THE EFFECT OF FREE ACCESS ON THE DIFFUSION OF SCHOLARLY IDEAS

Completed Research Paper

Heekyung Hellen Kim
The National Bureau of Economic Research
Cambridge, MA 02138
hkim@nber.org

Abstract

This study investigates a relationship between free access to research articles and the diffusion of their ideas as measured by citation counts. By using a dataset from the Social Science Research Network (SSRN), an open repository of research articles, and employing a natural experiment that allows the effect of free access separate from other confounding factors, this study identifies the causal effect of free access on the citation counts. The natural experiment in this study is that a select group of published articles is posted on SSRN at a time chosen by their authors' affiliated organizations or SSRN, not by heir authors. Using a difference-in-difference method and comparing the citation profiles of the articles before and after the posting time on SSRN against a group of control articles with similar characteristics, I estimated the effect of the SSRN posting on citation counts to be 10-20% of total citations.

Keywords: Open access, knowledge management, value of free access, diffusion of knowledge

Introduction

Scientific knowledge progresses as science challenges and builds upon prior beliefs and findings. The ability of a society to make scientific progress, therefore, depends on how well the society can validate, organize, maintain, and offer researchers access to the prior knowledge (Rosenberg 1963; Rosenberg 1979; Heller and Eisenberg 1998; David 2005; Mokyr 2002). Many researchers have reported the roles of various institutions in effectively diffusing prior knowledge. For example, Furman and Stern (2011) showed that biological resource centers (BRCs), "living libraries" where biological materials are deposited for researchers, improved the diffusion of knowledge by both certifying and offering access to the deposited materials. Zuckerman and Merton (1971) defined scholarly journals as an institution: the scholarly journals certify the content of the research articles they publish while offering access to the articles for other researchers. These institutions validate the prior knowledge as well as offer access to the knowledge for other researchers and scholarly ideas are distributed through these institutions after being validated.

The advent of the Internet and digital publishing has, however, changed the way that scholarly ideas are disseminated. The validation and distribution of scholarly ideas are no longer provided by the same institution. Scholarly ideas are disseminated through working papers, blogs, Twitter, and many forms other than refereed journals, regardless of their contents being validated. As of 2010 Blogpulse tracked over 166 million blogs, and estimated that 70,000 blog posts were created per day. Research articles are often available for free access in open repositories in the Internet. For example, arXiv.org provides an online repository for researchers in physics, mathematics, computer science, and quantitative biology to post unpublished and published manuscripts for public viewing. As of May 2011, arXiv.org has posted approximately 6,000 new articles per month and total of 677,000 research articles since its inception in 1991. Social Science Research Network (SSRN), an open online repository of research articles in the field of social science, has archived 300,000 research articles as of 2010 since its inception in 1993 and provides free access to those unpublished and published articles. This paper examines how such an online non-refereed repository of research articles has changed the diffusion of scholarly ideas.

The open online repositories of research articles provide two functions that most of traditional journals do not, free access[1] and early exposure, at the expense of no referees or quality control. Anyone who has the access to the Internet can download the full text of articles from such an on-line repository for free while most of traditional journals allow access to their articles only for their (paid) subscribers. Research articles submitted to the online repository is exposed to readers immediately upon submission because there is no refereeing process. Because it often takes a few years from submission to publication (in a refereed journal), many authors post their unpublished manuscripts in an online repository or their own website for free readership before their articles are published in a journal or book (refereed or not). The free access and early exposure should, in theory, accelerate the diffusion of ideas. However, previous research findings (e.g., Lawrence 2001a; Gaule and Maystre 2011; Davis et al. 2008; Eysenbach 2006) have not reached a consensus on the issue. This study aims to reconcile the difference by identifying the effect of free access, separating it from confounding factors.

Ever since Lawrence (2001a) reported that research articles published in online free computer science proceedings received over an 300% of citations of non-free articles based on a cross-sectional study, many researchers have challenged the findings using a more rigorous identification strategy such as a randomized experiment and an instrumental variable method (e.g., Davis et al. 2008; Gaule and Maystre 2011). The difficulty to measure the effect of free access arises because two other factors are often confounded with free access. The first is selection bias. Often authors select their better articles to post in the Internet for free readership. An apparent difference in citations between free and non-free articles may be, therefore, due to the inherent quality difference. The second is early viewership. Most research articles are available as a working paper for readership well before they are published. Depending on a journal, it takes a few years for a research article to be published. The open repositories of research articles such as arXiv.org and SSRN (Social Science Research Network) allow authors to post any research article. Therefore, a difference in citations between free and non-free articles may arise from that the free

[1] Refereed journals increasingly offer free readership. The first type is not subscription-based; the fees are charged to authors for their submission of articles, not readers, and the journals are available for free readership. The journals of the Public Library of Science such as *PLoS Biology* and *PLoS Medicine* are the most notable examples of such type. The second type is that subscription-based journals offer authors to buy open access to their articles. These journals have, therefore, both open access articles to the readers and articles with pay-wall in the same issue.

articles had a longer time to receive a citation than the non-free articles (Moed 2007). A cross-sectional study to compare citations between free and non-free articles fails to separate the confounding factors and what appears to be the effect of free access can be the effect of other factors.

To identify the effect of free access, this study employs a natural experiment allowed by a unique feature of data from an online repository of research articles, Social Science Research Network (SSRN). SSRN was established in 1993 to enhance the dissemination of research ideas in social science; it covers 24 research disciplines, emphasizing law, economics, and finance. It had archived approximately 300, 000 research articles as of 2010, hosting over 300 working paper series. I exploit three aspects of SSRN to identify the effect of free online access. First, when an organization joins SSRN and starts a research paper series for the organization, a large number of research papers, often over 100 papers, is submitted to SSRN and posted at once, the timing of posting being exogenous to the authors or the quality of the papers. The articles may be chosen by the authors, but the timing of posting those articles is not decided by authors. I do not assume that there is no inherent difference in quality between the articles selected for posting on SSRN and the articles not selected. The assumption here is that the inherent difference in quality between the selected articles to post on SSRN and unselected articles up to the timing of posting should not change after the timing of posting. Secondly, some portion of research papers posted on SSRN has already been published in refereed journals prior to being posted on SSRN, having a citation profile over time before being posted on SSRN. The citation profile over time before posting on SSRN allows the estimation of quality difference between the articles selected for posting on SSRN and the unselected articles. The third relevant trait of SSRN arises from its having been established over 15 years, with many articles available on SSRN for more than 5 years; this relative longevity enables me to compile their citation trajectories over 5 years after they were freely available on SSRN. Because receiving citations from other published articles takes time, allowing enough time for the free access articles to accrue citations is important.

