from investing in reputation, authors with good papers submit around the deadline only if readers’ attention is sufficiently limited. From studying authors’ citing behavior, we find evidence for readers’ limited attention to positions in the middle of the list.
We expect many other market places have a similar structure as arXiv.org for the field of high energy physics-theory. Therefore, our empirical observations, the theoretical model and the empirical testing can apply to many markets. For example, even though the underlying parameters for other fields using arXiv.org may differ from this field we study, under the same submission system, we may observe similar submission and citation pattern. The idea can apply to financial markets. For example, companies in the stock market may strategically choose when to announce earning news during any given trading day. Managers’ reputation concerns for their company may play an important role in the timing decision.

Our reduced form empirical analysis for authors’ submission behavior is consistent with the equilibrium strategy profile in the first period of the theoretical model. The second period is redundant because the equilibrium where everyone submits around 4pm is the unique equilibrium. A dynamic model, which takes into account factors such as authors’ entry and exit, and the corresponding structural analysis is waiting to be explored.
Appendix A

A1. Proof of Crossing-Order Condition: Define $g^*(\theta) = 1 - G(\theta)$, $h(\theta) = (1 - \epsilon)(1 - G(\theta))$, $f^*(\theta) = 1 - F(\theta)$ and $l(\theta) = (1 - \epsilon)(1 - F(\theta))$. If authors with bad paper wants to submit around the deadline, then

$$g^*(\bar{\theta}_{HV}) \geq h(\bar{\theta}_{LV}) \Leftrightarrow \bar{\theta}_{HV} \leq (g^*)^{-1}[h(\bar{\theta}_{LV})]$$

So, we have

$$f^*(\bar{\theta}_{HV}) - l(\bar{\theta}_{LV}) \geq f^*[(g^*)^{-1}(h(\bar{\theta}_{LV}))] - l(\bar{\theta}_{LV})$$

Let

$$L(\theta) = f^*[(g^*)^{-1}(h(\theta))] - l(\theta)$$

for $\theta \in (-\infty, +\infty)$. If $\theta \to +\infty$, $L(\theta) \to 0$.

$$L'(\theta) = (1 - \epsilon)g(\theta)[\frac{f(\theta)}{g(\theta)} - \frac{f((g^*)^{-1}(h(\theta)))}{g((g^*)^{-1}(h(\theta)))}$$

$g^*(\theta) > h(\theta)$ implies $\theta < (g^*)^{-1}(h(\theta))$. By MLRP, $L'(\theta) < 0$. Hence, $L(\theta) > 0$ for all $\theta$. Hence, $f^*(\bar{\theta}_{HV}) > l(\bar{\theta}_{LV})$.

A2. Proof of Proposition 3: Given the prior $q$, denote $V_2(s_2 \mid a, \theta_i, \theta_j, s_1, q)$ as any author’s payoff at $t = 2$ from position $s_2$ when he is type $a$ author, he has paper with quality $\theta_j$ at $t = 2$, paper with quality $\theta_i$ and position $s_1$ at $t = 1$, where $i,j \in \{H,L\}$. Then, his first period expected payoff from position $s_1$ is

$$V_1(s_1 \mid a, \theta_i, q) = u(s_1, \theta_i) + \beta E_{(\theta_j,s_2)}[V_2(s_2 \mid a, \theta_i, \theta_j, s_1, q)]$$
where the first period current payoff satisfies

\[
u(s_1, \theta_i) = \begin{cases} 
1 - F(\overline{\theta}_{HV}^1), & \text{if } s_1 = HV \text{ and } \theta_i = \theta_H; \\
1 - G(\overline{\theta}_{HV}^1), & \text{if } s_1 = HV \text{ and } \theta_i = \theta_L \\
(1 - \epsilon)(1 - F(\overline{\theta}_{LV}^1)), & \text{if } s_1 = LV \text{ and } \theta_i = \theta_H; \\
(1 - \epsilon)(1 - G(\overline{\theta}_{LV}^1)), & \text{if } s_1 = LV \text{ and } \theta_i = \theta_L; 
\end{cases}
\]

Since in the second period, all authors submit around the deadline, with probability \(1 - \eta\), any paper gets position \(LV\). However, if at position \(LV\), with probability \(\epsilon\), the paper will be lost. We have

\[
E(\theta_i, s_2)[V_2(s_2 | a, \theta_i, \theta_j, s_1, q)] = (1 - (1 - \eta)\epsilon)[p_a(1 - F(\overline{\theta}_{s_1}^{2H})) + (1 - p_a)(1 - G(\overline{\theta}_{s_1}^{2L}))]
\]

The first period expected payoff from taking action \(D\) and \(ND\) are respectively

\[
V(ND | a, \theta_i, q) = p_{nd}V_1(HV | a, \theta_i, q) + (1 - p_{nd})V_1(LV | a, \theta_i, q)
\]

\[
V(D | a, \theta_i, q) = (1 - p_d)V_1(HV | a, \theta_i, q) + p_dV_1(LV | a, \theta_i, q)
\]

So,

\[
V(D | a, \theta_i, q) - V(ND | a, \theta_i, q) = (1 - p_d - p_{nd})[V_1(HV | a, \theta_i, q) - V_1(LV | a, \theta_i, q)]
\]

Denote \(V_a^H(q) = V_1(HV | a, \theta_L, q) - V_1(LV | a, \theta_L, q)\) and \(V_a^L(q) = V_1(HV | a, \theta_H, q) - V_1(LV | a, \theta_H, q)\). Then,

\[
V_a^H = \epsilon + (1 - \epsilon)F(\overline{\theta}_{LV}^1) - F(\overline{\theta}_{HV}^1) + \beta(1 - (1 - \eta)\epsilon)[p_a(F(\overline{\theta}_{LV}^{2H}) - F(\overline{\theta}_{HV}^{2H})) + (1 - p_a)(G(\overline{\theta}_{LV}^{2H}) - G(\overline{\theta}_{HV}^{2H}))]
\]
\[ V_L^a = \epsilon + (1 - \epsilon)G(\bar{\theta}_{LV}^1) - G(\bar{\theta}_{HV}^1) + \beta(1 - (1 - \eta)\epsilon)[p_a(F(\bar{\theta}_{LV}^2) - F(\bar{\theta}_{LV}^2)) + (1 - p_a)(G(\bar{\theta}_{LV}^2) - G(\bar{\theta}_{HV}^2))] \]

\[ V_L^a \] for \((1, 0, 1, 1)\) satisfies the following properties.

