The Effect of Open Access on the Diffusion of Scholarly Ideas

By HEEKYUNG HELLEN KIM*

This study investigates the impact of open access on the diffusion of scholarly ideas measured by citations. The key questions that this paper answers are that: 1) does open access increase citations?; 2) who benefits from the open access? By using a natural experiment on a dataset from the Social Science Research Network (SSRN), an open repository of research articles, this study identifies that the open access increases the citation counts approximately by 10% on average. The effect is not, however, homogeneous, being larger in the articles published in low-tiered journals and citing scholars in the low-income countries.

* Sloan School of Management, E62-412, Massachusetts Institute of Technology, Cambridge, MA 02139 (e-mail: hekim@mit.edu). I thank Erik Brynjolfsson and Scott Stern at MIT Sloan and Ajay Agrawal at U. Toronto 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, NBER productivity seminar, and MIT Organizational Economics Lunch for their feedback.

I. Introduction

The advent of the Internet and digital publishing has created a major disruption in the academic publishing industry, which is the emergence of open access. The open access is available in many types. For example, there are open access journals that offer free readership to their readers while charging upfront submission fees to the researcher submitting work. Traditional subscription-based journals also allocate some space for open access articles that are free to readers, but charged to the authors. Some academic fields often have open online repositories of research articles where published or unpublished manuscripts are

posted for free readership. The availability of open access has prompted many research questions. First and foremost, what is the impact of the open access on citation counts, a proxy for the diffusion of scholarly ideas? The research questions that this paper answers are 1) how much the open access changes citation counts; 2) how different in characteristics the articles with open access are from those without open access; 3) who benefits from the open access.

The academic publishing industry, approximately $20 billion market, has only a few large publishers who have attained a monopoly-like stranglehold on the academic world and consequently commands between 30 and 40% profit margin with 7% increase per year in subscription price for the last decade, despite of the marginal cost for electronic copies being zero. With the price hike, many universities encouraged their researchers to submit their work to open access journals. The open access journals had grown in numbers dramatically from 1,100 in 2004 to 9,800 in 2013, as listed in the directory of open access journals. Some subscription-based journals also offer authors to buy open access to their articles, having both open access articles to the readers and articles with pay-wall in the same issue. Many institutions also practice open access mandate. Most notably, the National Institute of Health (NIH) requires all scholars funded by NIH to submit an electronic version of their final, peer-reviewed manuscripts upon acceptance for publication, to the National Library of Medicine's PubMed Central since 2009. More than 50 research funders and over 150 universities in the world mandate open access self-archiving[1]. While the open access has become a new emerging outlet for research articles, scholars hardly reached a consensus on the impact of the open access and very few studies have investigated the beneficiaries of the open access. Measuring the effect of open

[1] This open access mandate was challenged by the introduction of a bill in the US, the Research Works Act, which eliminates such a requirement by federal government agencies, not surprisingly with the sponsorship of a large publishing company. This bill prompted many scientists to boycott the publishing company, which withdrew the sponsorship for the bill, and was later abolished.

access separate from other confounding factor such as quality difference on research articles posted for free readership on a free online repository, this paper reconciles the differences among scholars. Moreover, the analysis of the readers and authors in this paper demonstrates a heterogeneous effect of the open access.

The difficulty to measure the effect of open access arises because two other factors are often confounded with open access. The first is quality difference. The open access journals require authors to pay upfront submission fees, resulting in authors' selecting the articles appropriate for open access journals. Even among the articles that have been already published somewhere else, authors often select their better articles to self-archive in the Internet for free readership. More prominent and research-oriented institutes and universities tend to have a better-maintained and well-organized website to facilitate posting their members' research articles as well. Therefore, an apparent difference in citations between the articles with and those without open access may be 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. There are several large open online repositories of research articles such as arXiv.org and SSRN, which allow authors to post any research article without peer-review. Therefore, a difference in citations between article with and those without open access may arise from that the articles with open access had a longer time to receive a citation than the articles without open access (Moed 2007). A cross-sectional study to compare citations between the articles with and those without open access fails to separate the confounding factors and what appears to be the effect of open access can be the effect of other factors. This paper is, to the best of the author's knowledge, the first to disentangle the confounding factors and separately measures the effect due to the quality difference and open access.