The natural experiment in this study is that some articles, which had been published in refereed journals at least 4 years earlier and thus had a citation profile over time prior to being posted on SSRN, were posted on SSRN at an exogenously chosen time ("treated articles") and other articles, which had been published in the same journal, volume, and issue as their counterpart treated articles, were never posted on SSRN ("control articles"). The treatment in this natural experiment is, therefore, the posting on SSRN. The identification of free access comes from the fact that the timing of posting the treated articles was not decided upon by their authors. By comparing the citation profiles of the treated articles before and after the posting time on SSRN with those of control articles with similar characteristics using a difference-in-difference method, I estimated the effect of free access on citations, not confounded with the quality of the treated articles. The results show that SSRN articles receive 60-80 % higher citations than their matched control articles even before being posted on SSRN, indicating that the articles are of higher quality. They receive an additional 10-20% of citations after being posted on SSRN, which is likely to driven by the free access that SSRN provides. The contribution of this paper is to identify a causal relationship between free access and the diffusion of scholarly ideas and quantify the impact of free access for the first time.

Open Access Debate

The theoretical proposition that I test is that scholarly ideas with free access should diffuse more widely and, as a result, their citation counts, a proxy for diffusion, should increase. However, the empirical findings on this relationship have been inconsistent, and researchers are still debating whether there is a causal relationship between free access and citations. The debate on the existence of an "open access advantage" for scholarly communications started when Lawrence (2001b) reported that freely-accessible online computer science proceedings received more than three times the average number of citations of papers as their counterpart paper journals. The open access advantage refers to the fact that free and unrestricted access to research papers gives them an advantage in receiving citations. Many researchers have reported that freely available papers in the Web receive more citations in a variety of disciplines (Norris 2008), such as computer science (Lawrence 2001b), astrophysics (Schwarz and Kennicutt 2004; Metcalfe 2005), physics (Harnad and Brody 2004), mathematics (Antelman 2004; Davis and Fromerth 2007), philosophy (Antelman 2004), political science (Antelman 2004), engineering (Antelman 2004), law (Donovan and Watson 2011), and multi-disciplinary sciences (Eysenbach 2006). Researchers, however, argue that what appeared to be an open access advantage may be attributable to either early viewing or self-selection or both. For example, Moed (2007) reports that arXiv.org accelerates citation due to the fact that it makes papers available *earlier* rather than by making them *freely* available. On the other hand, Kurtz et al. (Kurtz et al. 2005; Kurtz and Henneken 2007) report that authors tend to make more citable papers such as those published in journals with higher impact factors freely available, suggesting that self-selection, not unrestricted accessibility, causes the increased citation of open access papers. Conducting an experiment that randomly assigned certain articles for free access at publishers' websites, Davis et al. (2008) reported that there is no evidence of an open access advantage for citation counts in the 2 years subsequent to publication. In summary, the open access advantage ranges, researchers argue, from zero to 300% of citations of non-free research articles; early exposure and quality difference have been identified as the potential confounding factors for the overestimation of the effect of free access. Without an identification strategy capable of separating the confounding factors as well as allowing a reasonably long time for articles to receive citations after publication, an unbiased estimate on the value of free access cannot be made.

Identification Strategy

Previous studies have identified three factors may cause the increases in citations for research articles with open access: 1) free access; 2) early exposure; and 3) quality difference. These three factors are often confounded, causing a biased estimate of the effect of each factor. I employ an identification strategy to separate the effect of free access from other potential confounding factors with a longitudinal dataset. In this study, I focus on a setting where two requirements are met. The first is that an exogenous shock exists to make the research articles available for free access. In this setting, authors do not choose the time to post their articles for free access; instead, the organizations that the authors are affiliated with or the websites that host those articles choose the time to post articles for free access even if the authors choose which of their articles are posted. The second requirement is that these articles were already published for some time before being posted for free access. This requirement served two purposes: 1) the effect of early exposure is removed and, 2) more importantly, these articles have an observable citation trajectory over time before the posting, allowing the comparison of the citation trajectory before and after the posting. The research articles I chose were posted at a time decided upon not by their authors but by the authors' affiliated organizations or their hosting website and had been already published at least 4 years before being posted for free access, meeting the two requirements.

An inference on the effect of any event based on a comparison before and after the event should address a time trend that may concur with the effect of the event. A standard approach to address the time trend is to include time dummy variables in empirical equations. Merely including time dummy variables, however, is not enough to address the time trend when the dependent variable is citation counts. This inadequacy of merely including time dummy variables to account for the time trend exists because the citation profile over time is often specific for each research article, depending on when the articles were first published, when their citing articles were published, the interaction between the publication time of cited articles and citing articles, and the quality of the article. To separate the time and age effects, I used a difference-in-difference estimator by including a set of control articles with characteristics similar to their counterpart treated articles. I chose the control articles based on the following criteria: that they were published in the same journal, volume, and issue; and that they have their own observable citation trajectory over time, as their counterpart treated samples do. The difference-in-difference method that I used in this study was illustrated as in Figure 1. For example, a research article posted on SSRN in 2000 at an exogenous timing is selected for the study. Because the journal and volume published the article are known, the citation profiles for all the articles in the volume before and after the year of posting on SSRN are constructed. The selection bias is determined from the difference in the citations between the articles posted on SSRN, as indicated as SSRN paper, and the articles published in the same volume but not posted on SSRN, as indicated as control paper. The counterfactual citations, as denoted in a dotted line, that SSRN paper would have received after being posted on SSRN was constructed from the citations that the control papers received, on the assumption that their citation trend would be similar to that before the posting event. The difference in citations between the observed citations and the counterfactual citations is interpreted as SSRN effect or the effect of free access.