First, if \( q > \frac{1 - \eta}{1 - p} \), \( V_L^a(q) \) strictly increases with \( q \). In other words, the incentive for authors with bad paper to invest in reputation decreases with \( q \). This immediately follows from the following factors. On one hand, the sacrifice for the investment in reputation increases with reputation \( q \). Papers at LV positions are not cited, that is, \( \bar{\theta}_{LV}^1 = +\infty \). Papers at HV positions on average have higher and higher quality, that is, \( \bar{\theta}_{HV}^1 \) decrease with \( q \), where \( p(\theta_H \mid HV) = \frac{pq + (1 - p)(1 - q)}{\eta} \). On the other hand, the reputation gain decreases with reputation \( q \). \( \bar{\theta}_{LV}^2 \) decreases with \( q \) because \( p(G \mid \theta_L, HV) = \frac{p_{nd}q(1 - p)}{p_{nd}q(1 - p) + p(1 - q)} \) increases with \( q \) for any given \( p_{nd} \) and \( p_{nd} = 1 - \frac{1 - \eta}{q(1 - p)} \) also increases with reputation \( q \).

Second, if \( R > 1/2 \), there exists \( \bar{q} > \frac{1 - \eta}{1 - p} \) such that \( V_L^G(\bar{q}) = V_L^B(\bar{q}) \) and \( V_L^G(q) \leq V_L^B(q) \) if and only if \( q < \bar{q} \) for all \( q > \frac{1 - \eta}{1 - p} \), where

\[ \bar{q}(p, \eta, R) = \frac{1 - \eta + p(x(p, R) - 1)}{1 - p + p(x(p, R) - 1)} \] (5.0.1)

and \( x(p, R) = \frac{2p - 1}{1 - p(1 - R)^2} \).

\[ V_L^G - V_L^B = (2p - 1)(M(\bar{\theta}_{HV}^2) - M(\bar{\theta}_{LV}^2)), \] where \( M(x) = G(x) - F(x) \) for \( x \in (-\infty, +\infty) \). \( M \) is symmetric with respect to \( x = \frac{\theta_H + \theta_L}{2} \) and also reaches its peak at \( x = \frac{\theta_H + \theta_L}{2} \). Therefore, for \( \bar{q} \geq \frac{1 - \eta}{1 - p} \),

\[ V_L^G(\bar{q}) = V_L^B(\bar{q}) \iff \bar{\theta}_{LV}^2 + \bar{\theta}_{LV}^2 = \theta_H + \theta_L \iff \]

\[ p(\theta_H \mid \theta_L, HV) = \frac{1}{1 + \frac{p}{1 - p(1 - R)^2}} \iff \]
\[ \bar{q}(p, \eta, R) = \frac{1 - \eta + p(x(p, R) - 1)}{1 - p + p(x(p, R) - 1)} \]

As \( R > 1/2 \), \( x(p, R) > 1 \) and then \( \bar{q} > \frac{1 - \eta}{1 - p} \). Therefore, if \( R > 1/2 \), equation (5) holds.

Since \( x(p, R) \) increases with \( R \), \( \bar{q} \) increases with \( R \). As \( R \to p \), \( x(p, R) \to +\infty \) and hence \( \bar{q} \to 1 \). For \( q \in (\frac{1 - \eta}{1 - p}, \bar{q}) \), \( V_L^G < V_L^B \); for \( q > \bar{q} \), \( V_L^G > V_L^B \). This is because \( \bar{\theta}_{LV}^G \) does not change with \( q \) for \( q > \frac{1 - \eta}{1 - p} \) and \( \bar{\theta}_{HV}^G \) decreases with \( q \). Therefore, \( M(\bar{\theta}_{LV}^G) \) does not change with \( q \) for \( q > \frac{1 - \eta}{1 - p} \) and \( M(\bar{\theta}_{HV}^G) \) increases with \( q \) as long as \( \bar{\theta}_{HV}^G \geq \frac{\theta_H + \theta_L}{2} \). For \( q > \bar{q} \), \( M(\bar{\theta}_{LV}^G) > M(\bar{\theta}_{HV}^G) \) and for \( q < \bar{q} \), \( M(\bar{\theta}_{LV}^G) < M(\bar{\theta}_{HV}^G) \).

Finally, \( V_H^a > V_L^a \). This is due to \( G(\bar{\theta}_{HV}^G) > F(\bar{\theta}_{HV}^G), \bar{\theta}_{LV}^G \geq \bar{\theta}_{LV}^G \) and \( \bar{\theta}_{HV}^G < \bar{\theta}_{HV}^G \).

To prove proposition 3, we will first show that for all \( \eta > \frac{4}{5} \), there exists \( R(\eta)(> 1/2) \) such that if \( R > R(\eta) \), \( p(\theta_H | HV) < R \) for \( q \leq \frac{1 - \eta}{1 - p} \). To show it, first suppose \( R > 1/2 \) such that according to the second property of \( V_L^a, \bar{q}(< 1) \) exists. Define \( \eta(R) = \frac{\frac{p^2}{(R+1)(1-p)^2}}{\bar{q}(R)} \).

When \( \eta = \eta(R) \), \( p(\theta_H | HV) = R \) for \( q = \frac{1 - \eta}{1 - p} \). Since \( p(\theta_H | HV) = \frac{\eta + (1-p)(1-q)}{\eta} \) decreases with \( \eta \) and increases with \( q \), for all \( q < \frac{1 - \eta}{1 - p} \) and \( \eta > \eta(R) \), \( p(\theta_H | HV) < R \). Since \( R < p \), \( \eta(R) > \frac{\frac{p^2}{3p+3p-1}}{3p+p^2} \). If \( p < 2/3 \), \( \frac{\frac{p^2}{3p+3p-1}}{3p+p^2} \) decreases with \( p \); if \( p > 2/3 \), \( \frac{\frac{p^2}{3p+3p-1}}{3p+p^2} \) increases with \( p \). Therefore, \( \frac{\frac{p^2}{3p+3p-1}}{3p+p^2} \) reaches the lowest level 4/5 at \( p = 2/3 \). Hence, \( \eta(R) > 4/5 \).

Therefore, for all \( \eta > 4/5 \), there exists \( R(\eta) \) such that for all \( R > R(\eta) \), \( p(\theta_H | HV) < R \) when \( q > \frac{1 - \eta}{1 - p} \), where \( \bar{R}(\eta) = \frac{p^2 - \eta(2p-1)}{\eta(1-p)} \).

Next, we show that there exist \( \bar{R}^*(\eta)(> \bar{R}(\eta)) \) and \( \sigma^*(R, \eta) \) for all \( R > \bar{R}^*(\eta) \) and \( \eta > 4/5 \) such that \( V_L^G(q) < 0 \) for all \( q \leq \frac{1 - \eta}{1 - p} \) and \( V_L^G(q) > 0 \) for all \( q \geq \bar{q} \). Let \( R > \bar{R}(\eta) \). Since \( p(\theta_H | HV) < R \) and \( \bar{\theta}_{LV}^G < \frac{\theta_L + \theta_H}{2} < \bar{\theta}_{HV}^G \) for \( q = \frac{1 - \eta}{1 - p} \), as \( \sigma \to +\infty \), \( V_L^G(\frac{1 - \eta}{1 - p}) \to -p \).