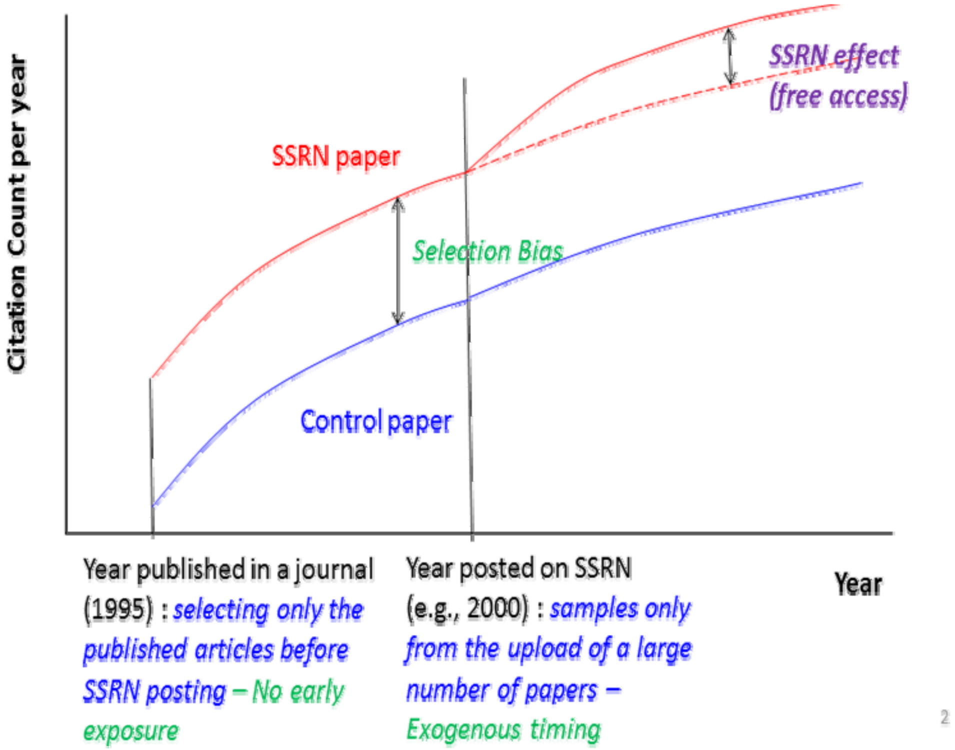

Figure 1. Identification Strategy

The limitation of the difference-in-difference estimator is that the counterfactual trajectory of the treated articles is accounted for by the control articles and the quality of match between the treated and control articles is critical. Therefore, in the next analysis for a tighter match between control and treated articles, I used the coarse exact matching method (CEM) to choose a subset of treated articles that can be matched to their control articles with respect to citation profiles over time and total citation counts up to the year when their matching treated articles were posted on SSRN. The citation profile of the control articles provide the counterfactual citation profiles over time that the treated articles would have without being posted for free access. It is, however, possible that the inherent difference between treated and control articles results in the different trajectory after the posting event. Because a traditional fixed effect estimator without using any control unit does not rely on the quality of match between treated and control units, it can be an alternative to the difference-in-difference estimator. Using a traditional fixed effect estimator and some common functional forms that other researchers have used for citation profiles, I also estimated the effect of free access.

For the statistical analysis, I used a conditional fixed effect negative binomial model and a conditional fixed effect Poisson model. While some studies have successfully used the conditional negative binomial model for panel estimation of overdispersed count data (e.g., Hausman et al. 1984; Furman and Stern 2011), it has been reported that the conditional fixed-effects negative binomial model is not a true fixed-effects model because it fails to control for all of the predictors that are fixed over time (Allison and Waterman 2002; Guimaraes 2008; Hilbe 2007). An alternative is to use a conditional fixed effect Poisson model but handle overdispersion of data by bootstrapping the sample without assuming any distribution of data or using a quasi-maximum-likelihood estimator to estimate a robust standard error (Hilbe 2007). As there are trade-offs in using one over the other model specification, I present the result using all of them in the first analysis. For the following analyses, I present the results from using only the conditional fixed effect Poisson model with robust standard error.

Data

Data Construction and Source

The data source I relied on for this study is the SSRN, complemented with the Web of Science. The SSRN was established in December of 1993 by Social Science Electronic Publishing Inc. to facilitate worldwide dissemination of social science research. Since then, the number of archived papers and delivered downloads has increased exponentially (Figures 2 and 3). For the year from May 2010 to May 2011, SSRN received 56,000 papers and delivered 8.6 million downloads. As of August 2010, SSRN had archived 298,243 research articles, of which 189,625 articles had full texts free of charge. Downloading and posting a research article on SSRN is free and open to anyone. However, a research organization is charged when SSRN hosts a research paper series for the organization. In addition, certain user services are charge-based: for example, an email alert or delivery service for research articles on certain topics or written by certain authors, suited to users' preferences, is provided to users at a charge. SSRN records posting and revision date of posted articles, tracks citations and number of downloads even before citing or cited articles are published in traditional scholarly journals, and identifies whether some papers in multiple versions are in fact the same paper, removing any erroneous counts of citation or posting of the same paper.

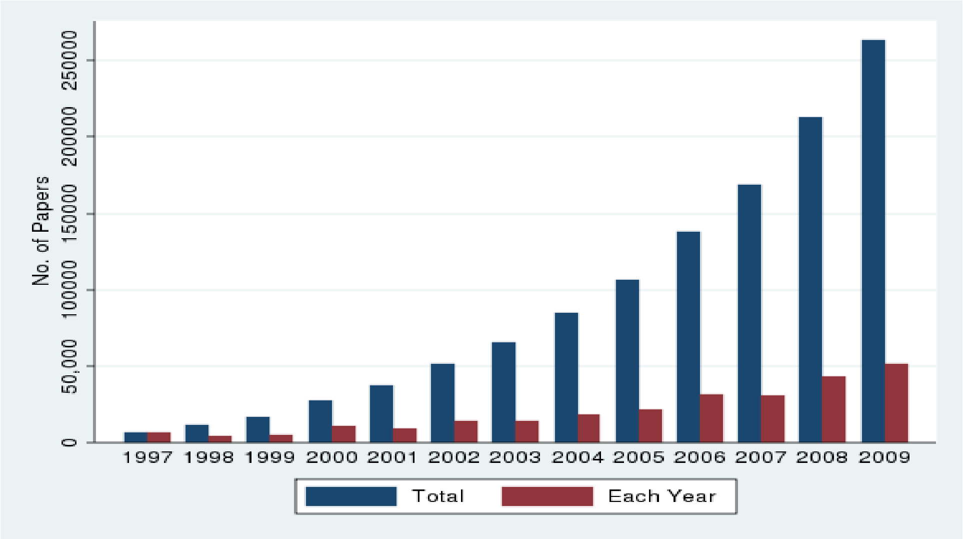

Figure 2. Number of papers posted on SSRN.

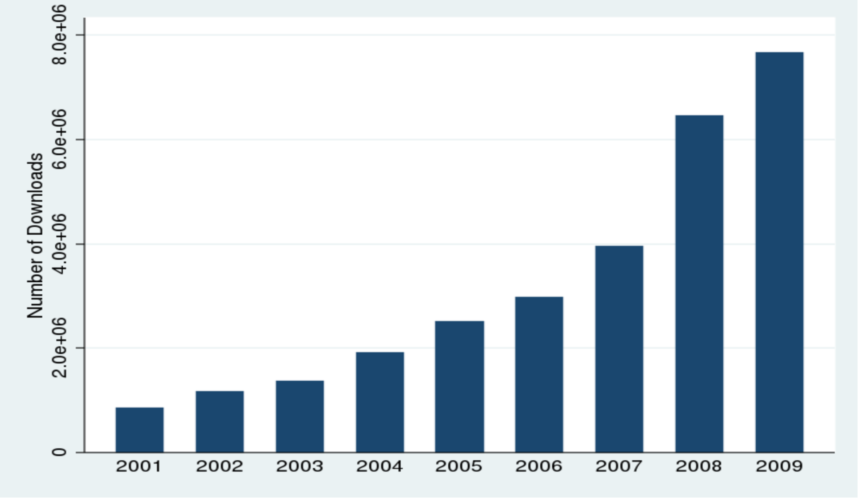

Figure 3. Number of downloads.

SSRN does not report whether their posted articles are published in refereed journals unless the authors or the organizations indicate it. In order to identify the publication status of SSRN articles, I matched the title and the names of authors with those in Web of Science and collected the information on the publication status. If they were identified as published, I collected data on the total citations, the publication source, the names of authors of citing papers, and the publication source of the citing papers. The matching method I used is hardly perfect: some research articles from SSRN may have been published with slightly different titles and erroneously identified as unpublished. The imperfect matching error can lead to two cases: 1) some published SSRN articles may be excluded from the study erroneously or 2) they may be categorized as control articles erroneously if they happened to be published in the same journal, volume, and issue as the other SSRN articles identified as published. In the first case, the exclusion of those published articles should not affect the estimate in one way or the other as their exclusion from the study is random. The second case can lead to a biased estimate on the effect of the free access. The estimate to which this error leads is, however, an underestimate, not an overestimate, of the effect of free access. The difference in citations between the control and treated articles that I observe may be smaller than the true difference without the error because the control articles are contaminated by the treated articles. What I report from the analysis is, therefore, going to be a downward bias, if any, due to this imperfect matching.

The identification strategy exploits a unique feature of SSRN's practice of posting articles. While authors can post their papers at any time of their choice, there is a general trend of a large number of papers submitted to SSRN for posting at once by organizations, especially when SSRN starts a new paper series for the organizations. As Figure 4 shows, the number of newly submitted papers to SSRN per month spikes whenever an organization starts a new research working paper series or submits a large number of papers for the series. This spike in submissions of papers from organizations is also illustrated in Figure 5. For example, an organization, A, submitted 445 papers in one day, May 4, 2000, when it started a new research paper series. The timing of posting these papers is decided by the organization or SSRN, not by the authors of these papers. In other words, the timing of posting is exogenous to the quality of papers. Therefore, the increase in citations after posting on SSRN can be attributed to either the time trend or free accessibility available on SSRN. The time trend is accounted for by matching the SSRN articles with non-SSRN articles that were published in the same journal, volume, and issue as described in more detail in the next paragraph. I identified 13,000 articles that were posted in a large number at once, at least more than 100 articles from the same organization in the same month. I confirmed with SSRN that these articles were posted either by SSRN or the organization, not by the authors, typically at the start of a new research paper series for the organization. Among those articles, I chose 385 articles that had been published at least 4 years prior to the posting year on SSRN.

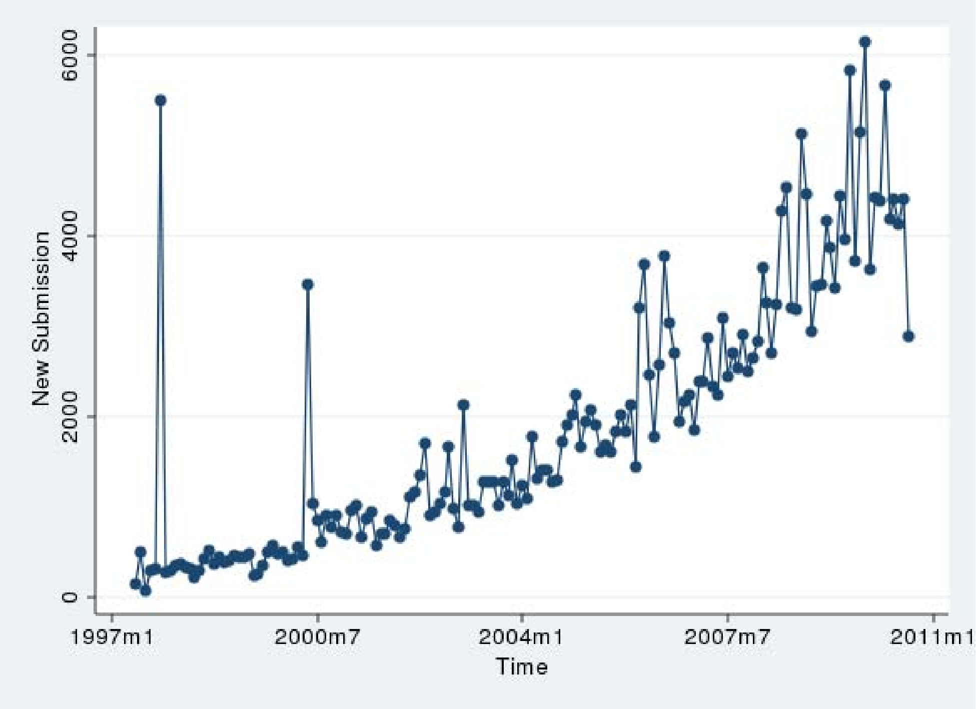

Figure 4. Number of new papers posted on SSRN per month.

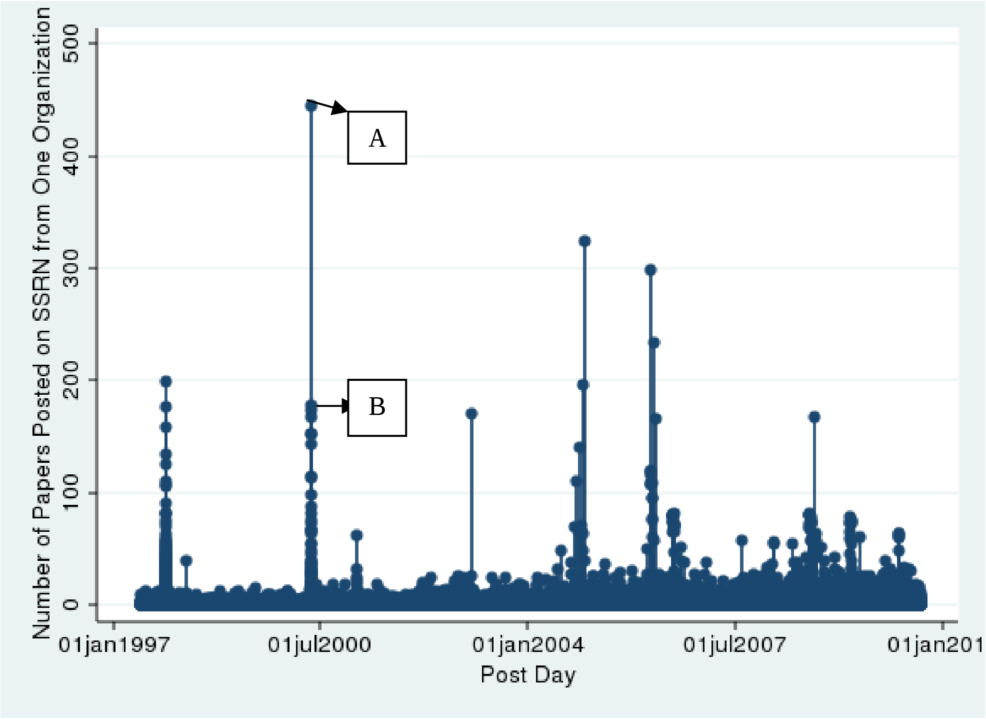

Figure 5. Number of new papers submitted from one organization in one day. Each circle represents the number of papers submitted from one unique organization on that day. For example, an organization, A, posted 445 papers in May 4, 2000, and an organization, B, posted 178 papers in the same day.