Therefore, there exists \( \sigma^*(\eta, R) \) such that, for all \( \sigma > \sigma^*(R, \eta) \), \( V_L^G(\frac{1 - \eta}{1 - p}) < 0 \). For any
σ > σ^∗(R, η), since as R → p, \(\bar{q} \rightarrow 1\), there exists \(R_2(\eta)\) such that for all \(R > R_2(\eta)\), \(V_\ell^G(\bar{q}) > 0\). We can infer that at \(q = \bar{q}\), \(p(\theta_H | HV)(\bar{q}) > R\). Otherwise, for \(σ > σ^∗(R, η)\), \(V_\ell^L < 0\), contradiction. Since \(V_\ell^a\) strictly increases with \(q\), \(V_\ell^G(q) > 0\) for all \(q ≥ \bar{q}\).

Therefore, there exists an unique \(\bar{q}\) such that \(V_\ell^G(\bar{q}) = 0\) with \(\frac{1-σ}{1-p} < \bar{q} < \bar{q}^∗\). For \(q \in (\frac{1-σ}{1-p}, \bar{q})\), \(V_\ell^G < V_\ell^B\). There exists \(q ≥ \frac{1-σ}{1-p}\) such that \(V_\ell^B ≥ 0\) and \(V_H^L ≥ 0\). The latter follows from \(V_H^a > V_\ell^a\). So, \((1, 0, 1, 1)\) is equilibrium. Since \(p_{nd} > 0\), it is consistent with the data.

A3. Proof of Proposition 4: There exists \(σ^∗ > 0\) such that for \(σ < σ^∗\), \(V_H^i > 0\) for any \(i \in \{G, B\}\), where \(σ^∗ = \inf\{σ | \exists(i, σ, R, β, η) s.t. V_H^i ≤ 0\}\). Since \(V_H^i → ε\) as \(σ → 0\) for any \(i \in \{G, B\}\), \(σ^∗ > 0\) is well-defined. Therefore, if \(σ < σ^∗\), all good papers are submitted around the deadline. If authors with good papers invest in reputation, we must have \(σ > σ^∗\).

Suppose \(σ > σ^∗\). We will first show that there exists \(q_T\) such that, for \(q > q_T\), no equilibrium is consistent with the motivating facts. Define \(q_T = \sup\{q | \exists(i, σ, R, β, η) s.t. V_H^i ≤ 0\}\).

\(q_T < 1\) because as \(q → 1\), \(V_H^i → ε + (1-ε)G(\theta_{LV}) - G(\theta_{HV}) > ε(1-G(\frac{θ_H+θ_L}{2})) > min_{σ>σ^∗}\epsilon\{1-Φ(\frac{θ_L-θ_H}{2\sigma})\} = \epsilon(1-Φ(\frac{θ_L-θ_H}{2\sigma^∗})) > 0\) for any \(i \in \{H, L\}\). For any \(q > q_T\), no equilibrium is consistent with the motivating facts. Define \(q_{T_1} = \sup\{q | \exists(i, σ, R, β, η) s.t. V_H^i ≤ 0\}\).

We have \(q_{T_1} < 1\) because, for any strategy profile with \(x_H^G < x_H^B\), as \(q → 1\), \(V_H^i > min_{σ>σ^∗}\{ε + (1-ε)F(\theta_{LV}) - F(\theta_{HV})\} > min_{σ>σ^∗}\{ε(1-F(\frac{θ_H+θ_L}{2}))\} = min_{σ>σ^∗}\epsilon(1-Φ(\frac{θ_L-θ_H}{2\sigma^∗})) > \frac{1}{2}ε\) for any \(i \in \{H, L\}\). Therefore, for any \(q > q_{T_1}\), only equilibria with \(x_H^G = x_H^B = 1\) can be consistent with the motivating facts. If \(q_T > q_{T_1}\), for any reputation \(q ∈ (q_{T_1}, q_T)\), only equilibria with all good papers submitted around the deadline can be consistent with the motivating facts.
Now we show that there exists $q_B$ such that for all $q < q_B$, no equilibrium satisfying $x^B_L = 1$ or $x^G_H \in (0, 1)$ can be consistent with the motivating facts. By contradiction, suppose there does not exist such a lower bound. Let $q'$ satisfy $M(\overline{\theta}^2_1) = M(\overline{\theta}^2_2)$, where $\overline{\theta}^2_1$ and $\overline{\theta}^2_2$ are readers’ strategies when they believe the paper is from good authors and when they believe the paper’s author is a good author with probability $pq' + (1-p)(1-q')$. For any $q \leq q'$, given paper quality, all good authors have a strictly stronger incentive to submit away from the deadline than bad authors. Therefore, if all bad authors with bad papers submit around the deadline, for Fact Three to be satisfied, good authors with good papers must submit around the deadline with positive probability, which implies all bad authors with good papers submit around the deadline. The equilibrium is consistent with Fact Two only if the reputation is sufficiently high. Contradiction. Now suppose good authors with good papers randomizes and bad authors with bad papers invest in reputation with positive probability. Since $q \leq q'$, all good authors with bad papers invest in reputation and all bad authors with good papers submit around the deadline. As the reputation $q$ decreases, the reputation gain decreases for authors with bad papers and increases for authors with good papers. For bad authors with bad papers to randomize, the sacrifice for the investment in reputation must be sufficiently small as the reputation becomes sufficiently low, which only happens if papers’ noisy signals are very noisy. However, if papers’ signals are very noisy, to authors with good papers, the sacrifice for the investment in reputation is also very small, which implies authors with good papers have a strict incentive to submit away from the deadline. Again, we reach a contradiction.
A4. Proof of Proposition 5: First, let’s define the social welfare function. The net social payoff (the social payoff minus $\theta^*$, the payoff from being not cited) at any given information set is

$$w(b) = b(\theta_H - \theta^*)(1 - F(\theta^*)) + (1 - b)(\theta_L - \theta^*)(1 - G(\theta^*))$$

where $b$ is the proportion of good papers at certain information set. Let $p(\theta_H)$ be the proportion of good paper at the beginning of $t = 1$ and $p(\theta_H | s)$ be the proportion of good papers at position $s$. Then we have $p(\theta_H) = \eta p(\theta_H | HV) + (1 - \eta)p(\theta_H | LV)$. So, the first period net social payoff is

$$w_1 = \eta w(p(\theta_H | HV)) + (1 - \eta)(1 - \epsilon)w(p(\theta_H | LV))$$

Conditional on paper’s quality $\theta_i$ at $t = 1$, the second period net social payoff is

$$w^i_2 = [p(HV | \theta_i)w(p(\theta_H | \theta_i, HV)) + p(LV | \theta_i)w(p(\theta_H | \theta_i, LV))](1 - (1 - \eta)\epsilon)$$

where $i = \{H, L\}$, $p(HV | \theta_i)$ is the probability of $\theta_i$ papers at $t = 1$ having position $HV$. $p(\theta_H | \theta_i, s)$ is the probability of a paper at $t = 2$ being good, given that its author submitted a paper at $t = 1$ with position $s$ and quality $\theta_i$. Denote $p(\theta_H | \theta_i)$ as the probability of a second period paper being good, given that its author having paper $\theta_i$ at $t = 1$. That is,

$$p(\theta_H | \theta_H) = \frac{pq}{pq + (1 - p)(1 - q)(2p - 1) + 1 - p}$$

$$p(\theta_H | \theta_L) = \frac{(1 - p)q}{1 - pq - (1 - p)(1 - q)(2p - 1) + 1 - p}$$

We have

$$p(\theta_H | \theta_i) = p(HV | \theta_i)p(\theta_H | \theta_i, HV) + p(LV | \theta_i)p(\theta_H | \theta_i, LV)$$
Hence, any strategy profile’s net social payoff is

\[ W = (1 - u)w_1 + u[p(\theta_H)w_H^2 + p(\theta_L)w_L^2] \]

where \( u \in [0, 1] \) is the weight on the second period social payoff.