When SSRN articles were identified as published in refereed journals and subsequently chosen to be included in this study, their publication source such as journal name, volume, issue, and publication date was also identified from Web of Science. Once the publication source was identified, I collected the title, the total citations, and the authors of the articles, other than SSRN articles, published in *the same publication source* as the SSRN articles. I chose those articles that were published in the same journal, volume, and issue as SSRN articles were published as "control articles," counterparts to the SSRN articles which are thought to be the "treated" articles. The treatment is whether the articles are posted on SSRN after having been published in refereed and non-free journals for at least 4 years: the treated articles are posted on SSRN and the control articles are not, while they both were published in the same journal, volume, and issue at the same publication time. I compiled the list of the control and treated articles with their titles, authors, and publication sources. From each article in the complied list, I collected data from Web of Science on citing papers such as name of authors and publication source if each article in the complied list received any citation from other articles published in journals that Web of Science tracked. The final data set I compiled from both Web of Science and SSRN included counts of self-citations over time for both treated and control articles, counts of non-self citations over time for both treated and control articles, posting dates on SSRN for treated articles, posting organizations for treated articles, publication sources for both treated and control articles, and names of authors for both treated and control articles.

Descriptive Statistics

The descriptive statistics on both treated and control are shown in Tables 1, 2, and 3. The numbers of SSRN articles and of their matched control samples are 385 and 3,820, respectively (Table 1). The total citations that SSRN articles received, 47.2, was twice as high as those of their matching control articles, 24.4. The publication year of the SSRN articles ranged from 1970 to 2006. The average number of years for which the SSRN articles had been published when they were posted on SSRN was 10.3. The numbers of the journals and issues in which these SSRN articles were published were 165 and 337, respectively. The average journal impact factor of the journals of the SSRN articles was 2.6, ranging from 0.3 to 7.4.

In order to show the differences in the characteristics of the SSRN and their control articles before posting on SSRN, I tabulated the descriptive statistics on those before and after the posting (Table 2 and 3) separately. The SSRN articles received more citations (2.0 on average, Table 2) than their control articles (1.1 on average, Table 2), even prior to the posting year. The difference in cumulative citations, which are non-self citation counts that the articles received up to the posting year of the treated articles, is more pronounced: 12.5 for the SSRN articles and 6.8 for the control articles. These differences between the SSRN articles and their control articles even prior to the posting suggest that the SSRN articles are higher quality than their control articles. After posting on SSRN, the differences in both citations per year and cumulative citations between the SSRN articles and their control articles seem to become greater, suggesting that the posting may cause the increased gap in citation counts between the SSRN articles and their control articles (Table 3). The effect of the posting on the citation counts is quantified by the empirical equations described in the next section.

Results and Discussion

The difference in citations between SSRN-articles and non-SSRN articles before and after posting year is quite clear, as shown in Figure 5. The first graph (Figure 5a) shows all the samples, where the posting year of all of the SSRN articles is set to be zero. At the year -20, the citation of the SSRN articles seems to increase with time more than their matching control samples (Figure 5a). Therefore, excluding the articles which were published longer than 20 years before being posted on SSRN, I graphed the citation changes over time of SSRN articles and their matching control articles (Figure 5b). Even prior to being posted on SSRN, the SSRN articles showed a higher number of citations than their matching control articles. This finding is consistent with the reports of numerous studies that articles with free access tend to be of higher quality (e.g., Davis and Fromerth 2007; Kurtz et al. 2005). In this setting, the authors did

Table 1. Article characteristics for samples used in the longitudinal study. Control articles were drawn from the same journal, volume, and issue where SSRN-articles were published.

Treated Samples (SSRN articles, n=385)				
	Mean	Std.	Min.	Max.
Total citations up to 2010 since published	47.2	126.9	0	1898
Publication year	1994.5	7.3	1970	2006
Year posted on SSRN	2004.8	3.7	2000	2010
Years since publication when posted on SSRN	10.3	6.6	4	35
Number of journals where sample articles were published	165			
Number of Journal/Vol/Issue where sample articles were published	337			
Journal Impact Factor	2.6	1.5	0.3	7.4
Observations	385			
Control Samples (non-SSRN articles, n=3820)				
	Mean	Std.	Min.	Max.
Total citations up to 2010 since published	24.4	62.9	0	1387
Publication year	1992.9	7.4	1970	2006
Year posted on SSRN	Not Applicable			

Years since publication when posted on SSRN	Not Applicable			
Number of journals where sample articles were published	165			
Number of Journal/Vol/Issue where sample articles were published	337			
Journal Impact Factor	2.7	1.3	0.3	7.4
Observations	3820			

Table 2. Article-year characteristics BEFORE posting on SSRN for samples used in the longitudinal study. Control articles were drawn from the same journal, volume, and issue where SSRN articles were published.

	Treated Samples (SSRN articles)			
	Mean	Std.	Min.	Max.
Citations per year*	2.0	5.2	0	117
Cumulative citations	12.5	35.7	0	740
Year	1997.2	6.8	1971	2009
Years since publication	6.8	6.2	0	34
Years since posting on SSRN	-7.8	6.2	-35	-1
Observations	3979			
	Control Samples (Non-SSRN articles)			
	Mean	Std.	Min.	Max.
Citations per year	1.1	3.0	0	102
Cumulative citations	6.8	19.1	0	804
Year	1996.1	6.8	1971	2009
Years since published	6.9	6.1	0	34
Years since posting on SSRN	Not Applicable			
Observations	42053			

Table 3. Article-year characteristics AFTER posting on SSRN for samples used in the longitudinal study. Control articles were drawn from the same journal, volume, and issue where SSRN articles were published.

	Treated Samples (SSRN articles)			
	Mean	Std.	Min.	Max.
Citations per year	5.0	14.4	0	289

Cumulative citations	48.9	105.6	0	1898
Year	2006.6	2.8	2001	2010
Years since publication	14.4	6.4	5	40
Years since posting on SSRN	4.4	2.8	1	10
Observations	1998			
	Control Samples (Non-SSRN articles)			
	Mean	Std.	Min.	Max.
Citations per year	2.1	5.8	0	149
Cumulative citations	22.2	53.3	0	1387
Year	2006.3	2.8	2001	2010
Years since publication	15.6	6.5	5	40
Years since posting on SSRN	Not Applicable			
Observations	23235			

* Citation in all tables is non-self citation.

not choose the timing of posting but the authors may have chosen which of their articles would be posted on SSRN. Even if it was the authors' affiliated organization that chose which articles to choose, it is likely that they chose better articles for posting. Many researchers reported that the selection bias may explain the observed difference in citations between open access articles and other articles (e.g., Schwarz and Kennicutt 2004; Davis and Fromerth 2007; Kurtz et al. 2005; Moed 2007; Metcalfe 2005). Because the timing of posting was not chosen by the authors, the difference between pre-posting and after-posting, however, can be attributed to the posting on SSRN after the natural citation trend with aging is accounted for by the matching control articles.

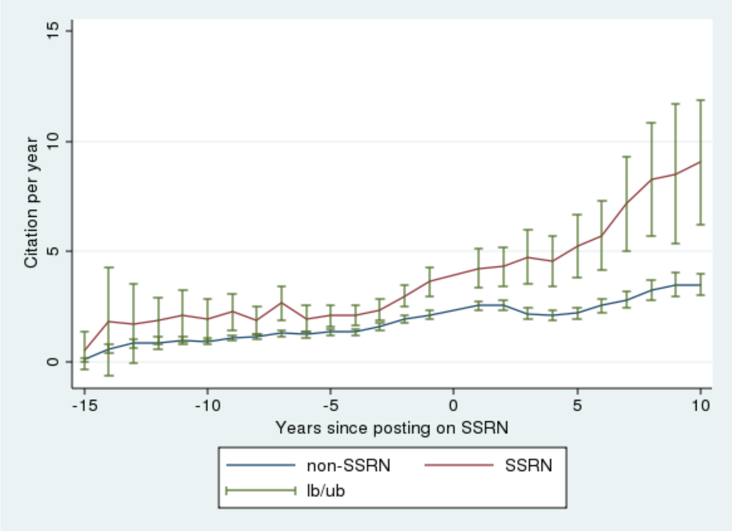

Figure 5. Citation-age profile (a) all samples; (b) a subset of SSRN and non-SSRN articles that were published less than 20 years before SSRN articles were posted on SSRN. The error bar is one standard deviation.

For a statistical analysis, I used a difference-in-difference method for panel data (Wooldridge 2007) as in the following empirical equation, similar to what was used by Furman and Stern (2011):

$$Cite_{igt} = f(\varepsilon_{igt}; \alpha_g + \lambda_t + \beta_1 SSRN_i + \beta_2 (SSRN \times After\_Posting)_{it}) \quad \text{---- (1)}$$

where $Cite_{igt}$ is citation counts that an article, *i*, received at a year, *t*, when it was published in a journal, volume, and issue, *g*. The subscripts *i, g*, and *t* indicate article, group, and time, respectively. Each group means the same journal, volume, and issue. The α_g and λ_t indicate a fixed effect for the group and citation year, respectively. *SSRN* is a binary variable, 1 if posted on SSRN at some point and 0 if not posted on SSRN. This variable is time-invariant and for all time periods it is either 1 or 0. *After_Posting* is a time-variant binary variable, equal to 1 only for years after the treated article is posted on SSRN and 0 otherwise. In this specification, I am interested in not only the effect of posting on SSRN, which is captured by the interaction term, *SSRN x After_Posting*, but also inherent differences between SSRN and non-SSRN articles, captured by the term, *SSRN*, alone. In order to show the average difference in citation counts between SSRN articles and non-SSRN articles even prior to being posted on SSRN in this specification, I included as control articles all of the research articles published in the same journal and issue as a SSRN article was published. The coefficient β_1, for the binary variable, *SSRN*, captures the possible differences between the SSRN articles and non-SSRN articles prior to being posted on SSRN.

In the conditional fixed-effect negative binomial model (4-1 in Table 4), SSRN articles appear to receive 164.5% of citations of their matching control articles, even prior to being posted on SSRN, consistent with the earlier figure. The coefficient for (*SSRN x After_Posting*), 0.158 or 1.171 as the exponentiated value, tells that the SSRN articles gained an additional 17% citation counts after being posted on SSRN compared to their counterpart control samples that were never posted on SSRN. In the model 4-2 and 4-3 where a conditional fixed effect Poisson model was used, the estimate on the coefficient for (*SSRN x After_Posting*) was 0.099 (Model 4-2 and 4-3). The standard error for the model 4-3 becomes large because the model accommodates distribution of data other than Poisson. Nonetheless, the posting on SSRN seems to increase the citation counts over 10% across all of the models at a statistical significance level of p<0.10. I attribute this gain to free access offered by SSRN. Among the three potential factors to increase citations for articles with open access identified by previous researchers, which are free access, early exposure, and quality difference, I excluded the early exposure factor because all of these articles were already published before posted on SSRN. Conditional on the assumption that the quality of articles is not correlated with the timing of the posting, the quality difference should be accounted for by the coefficient for *SSRN* but not by the coefficient for *SSRN x After_Posting*. The control articles may be available as well for free access somewhere other than SSRN. If this is the case, what is estimated by the *SSRN* coefficient in this model is an underestimate, not an overestimate of the effect of free access. It is, however, possible that what SSRN provides is not a passive free access to a research article but an active promotion. Knowing that there is no barrier to access to posted articles, the authors or the organizations that the authors are affiliated with may cite their own articles more than they would otherwise and put a link to their articles on SSRN whenever they cite these articles. In addition, SSRN provides some services to users to draw attention to popular papers or papers to suit users' specific interests. This kind of service may give additional readership for the articles posted on SSRN and increase citations. However, the promotion is not a cause but a consequence of free access.