**Second**, \( w \) is strictly increasing, non-negative and strictly convex.

Trivially, \( w(0) = 0 \).

\[ w'_b = [(\theta_H - \theta^*)(1 - F(\bar{\theta})) - (\theta_L - \theta^*)(1 - G(\bar{\theta}))] - [b(\theta_H - \theta^*)f(\bar{\theta}) + (1 - b)(\theta_L - \theta^*)g(\bar{\theta})] \frac{d\bar{\theta}}{db} \]

By \( E(\theta \mid b, \bar{\theta}) = \theta^* \), \( b(\theta_H - \theta^*)f(\bar{\theta}) + (1 - b)(\theta_L - \theta^*)g(\bar{\theta}) = 0 \). Hence, \( w'_b = (\theta_H - \theta^*)(1 - F(\bar{\theta})) - (\theta_L - \theta^*)(1 - G(\bar{\theta})) > 0 \). Hence, \( w \) is strictly increasing. We have \( w(b) > 0 \) for all \( b > 0 \). \( w \) is also strictly convex because \( w''_b = -[(\theta_H - \theta^*)f(\bar{\theta}) + (\theta^* - \theta_L)g(\bar{\theta})] \frac{d\bar{\theta}}{db} > 0 \).

**Finally**, Denote \( x = p(\theta_H), x_1 = p(\theta_H \mid HV) \) and \( y_1 = p(\theta_H \mid LV) \). We have \( \eta x_1 + (1 - \eta)y_1 = x \).

\[ w_1(1, 1, 1, 1) = (\eta + (1 - \eta)(1 - \epsilon))w(x) \]

For any equilibrium consistent with the data, \( x_1 > y_1 \). Since \( w \) is strictly increasing, we have \( w(x_1) > w(y_1) \). \( w_1 = \eta w(x_1) + (1 - \eta)(1 - \epsilon)w(y_1) \). By convexity and positiveness of \( w \),

\[ \eta w(x_1) + (1 - \eta)w(y_1) > w(x) \iff \]

\[ [\eta + (1 - \epsilon)(1 - \eta)]\eta w(x_1) + (1 - \eta)w(y_1)] > [\eta + (1 - \epsilon)(1 - \eta)]w(x) \]

\[ w_1 - [\eta + (1 - \epsilon)(1 - \eta)]\eta w(x_1) + (1 - \eta)w(y_1)] = \epsilon \eta(1 - \eta)[w(x_1) - w(y_1)](> 0) \]

Hence, \( w_1 > w_1(1, 1, 1, 1) \). By convexity, in the second period, the social payoff of the uninformative equilibrium is always lower. So, its total social payoff is also lower.
Appendix B: Supplemental Theoretical Results

B1. Welfare Analysis: Positive Weight on Authors

Suppose the social payoff has weight $x \in [0, 1)$ on readers and $1-x$ on authors.

$$w(b) = (1-x)[b(\frac{x}{1-x} (\theta_H - \theta^*) + 1)(1-F(\bar{\theta}(b))) + (1-b)(\frac{x}{1-x} (\theta_L - \theta^*) + 1)(1-G(\bar{\theta}(b)))]$$

Define $\theta'_i = \frac{x}{1-x} \theta_i$ and $\theta^{*'} = \frac{x}{1-x} \theta^*$, $\theta = \frac{\theta^{*'} - \theta'_L}{\theta^*_H - \theta'_L}$.

$$w(b) = (1-x)[b((\theta'_H - \theta^{*'}) + 1)(1-F(\bar{\theta}(b))) + (1-b)((\theta'_L - \theta^{*'}) + 1)(1-G(\bar{\theta}(b)))]$$

**Lemma 1** $w$ satisfies the following properties:

1. $w(0) = 0$. $w(b) > 0$ for all $b \in [0, 1]$.
2. $w'(b) > 0$.
3. $w$ is strictly convex for $b < \theta$ for all $\sigma > 0$, and strictly concave for $b > \theta$ when $\sigma$ is sufficiently big.
4. $w''_b < 0$ if $b \leq \theta$ and $w''_b > 0$ if $b > \theta$ when $\sigma$ is sufficiently big.

**Proof:** (1) trivial except for $w > 0$ for all $b > 0$, which follows from (2).

(2)

$$w'_b = (1-x)[(\theta'_H + 1-\theta^{*'})(1-F(\bar{\theta})) - (\theta'_L + 1-\theta^{*'})(1-G(\bar{\theta}))] - [bf(\bar{\theta}) + (1-b)g(\bar{\theta})]$$
By \( \theta'_H > \theta'_L \) and \( F < G \), the first part is positive. By \( \frac{\theta'}{\theta} < 0 \), we have \( w'_b > 0 \) for all \( b \in [0, 1] \). Since \( w(0) = 0 \) and \( w'_b > 0 \), \( w > 0 \) for all \( b > 0 \).

\[
\frac{1}{1-x} w''(b) = \left[-(\theta'_H + 2 - \theta^*) f(\theta) + (\theta'_L + 2 - \theta^*) g(\theta)\right] \frac{d\theta}{db} - \\
\left[b f'(\theta) + (1 - b) g'(\theta)\right] \left(\frac{d\theta}{db}\right)^2 - \left[b f(\theta) + (1 - b) g(\theta)\right] \frac{d^2\theta}{db^2}
\]

\[
= \frac{d\theta}{db} \left[-(\theta'_H - \theta^*) f(\theta) + (\theta'_L - \theta^*) g(\theta) - \frac{\theta - \theta_L}{1 - b \theta_H - \theta_L} f(\theta) + \frac{\theta_H - \theta}{b \theta_H - \theta_L} g(\theta)\right]
\]

The last two steps come from \((\frac{\theta}{\theta})^\prime_b = -\sigma b(1-b)^{1-\sigma} \), \( f'(\theta) = f(\theta)(-\frac{\theta - \theta_L}{\sigma^2}), g'(\theta) = g(\theta)(-\frac{\theta - \theta_L}{\sigma^2}) \) and \(-\frac{\sigma}{\theta_H - \theta_L} \frac{1}{(1-b)} \). If \( b \leq \theta \), as \( \sigma \) being sufficiently noisy, \( \theta > \theta_H \) and hence \( w'' > 0 \). If \( b > \theta \), as \( \sigma \) being sufficiently big, \( w \) is strictly concave.

\[
\frac{dF(\theta)}{d\sigma} = f(\theta) \frac{d\theta}{d\sigma} + \int_{-\infty}^{\theta} \frac{df(x)}{d\sigma} dx
\]

\[
= f(\theta) \frac{d\theta}{d\sigma} + \frac{1}{\sigma} \left(\int_{-\infty}^{\theta} f(x) \frac{(x - \theta_H)^2}{\sigma^2} dx - F(\theta)\right) = f(\theta) \left(\frac{d\theta}{d\sigma} - \frac{\theta - \theta_H}{\sigma}\right)
\]

Similarly, we have

\[
\frac{dG(\theta)}{d\sigma} = g(\theta) \left(\frac{d\theta}{d\sigma} - \frac{\theta - \theta_L}{\sigma}\right)
\]

So, we have

\[
\frac{1}{1-x} w'_\sigma = -b(\theta'_H + 1 - \theta^*) [f(\theta) \left(\frac{d\theta}{d\sigma} - \frac{\theta - \theta_H}{\sigma}\right)] - (1 - b)(\theta'_L + 1 - \theta^*) [g(\theta) \left(\frac{d\theta}{d\sigma} - \frac{\theta - \theta_L}{\sigma}\right)]
\]

\[
= -[b f(\theta) + (1 - b) g(\theta)] \frac{d\theta}{d\sigma} + [b(\theta'_H + 1 - \theta^*) f(\theta) \frac{\theta - \theta_H}{\sigma} + (1 - b)(\theta'_L + 1 - \theta^*) g(\theta) \frac{\theta - \theta_L}{\sigma}]
\]

If \( b < \theta \), as \( \sigma \) being sufficiently big, \( w'_\sigma < 0 \); if \( b > \theta \), as \( \sigma \) being sufficiently big, \( w'_\sigma > 0 \). \( \square \)

**Proposition 1.** When the signals are sufficiently noisy and the reputation \( q \) is sufficiently low, any equilibrium consistent with the motivating facts generates higher social
payoff than \((1, 1, 1, 1)\).

This result follows directly from Lemma 1 and Proposition 5 from the main text. When the reputation \(q\) is sufficiently low, the social payoff is higher than that of \((1, 1, 1, 1)\) because the welfare function \(w\) is strictly convex for sufficiently low beliefs about the paper quality \((b)\). The result may still hold for higher reputations when the quality of papers at HV positions is higher than \(\theta\) in the first period. The reason is that the higher welfare \(w\) from papers at HV positions dominates the concavity of \(w\) for papers at HV positions. Since in the first period only a small fraction of papers are good papers, in the second period, the social payoff for authors who had bad papers in the first period dominates and hence the second period social payoff is also higher than that in \((1, 1, 1, 1)\).

**B2. Authors with Incomplete Information**

Suppose authors’ information about their own quality follows

\[
\hat{\theta}_a = \theta_i + \sigma_a \xi_a
\]

where \(i \in \{H, L\}\) and \(\xi_a \sim N(0, 1)\).

**Proposition 2.** In the second period, at equilibrium, all papers are submitted around the deadline.
**Proof:** Denote the expected payoffs of type $a$ author with signal $\hat{\theta}$ taking D and ND as $V(D \mid \hat{\theta},a)$ and $V(ND \mid \hat{\theta},a)$ respectively. Then, for any equilibrium,

$$V(D \mid \hat{\theta},a) - V(ND \mid \hat{\theta},a) = (1 - p_{nd} - p_d)[p(\theta_H \mid \hat{\theta},a)\Delta' + g^*(\bar{\theta}_{HV}) - h(\bar{\theta}_{LV})]$$

where $\Delta' = f^*(\bar{\theta}_{HV}) - l(\bar{\theta}_{LV}) - [g^*(\bar{\theta}_{HV}) - h(\bar{\theta}_{LV})]$.

We consider three cases:

1. If $\Delta' > 0$, since $p(\theta_H \mid \hat{\theta},a)$ increases with $\hat{\theta}$, for the same type of author, the ones with higher signals are more likely to submit around the deadline. So, authors use cutoff strategy, submitting around the deadline only if they have good signals. Denote the cutoffs for good authors and bad authors to be $\bar{\theta}_G$ and $\bar{\theta}_B$ respectively.

   If $\bar{\theta}_B = -\infty(+\infty)$, we have $\bar{\theta}_G = -\infty(+\infty)$. Suppose there exists finite cutoff $\bar{\theta}_B$. We have

   $$p(\theta_H \mid \hat{\theta},G) = \frac{pf(\hat{\theta})}{pf(\hat{\theta}) + (1 - p)g(\hat{\theta})}$$

   $$p(\theta_H \mid \hat{\theta},B) = \frac{(1 - p)f(\hat{\theta})}{(1 - p)f(\hat{\theta}) + pg(\hat{\theta})}$$

   By $p > 1/2$, $p(\theta_H \mid \hat{\theta},G) > p(\theta_H \mid \hat{\theta},B)$. So, $\bar{\theta}_G < \bar{\theta}_B$. Hence, if there are too many papers submitted away from the deadline ($p_{nd} > 0$), readers believe papers with position LV have quality

   $$p(\theta_H \mid LV) = \frac{qpF(\bar{\theta}_G) + (1 - q)(1 - p)F(\bar{\theta}_B)}{q(pF(\bar{\theta}_G) + (1 - p)G(\bar{\theta}_G)) + (1 - q)((1 - p)F(\bar{\theta}_B) + pG(\bar{\theta}_B))}$$
By $F < G$ and $\bar{\theta}_B > \bar{\theta}_G$, we have $p(\theta_H | LV) < qp + (1 - q)(1 - p)$. So, $\bar{\theta}_{LV} > \bar{\theta}_{HV}$. Hence, $V(D | \hat{\theta}, a) > V(ND | \hat{\theta}, a)$ for all $\hat{\theta}$, all $\epsilon$ and all $a$. All authors submit around the deadline for any signals. Contradiction. The logic also holds when $p_d > 0$.

(2) if $\Delta' = 0$, there are three cases: all D submission, all ND submission, and some submitting around the deadline, others submitting away from the deadline. If all authors submit away from the deadline, everyone has incentive to deviate. If it is the last case, each author is indifferent between D submission and ND submission, which contradicts the crossing-order condition.

(3) if $\Delta' < 0$, for the same type of authors, the ones with higher signals are more likely to submit late. If there exists finite cutoff $\bar{\theta}_B$, we have $f^*(\bar{\theta}_{HV}) < l(\bar{\theta}_{LV})$. By the crossing-order condition, $g^*(\bar{\theta}_{HV}) < h(\bar{\theta}_{LV})$. But if so, we have $\bar{\theta}_B = -\infty$. Contradiction.