While the above specification, (1), provides an estimate of the difference between SSRN articles and non-SSRN articles, the potential for substantial heterogeneity among articles (even though they are published in the same journal, volume, and issue) may lead to a biased estimate of the impact of SSRN posting on subsequent citation. Therefore, the article-specific fixed effect (c_i) is included as in the following specification:

$$Cite_{igt} = f(\varepsilon_{igt}; c_i + \lambda_t + \delta_{t-pubyear} + \beta_2 (SSRN \times After\_Posting)_{it}) \quad \text{----(2)}$$

This specification tests for the impact of posting on SSRN by estimating the changes in citations after an article is posted on SSRN. The age and time effect which may affect the citation counts are accounted for by including the year and age fixed effect, λ_t and $\delta_{t-pubyear}$, along with the control articles with similar characteristics. In this specification, only one control article, among the non-SSRN articles published in the same journal and issue as the SSRN article, was selected to match one SSRN article. Two other criteria for the selection of a control article, in addition to being published in the same journal and issue as the SSRN article, were used: 1) the control article should have a similar citation-year profile for 4 years prior to the posting year of its matching SSRN article and 2) the control article should have total citation counts close to its matching SSRN article up to the posting year. If no article meeting the criteria is found to match a SSRN article, the SSRN article was excluded from the analysis. As a result, the number of SSRN articles included in this analysis was smaller than in the earlier analysis. The resulting articles consist of 145 SSRN articles and 145 control articles.

Table 4. Value of free access: SSRN effect for longitudinal samples,

	Conditional Fixed Effect Negative Binomial	Conditional Fixed Effect Poisson	Quasi-ML Poisson
	(4-1)	(4-2)	(4-3)
SSRN	0.498***	0.585***	0.585***
	(0.0620)	(0.0122)	(0.1207)
	[1.645]	[1.795]	[1.795]
SSRN x After_Posting	0.158***	0.099***	0.099*
	(0.0548)	(0.0162)	(0.0552)
	[1.171]	[1.104]	[1.104]
Constant	-0.771*		
	(0.3341)		
	[0.462]		
Group Fixed Effect	Yes	Yes	Yes
Year Fixed Effect	Yes	Yes	Yes
N of article-years	71265	71265	71265
N of articles	4205	4205	4205
N of SSRN article-year	5977	5977	5977
N of SSRN articles	385	385	385
N of Journal/Vol/Is	337	337	337
N of Journal	165	165	165
Log-Likelihood	-95067	-141962	-141962

Exponentiated forms of coefficients (or Incidence-Rate Ratios) are reported in brackets.

*** p<0.01, ** p<0.05, * p<0.10

This specification was also tested with both a conditional fixed effect Poisson and negative binomial models. They were qualitatively similar and only the result from the conditional fixed effect Poisson model with robust standard error was presented in Table 5. The coefficient for *SSRN x After_Posting* was 0.122 or 112.9% (5-1). In other words, these articles gain approximately 13% in citation counts after being posted on SSRN. The magnitude is similar to what was obtained with the group fixed effect in the earlier specification (10% in Model 4-2 and 4-3). This interpretation, however, depends on the assumption that the SSRN and their control articles have the same aging profile. It is possible that SSRN articles may have longer-lived citation profiles, which would result in an upward bias on the estimate of *SSRN x After_Posting*. To address this possibility, I include a separate linear time trend term for SSRN articles, *SSRN x Age,* in (5-2) while all the other dummy variables are included as in (5-1). The coefficient for *SSRN x Age* is insignificant while the coefficient for *SSRN x After_Posting* increases, suggesting that the differences in citation profiles between SSRN articles and control articles do not cause an upward bias on the estimate of the posting effect.

In the next two models, 5-3 and 5-4, I estimate the posting effect only with SSRN articles, excluding the control articles. In the panel analysis, it is common not to include control samples. As the time-invariant fixed effect of an article is differenced out from the estimating equation, the citation change with time can be attributed to the posting on SSRN. To exclude the control articles, however, one should assume an underlying citation-age profile common to all articles. For example, McCabe and Snyder (2011) assumed citation counts to be a concave function of age, and Furman and Stern (2011) specified one of their models with a concave function of age and a polynomial expansion of calendar year. Following the functional forms in these previous studies, I included publication age and its square term in the model (5-3) along with calendar year dummy variables. The coefficient for *SSRN x After_Posting* increases in this model as the coefficients for both age and age-squared term are negative. In the next model, (5-4), a polynomial expansion of year variable was included in place of calendar year dummy variables. In both models, the coefficient for *(SSRN x After_Posting)* was significant at $p<0.05$. It seems that the estimate on the coefficient for (*SSRN x After_Posting)* seems robust to different model specifications, suggesting that the effect of free access on the diffusion of scholarly ideas is statistically significant as predicted by the theory.

Table 5. Value of Free Online Access with Article-Fixed Effect.

	Baseline diff-in-diffs specification	Interacting SSRN articles with age	Identification based only on variation within SSRN articles with age functions	Identification based only on variation within SSRN articles with age and year functions
	(5-1)	(5-2)	(5-3)	(5-4)
SSRN x After_Posting	0.122*	0.179*	0.307**	0.176**
	(0.0701)	(0.0925)	(0.1361)	(0.0885)
	[1.129]	[1.196]	[1.360]	[1.192]
SSRN x Age		-0.008		
		(0.0129)		
		[0.992]		
Age			-2.155***	-2.115***
			(0.2610)	(0.2473)
			[0.116]	[0.121]
Age-squared			-0.003***	-0.003***
			(0.0010)	(0.0010)
			[0.997]	[0.997]
Year				1.954***
				(0.2475)
				[7.056]

Year-squared				0.004***
				(0.0008)
				[1.004]
Age Fixed Effect	Yes	Yes	No	No
Calendar Year Fixed Effect	Yes	Yes	Yes	No
Article Fixed Effect	Yes	Yes	Yes	Yes
N of article-years	4425	4425	2153	2153
N of articles	290	290	145	145
Log-Likelihood	-3947	-3947	-2243	-2281

Exponentiated forms of coefficients (or Incidence-Rate Ratios) are reported in brackets.

*** p<0.01, ** p<0.05, * p<0.10

The results shown both in Table 4 and 5 are the increased citation upon posting on SSRN or SSRN-effect. Although I attribute the SSRN-effect to free access, there are other potential effects associated with posting on SSRN, except the early exposure and the selection bias that this study controlled for. SSRN is a repository, providing a database of research articles and allowing an easy search for a research article. Even if a research article is freely accessible at other sites such as its author's personal webpage, the free article may not be easily searchable and thus not be cited as it would be if posted on SSRN. This effect is not due to free access *per se*, but due to low search cost. This argument would be applicable to unpublished SSRN articles that are not available in other database. The SSRN articles in this study are, however, already published at least for four years and easily searchable in the Web of Science, a more commonly used and much more exhaustive database of published research articles. T0020he way with that I identified the publication source of a research article posted on SSRN was to match the title of the SSRN article and its authors to the Web of Science database. Therefore, by design, the SSRN article included in this study had to be searchable by the Web of Science. For a citing author to locate an old published research article only because it is available in SSRN although it is also searchable in the Web of Science, she must have an access to SSRN but not to the Web of Science. The difference between the two databases in this context is not a difference in the search cost but a difference in the access cost. SSRN is free to any user while the Web of Science is only available to subscribing individuals or the users affiliated with subscribing organizations.