In all, all papers are submitted around the deadline. □

So, we can write authors’ strategy profile as $(s_G, s_B)$. If authors use cutoff strategies, we can write it as $(\bar{\theta}_G, \bar{\theta}_B)$. There exists a functional $C : (s_G, s_B) \rightarrow (x^G_H, x^G_L, x^B_H, x^B_L)$ with

$$x^G_H = \int_{-\infty}^{+\infty} s_G(\sigma_a)f_a(\sigma_a) d\sigma_a, x^G_L = \int_{-\infty}^{+\infty} s_G(\sigma_a)g_a(\sigma_a) d\sigma_a$$

$$x^B_H = \int_{-\infty}^{+\infty} s_B(\sigma_a)f_a(\sigma_a) d\sigma_a, x^B_L = \int_{-\infty}^{+\infty} s_B(\sigma_a)g_a(\sigma_a) d\sigma_a$$

We call the functional value as the reduced strategy profile. Then $(1, 0, 1, 1)$ is robust to the perturbation of authors’ information, which can be formally stated as follows.
Proposition 3. Suppose \((1, 0, 1, 1)\) is equilibrium for some \(q\). There exists sufficiently small \(\sigma_a^0(q)\) and \(\eta_0 < \frac{\theta_H - \theta_L}{2}\) such that \(\forall \sigma_a < \sigma_a^0(q)\), for some \(\theta_G \in \left[\frac{\theta_H + \theta_L}{2} - \eta_0, \frac{\theta_H + \theta_L}{2} + \eta_0\right]\), \((\theta_G, -\infty)\) is equilibrium.

Proof: For sufficiently small \(\epsilon_0\) and \(\eta_0 < \epsilon_0 < \frac{\theta_H - \theta_L}{2}\), let \(\theta_G \in \left[\frac{\theta_H + \theta_L}{2} - \eta_0, \frac{\theta_H + \theta_L}{2} + \eta_0\right]\). There exists \(A\) and \(\sigma'_a > 0\) such that for all \(\sigma_a < \sigma'_a\), \(V^G_L < 0, V^G_H > 0, |\ln\left(-\frac{1-p}{p} V^G_G(\theta^0_G, \sigma_a)\right)| \leq A, V^H_H > 0\) and \(V^B_L > 0\). By the equilibrium condition,

\[
\bar{\theta}_G^l = \frac{\sigma_a^2}{\theta_H - \theta_L} \ln(\frac{1-p}{p} \frac{V^G_L(\theta^0_G, \sigma_a)}{V^G_H(\theta^0_G, \sigma_a)}) + \frac{\theta_H + \theta_L}{2}
\]

\(\bar{\theta}_G^l\) is a continuous function of \(\theta_G^0\). For all \(\sigma_a < \frac{\eta_0}{\theta_H - \theta_L} (= \sigma_a^0), \bar{\theta}_G^l \in \left[\frac{\theta_H + \theta_L}{2} - \eta_0, \frac{\theta_H + \theta_L}{2} + \eta_0\right]\).

By Brouwer fixed point theorem, there exists \(\bar{\theta}_G^l = \theta_G^0\). Hence, \((\theta_G^0, -\infty)\) is equilibrium. \(\square\)

Appendix C. Algorithm on Identifying Submitters and their publication history

Step One: We start from all (23641) papers submitted between Jan 2000 and Mar 2007 to arXiv.org and use the information on the submitter, email address and coauthors to get the initial set of names for submitters. Each submitter might be under multiple names. The procedure is as following: (1) removing papers submitted by none of the authors by matching submitters, submitters’ email addresses and authors (329 papers). (2) aggregating the names of submitters with the same email address. Except for a few cases (three
email addresses), authors submitted with the same email address are the same person. (3) aggregating submitters across different email addresses. Denote $A$ as the set of submitters’ names from (2) and $A_1$ as the subset with only last name (one word) and first name initial. Matching $A_1$ to $A/A_1$ and resulting in $A_{11}$ with $A_1 \subset A_{11} \subset A$. Denote $A_2$ as the subset of $A/A_{11}$ with only last name (one word) and full first name (one word). Matching $A_2$ to $A/(A_{11} \cup A_2)$. This process continues until we reach $A_n$ which cannot easily be matched to $A/(A_{11} \cup A_{21} \ldots \cup A_n)$. Manually matching within $A/(A_{11} \cup A_{21} \ldots \cup A_{(n-1)1})$. This procedure produces 5023 names and a list of papers submitted under each of these names. In total, there are 23312 papers.

Step Two: We identify names for submitters who ever submitted between 2002 and 2004. There are 23019 out of 23312 papers submitted between 2000 and Mar 2007 after the following restrictions. First of all, we remove all submissions during holidays as defined in arXiv.org What’s New because these submissions follow incentives outside of the interest of this paper. Second, the incomplete lists at the beginning of 2000 and the end of Mar 2007 are deleted. Third, we delete the lists around the change of the deadlines. Finally, 11 papers mistakenly announced on a later list are deleted. Assume for now a submission is from US or Canada if one of the affiliations of the submitter is located in US or Canada. Among these 23019 papers, there are 2569 papers submitted by 888 authors from US or Canada between 2002 and 2004. Among these 888 names, 81 are of Chinese or Korean descent. Therefore, there are in total 807 names of non-Chinese or non-Korean descent.

Step Three: We search the publication lists for the 807 names from Step Two in SPIRES and INSPIRE. The searching details are in Table 15. We divide names based on whether the name has a hyphen and the number of words in the name. Moreover, there are names
which are similar and we do not have enough information at this stage for whether they
belong to the same person. We then assume they are from the same author and define their
combination as an ambiguous name. Among the non-ambiguous names, there are 753 un-
hyphened and 22 hyphened. Among the un-hyphened names, most of them have last name
with only one word (742 names). For these names, we search through last name and first
name or first name initial (when we do not have information on the full first name). 284
names within this group have both first name and middle name (but not more). We also
search through the combination of the last name and the full middle name (if available) in
consideration of the fact that some authors may use the middle name instead of the first
name. The 11 un-hyphened names have last name with multiple words. For 10 of them, only
the last part of the last name matter and hence we use both the full last name and the last
part of the last name in the search. For the only one left, we use both the first part and the
last part of the last name (the last name has only two words). For all these 11 names, we
use only the first name or first name initial if only initials are available. For the hyphened
names with only either last name or first name hyphened but not both, we use both parts
of the last name and also the full last name. For the first name, we use only the first part.
There is only one name with both last name and first name hyphened. We use only the first
part or the second part of the last name (first name). Finally, for the 32 ambiguous names,
we search each name in the ambiguous name separately by the same criterion as above and
combine the searching results together.

_Step Four:_ We clean the searched publication lists from Step Three according to the
following algorithm. First, we remove all papers with more than 8 authors and only keep
papers published before September 2011. Second, we remove all papers with none of the
authors matching the searching name. Third, we identify publication records from matching affiliations and coauthors by the following procedure. (i) we order papers based on publication times (by year and month, according to normal order of time). (ii) We put the earliest papers with this submitter having no affiliation at the end of the list until this condition is not satisfied. (iii) We match papers through coauthors and affiliations. We start from the first paper in the modified list. If in the next paper, the author matching the submitter’s name has the same affiliation, then this paper belongs to this submitter. Otherwise, checking the coauthors. If some authors of the next paper match the submitter’s previous coauthors, then this paper belongs to the submitter. This process continues until no more paper is matched. We put the the matched papers at the beginning of the list and apply the same process repeatedly until no more paper gets matched. For the rest of papers, we use the same algorithm repeatedly. Finally, a partition \((B)\) of the publication list is generated. (iii) We remove sets from the partition based on the following steps. First, identifying sets \((B_1)\) including papers submitted between 2002 and 2004. If there are sets \(B_{11} \subset B_1\) overlap each other in terms of publication date, then we manually check whether they belong to the same submitter. If they come from different submitters, we create a new submitter for each set in \(B_{11}\). Each of these submitters’ other publications are taken as \(B/B_{11}\). Without loss of generality, for now, we assume the sets in \(B_1\) do not overlap with each other. Second, we remove sets \(B_2 \subset B/B_1\) which significantly overlap any set in \(B_1\) in terms of publication date. Third, we choose the set \(P_r\) in \(B_1\) with the most recent publications in \(B_1\) and the set \(P_{or}\) in \(B/(B_1 \cup B_2)\) which are the oldest publications after \(P_r\). If there is big time gap (> 3 years) between the last paper in \(P_r\) and the first paper in \(P_{or}\), then after manually checking, we throw away all papers published after \(P_r\). Otherwise, adding \(P_{or}\) to \(B_1\). Fourth, consid-
ering the set $P_o$ in $B_1$ with the oldest publications in $B_1$ and the set $P_{ro}$ which are the most recent publications before $P_o$. If there is a big time gap ($> 3$ years) between the first paper in $P_o$ and the last paper in $P_{ro}$, after manually checking, we throw away all papers published before $P_o$. Otherwise, adding $P_{ro}$ to $B_1$. Repeating step second, third and fourth under (iii) until all sets in $B/(B_1 \cup B_2)$ are exhausted. We get $B_3$. (iv) ordering all papers in $B_3$ according the normal order of time and manually checking the time gap between adjacent papers. We get the publication history including submission history for each submitter.
### Table 5.1: Searching Details On The Initial Publication History.

We assume here an submitter is from US or Canada if one of the affiliations under the name is located in US or Canada. With names of Chinese or Korean descent excluded, there are in total 807 names to put in the search engine.

<table>
<thead>
<tr>
<th>Category of Names</th>
<th>Quantity</th>
<th>Searching Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Names without hyphen</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1) Last name (one word)</td>
<td>446</td>
<td>A (440)</td>
</tr>
<tr>
<td>First name</td>
<td></td>
<td>EA (full, 2; part, 4)</td>
</tr>
<tr>
<td>(2) Last name (two words)</td>
<td>6</td>
<td>A (full ln+fn, last ln+fn) (5)</td>
</tr>
<tr>
<td>First name</td>
<td></td>
<td>A (first ln+fn, last ln+fn) (1)</td>
</tr>
<tr>
<td>(3) Last name (one word)</td>
<td>284</td>
<td>A (ln+fn, ln+mn) (221)</td>
</tr>
<tr>
<td>First name, Middle name</td>
<td></td>
<td>EA (4,8) (20)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A (ln+fn, ln+mn) (39 full, 4 part)</td>
</tr>
<tr>
<td>(4) Last name (multiple words)</td>
<td>5</td>
<td>A (full ln+fn, last ln+fn)</td>
</tr>
<tr>
<td>First Name, Middle name</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(5) Last name (one word)</td>
<td>12</td>
<td>A (ln+fn)</td>
</tr>
<tr>
<td>First Name, Middle Name, Third Name</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
B. Names with hyphen

<table>
<thead>
<tr>
<th>Last name (one word with hyphen)</th>
<th>18</th>
<th>A (full ln+fn)</th>
</tr>
</thead>
<tbody>
<tr>
<td>First name</td>
<td></td>
<td>A (first ln+fn, second ln+fn)</td>
</tr>
<tr>
<td>(2) Last name (one word)</td>
<td>3</td>
<td>A (ln+first fn)</td>
</tr>
<tr>
<td>First name (with hyphen)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(3) Last name (one word with hyphen)</td>
<td>1</td>
<td>A (first ln+first fn, second fn)</td>
</tr>
<tr>
<td>First name with hyphen</td>
<td></td>
<td>A (second ln+first fn, second fn)</td>
</tr>
</tbody>
</table>

C. Ambiguous names

| multiple names (potentially the same author) | 32 | search each name separately according to A and B |

| Total | 807* |

(Table 5.1 continued.)
Table 5.2: Ranksum test for reputation effect. The null hypothesis is there is no difference for the coauthor-adjusted h-index between D and ND submissions. The top table shows reputation effect by aggregate data. The bottom table shows reputation effect by different reputations.

<table>
<thead>
<tr>
<th>Paper quality (Percentile)</th>
<th>0-10</th>
<th>10-25</th>
<th>25-50</th>
<th>50-75</th>
<th>75-90</th>
<th>90-100</th>
</tr>
</thead>
<tbody>
<tr>
<td>z-value</td>
<td>0.562</td>
<td>1.34</td>
<td>2.506</td>
<td>1.648</td>
<td>2.497</td>
<td>3.519</td>
</tr>
<tr>
<td>p-value</td>
<td>0.5741</td>
<td>0.1804</td>
<td>0.0122</td>
<td>0.0994</td>
<td>0.0125</td>
<td>0.0004</td>
</tr>
<tr>
<td>sample size (D)</td>
<td>52</td>
<td>88</td>
<td>415</td>
<td>213</td>
<td>152</td>
<td>90</td>
</tr>
<tr>
<td>sample size (ND)</td>
<td>377</td>
<td>534</td>
<td>1811</td>
<td>834</td>
<td>482</td>
<td>330</td>
</tr>
</tbody>
</table>
(Table 5.2 continued.)