Conclusion

The main contribution of this study is to report a causal relationship between free access and citations. In theory, free access to ideas should help their diffusion, and research articles with free access should receive more citations, a proxy for diffusion. However, previous empirical studies have not been able to separate the effect of free access from selection bias and have reported inconsistent findings from no or negative effect to an over-300% increase vs. citations of non-free articles. By using a natural experiment that estimates the effect of free access separate from that of confounding factors, this study identifies the effect of free access to research articles on citation counts.

Among free access, quality differential, and early viewership, which previous studies have identified as three potential factors in increasing citation counts for research articles with open access, I attribute the increase in citations after posting on SSRN to free access. However, free access may entail more than one factor. First, it can be literally the case that free, as opposed to charged, access makes a difference in number of citations. If this is the case, the composition of the authors citing the free articles after the posting may change. Readers with a limited access to non-free scholarly journals, such as researchers in industries and less developed countries where the availability of scholarly journals may be limited, may benefit from free access most. Secondly, SSRN may offer not a passive free access but active promotion of the articles with free access. For example, SSRN features some articles as "top downloaded" or "top-cited papers" and also provides users an email alert or delivery service for research articles on certain topics or those written by certain authors, suited to users' preferences. These services may actively draw more readership and citation for SSRN articles. Furthermore, searching to locate a research article may cost little because SSRN is a repository of only those research articles equipped with some search tools, and it is much easier to locate a free research article inside SSRN than searching through all the sites on the Internet. Elucidating a mechanism by which the diffusion of scholarly ideas is enhanced by free access and information technology tools is not, however, within the scope of this study and is deferred for future work.

Acknowledgements

I thank Erik Brynjolfsson and Scott Stern at MIT Sloan for their advices on the course of the research project, Gregg Gordon and Ralph Dandrea at SSRN for their data and feedback on the results, and the audience at MIT Sloan Economic Sociology Working Group Seminars and MIT Organizational Economics Lunch for their feedback.

References

Allison, P.D., and Waterman, R.P. 2002. "Fixed-Effects Negative Binomial Regression Models," *Sociological Methodology* (32:1), pp. 247-265.

Antelman, K. 2004. "Do Open-Access Articles Have a Greater Research Impact?," *College & research libraries* (65:5), p. 372.

David, P.A. 2005. "From Keeping Â€˜Natureâ€™S Secretsâ€™ to the Institutionalization of Â€˜Open Scienceâ€™," *Code: Collaborative ownership and the digital economy*), p. 85â€"108.

Davis, P.M., and Fromerth, M.J. 2007. "Does the Arxiv Lead to Higher Citations and Reduced Publisher Downloads for Mathematics Articles?," *Scientometrics* (71:2), pp. 203-215.

Davis, P.M., Lewenstein, B.V., Simon, D.H., Booth, J.G., and Connolly, M.J.L. 2008. "Open Access Publishing, Article Downloads, and Citations: Randomised Controlled Trial," *BMJ: British Medical Journal* (337).

Donovan, J.M., and Watson, C.A. 2011. "Citation Advantage of Open Access Legal Scholarship," *SSRN eLibrary*).

Eysenbach, G. 2006. "Citation Advantage of Open Access Articles," *PLoS biology* (4:5), p. e157.

Furman, J., and Stern, S. 2011. "Climbing Atop the Shoulders of Giants: The Impact of Institutions on Cumulative Knowledge Production," *American Economic Review* (101:5), pp. 1933-1963.

Gaule, P., and Maystre, N. 2011. "Getting Cited: Does Open Access Help?," *Research Policy*).

Guimaraes, P. 2008. "The Fixed Effects Negative Binomial Model Revisited," *Economics Letters* (99:1), pp. 63-66.

Harnad, S., and Brody, T. 2004. "Comparing the Impact of Open Access (Oa) Vs. Non-Oa Articles in the Same Journals," *D-lib Magazine* (10:6).

Hausman, J., Hall, B.H., and Grtltches, Z. 1984. "Econometric Models for Count Data with an Application to the Patents-R&D Relationship," *Econometrica* (52:4), pp. 909-938.

Heller, M.A., and Eisenberg, R.S. 1998. "Can Patents Deter Innovation? The Anticommons in Biomedical Research," *science* (280:5364), p. 698.

Hilbe, J.M. 2007. *Negative Binomial Regression*. Cambridge: Cambridge University Press.

Kurtz, M.J., Eichhorn, G., Accomazzi, A., Grant, C., Demleitner, M., Henneken, E., and Murray, S.S. 2005. "The Effect of Use and Access on Citations," *Information Processing & Management* (41:6), pp. 1395-1402.

Kurtz, M.J., and Henneken, E.A. 2007. "Open Access Does Not Increase Citations for Research Articles from the Astrophysical Journal," *Arxiv preprint arXiv:0709.0896*).

Lawrence, S. 2001a. "Free Online Availability Substantially Increases a Paper's Impact," *Nature* (411:6837), p. 521.

Lawrence, S. 2001b. "Online or Invisible," *Nature* (411:6837), p. 521.

McCabe, M.J., and Snyder, C.M. 2011. "Did Online Access to Journals Change the Economics Literature?," *Social Science Research Network (SSRN), January* (23).

Metcalfe, T.S. 2005. "The Rise and Citation Impact of Astro-Ph in Major Journals," *Arxiv preprint astro-ph/0503519*).

Moed, H.F. 2007. "The Effect of “Open Access” on Citation Impact: An Analysis of Arxiv's Condensed Matter Section," *Journal of the American Society for Information Science and Technology* (58:13), pp. 2047-2054.

Mokyr, J. 2002. The Gifts of Athena: Historical Origins of the Knowledge Economy. Princeton Univ Pr.

Norris, M. 2008. "The Citation Advantage of Open Access Articles,").

Rosenberg, N. 1963. "Technological Change in the Machine Tool Industry, 1840â€“1910," *The Journal of Economic History* (23:04), pp. 414-443.

Rosenberg, N. 1979. "Technological Interdependence in the American Economy," *Technology and Culture* (20:1), pp. 25-50.

Schwarz, G.J., and Kennicutt, R.C. 2004. "Demographic and Citation Trends in Astrophysical Journal Papers and Preprints," *Arxiv preprint astro-ph/0411275*).

Wooldridge, J. 2007. "What's New in Econometrics? Imbens/Wooldridge Lecture Notes; Summer Institute 2007, Lecture 10: Difference-in-Differences Estimation," *NBER. http://www.nber.org/minicourse3. html, last accessed at March* (19), p. 2009.

Zuckerman, H., and Merton, R.K. 1971. "Patterns of Evaluation in Science: Institutionalisation, Structure and Functions of the Referee System," *Minerva* (9:1), pp. 66-100.