<table>
<thead>
<tr>
<th>Percentile</th>
<th>0-10</th>
<th>10-25</th>
<th>25-50</th>
<th>50-75</th>
<th>75-90</th>
<th>90-100</th>
</tr>
</thead>
<tbody>
<tr>
<td>z-value</td>
<td>1.541</td>
<td>2.836</td>
<td>1.141</td>
<td>2.79</td>
<td>3.356</td>
<td>1.986</td>
</tr>
<tr>
<td>p-value</td>
<td>0.1233</td>
<td>0.0046</td>
<td>0.2537</td>
<td>0.0053</td>
<td>0.0008</td>
<td>0.047</td>
</tr>
<tr>
<td>sample size (D)</td>
<td>16</td>
<td>46</td>
<td>96</td>
<td>121</td>
<td>70</td>
<td>46</td>
</tr>
<tr>
<td>sample size (ND)</td>
<td>156</td>
<td>274</td>
<td>464</td>
<td>429</td>
<td>230</td>
<td>148</td>
</tr>
<tr>
<td>z-value</td>
<td>-0.415</td>
<td>1.391</td>
<td>2.237</td>
<td>1.463</td>
<td>0.526</td>
<td>0.983</td>
</tr>
<tr>
<td>p-value</td>
<td>0.6783</td>
<td>0.1642</td>
<td>0.0253</td>
<td>0.1434</td>
<td>0.5986</td>
<td>0.3258</td>
</tr>
<tr>
<td>sample size (D)</td>
<td>30</td>
<td>31</td>
<td>41</td>
<td>29</td>
<td>21</td>
<td>17</td>
</tr>
<tr>
<td>sample size (ND)</td>
<td>187</td>
<td>201</td>
<td>218</td>
<td>152</td>
<td>77</td>
<td>43</td>
</tr>
<tr>
<td>z-value</td>
<td>-0.594</td>
<td>0.929</td>
<td>1.508</td>
<td>1.096</td>
<td>1.779</td>
<td>1.609</td>
</tr>
<tr>
<td>p-value</td>
<td>0.5522</td>
<td>0.3528</td>
<td>0.1315</td>
<td>0.2729</td>
<td>0.0753</td>
<td>0.1076</td>
</tr>
<tr>
<td>sample size (D)</td>
<td>8</td>
<td>10</td>
<td>34</td>
<td>69</td>
<td>52</td>
<td>25</td>
</tr>
<tr>
<td>sample size (ND)</td>
<td>26</td>
<td>68</td>
<td>185</td>
<td>249</td>
<td>179</td>
<td>143</td>
</tr>
<tr>
<td>Paper Quality (Percentile)</td>
<td>0-10</td>
<td>10-25</td>
<td>25-50</td>
<td>50-75</td>
<td>75-90</td>
<td>90-100</td>
</tr>
<tr>
<td>---------------------------</td>
<td>------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>--------</td>
</tr>
<tr>
<td>(1) young authors</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>z-value</td>
<td>1.652</td>
<td>1.872</td>
<td>0.3</td>
<td>-0.366</td>
<td>2.151</td>
<td>1.326</td>
</tr>
<tr>
<td>p-value</td>
<td>0.0985</td>
<td>0.0612</td>
<td>0.7642</td>
<td>0.7145</td>
<td>0.0314</td>
<td>0.1709</td>
</tr>
<tr>
<td>sample size (D)</td>
<td>13</td>
<td>25</td>
<td>42</td>
<td>73</td>
<td>45</td>
<td>26</td>
</tr>
<tr>
<td>sample size (ND)</td>
<td>84</td>
<td>115</td>
<td>203</td>
<td>219</td>
<td>125</td>
<td>81</td>
</tr>
<tr>
<td>(2) youngest authors</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>z-value</td>
<td>-1.769</td>
<td>-0.613</td>
<td>-0.281</td>
<td>-1.772</td>
<td>-0.135</td>
<td>0.604</td>
</tr>
<tr>
<td>p-value</td>
<td>0.0769</td>
<td>0.5396</td>
<td>0.7784</td>
<td>0.0764</td>
<td>0.8922</td>
<td>0.5457</td>
</tr>
<tr>
<td>sample size (D)</td>
<td>13</td>
<td>21</td>
<td>60</td>
<td>71</td>
<td>53</td>
<td>27</td>
</tr>
<tr>
<td>sample size (ND)</td>
<td>52</td>
<td>116</td>
<td>190</td>
<td>219</td>
<td>132</td>
<td>72</td>
</tr>
</tbody>
</table>

Table 5.3: Ranksum test for reputation effect by age. The null hypothesis is there is no difference for the coauthor-adjusted h-index between D and ND submissions.
(Table 5.3 continued.)

(3) old authors

<table>
<thead>
<tr>
<th></th>
<th>z-value</th>
<th>p-value</th>
<th>sample size (D)</th>
<th>sample size (ND)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-0.184</td>
<td>0.8538</td>
<td>15</td>
<td>114</td>
</tr>
<tr>
<td></td>
<td>-0.156</td>
<td>0.8759</td>
<td>15</td>
<td>127</td>
</tr>
<tr>
<td></td>
<td>0.618</td>
<td>0.5364</td>
<td>36</td>
<td>223</td>
</tr>
<tr>
<td></td>
<td>0.154</td>
<td>0.8791</td>
<td>33</td>
<td>204</td>
</tr>
<tr>
<td></td>
<td>0.176</td>
<td>0.8602</td>
<td>33</td>
<td>125</td>
</tr>
<tr>
<td></td>
<td>2.171</td>
<td>0.0299</td>
<td>27</td>
<td>102</td>
</tr>
</tbody>
</table>

(4) oldest authors

<table>
<thead>
<tr>
<th></th>
<th>z-value</th>
<th>p-value</th>
<th>sample size (D)</th>
<th>sample size (ND)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-1.075</td>
<td>0.2825</td>
<td>11</td>
<td>127</td>
</tr>
<tr>
<td></td>
<td>0.653</td>
<td>0.5135</td>
<td>27</td>
<td>175</td>
</tr>
<tr>
<td></td>
<td>-1.487</td>
<td>0.137</td>
<td>29</td>
<td>250</td>
</tr>
<tr>
<td></td>
<td>-1.386</td>
<td>0.1659</td>
<td>34</td>
<td>185</td>
</tr>
<tr>
<td></td>
<td>-0.712</td>
<td>0.4762</td>
<td>21</td>
<td>96</td>
</tr>
<tr>
<td></td>
<td>0.313</td>
<td>0.7542</td>
<td>10</td>
<td>72</td>
</tr>
</tbody>
</table>
Figure 5.1: Publishing in theory-HEP, 2002-2004. The data are retrieved from inspirehep.net. The calculation is based on monthly average for 11 months each year. We assume papers with the subject denoted as theory-HEP in inspirehep.net are from the field of high energy physics-theory. For papers missing information on subjects, we calculate the number of papers from this field based on the ratio of papers with subjects in theory-HEP and other subjects.
Figure 5.2: The evolution of submissions in hep-th. Note: (1) **hep-th** is a sub-archive storing papers from high energy theory-theory; (2) red bar denotes the year of 2002. Data Source: arXiv.org
Figure 5.3: Measurement of author quality and reputation. The top figure shows the measurement of author quality. The scaling line for the top figure is author $hindex = 0.18 \times authorage - 8$. For the bottom figure, the scaling line is submission $hindex = 0.2 \times submissionage + 0.001$. 

137