To identify the effect of open access and quality difference, 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 two aspects of SSRN to identify the effect of online open access and quality difference separately. 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 for posting 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 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 from the same journal and issue.

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 open 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 separated the effect of open access from the effect of quality difference on citations and estimated each effect. The first difference in citations between the treated articles (SSRN articles) and the control articles (non-SSRN articles) before the posting year on SSRN was attributed to the quality difference while the second difference in citations between the treated and the control articles after the posting year was attributed to the open access. 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 be driven by the open access that SSRN provides, much smaller than what was reported by previous studies (i.e., Lawrence, 2001a).

A previous study investigated the digitization effect of journals on the citation and reported that digitized articles tend to be cited more, probably due to easy access to the articles (Evans, 2008), but it has not been studied that the open access (among digitized articles) causes heterogeneous effects across the citing and cited authors. This study investigated the differential effect across the authors (supply) and the readers or citing scholars (demand). First, the cited articles were divided into two groups in various criteria: 1) high vs. low journal impact factor, 2) high vs. low profile institutes, and 3) high vs. low profile authors. Secondly, the readers or the citing authors were also divided into two groups: 1) high vs. low income countries, 2) across vs. within the discipline between the cited and citing authors, and 3) distant vs. close in geographic distance between the authors and

the readers. In the supply side, the observed increase appears to be driven by the research articles previously published in low-tiered journals or journals with low journal impact factor. It is consistent with the fact that top-tiered journals tend to be subscribed more widely and be accessible before the SSRN-posting while the low-tiered journals may not be widely distributed and the SSRN-posting may provide an additional outlet for the articles published in the low-tiered journals to the readers. In the demand side of the effect, SSRN-posting appears to increase the portion of citing authors from the low income countries.

The contribution of this paper is that: 1) to identify a causal relationship between open access and citations; 2) to separate the quality difference from the open access and quantify the two effects separately; and 3) most importantly, to demonstrate a heterogeneous effect of the open access on the authors and the readers of the articles. The result of this study reconciles the differences among scholars on the open access advantage, showing that the effect of open access is not as large as what was reported by some scholars (i.e. Lawrence, 2001a) but still larger than what was reported as unobservable by other scholars (i.e. Davis et al., 2008). This study shows that the apparent increase in the citations for the open access articles mostly come from the quality difference, much more than from the open access. More importantly, this study is the first to show the heterogeneous effect of the open access at the article level, advancing the current understanding about the effect of open access at journal or volume level (McCabe and Snyder, 2013).

II. Open Access Debate

The debate on the existence of an "open access advantage" for scholarly communications started when Lawrence (2001) 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, 2001), 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), multi-disciplinary sciences (Eysenbach, 2006), economics (McCabe and Snyder, 2011), and science (McCabe and Snyder, 2013). Other 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 open 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 closed access research articles; early exposure and quality difference have been identified as the potential confounding factors for the overestimation of the effect of open 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 open access cannot be made.

Moreover, the heterogeneous effect of the open access, if any, has not been fully investigated by the previous studies that reported the open access advantage. McCabe and Snyder (2013) reported a heterogeneous effect of the open access on citations at journal volume level, but not at article level. To the best of the author's knowledge, however, there has not been any study to report the heterogeneous effect of the open access at article level. The beneficiaries of the open access can be categorized in two groups in terms of supply and demand: the supply side is the authors of the open research articles and the demand side is the readers or other scholars to cite those articles. It is an empirical question whether the open access helps the articles written by well-known authors or published in well-known journals more than those with less-known authors or journals, in other words, the open access promotes Superstar effect. Alternatively, the open access can promote the readership of articles written by less-known authors or published in less-known journals, providing an additional outlet for those articles. This phenomenon would be similar to Long-tail effects shown in the sales and rentals in the digital market (Brynjolfsson et al., 2011). With respect to the demand side, the readers in the organizations and countries without access to a wide range of journals may get the benefit from the open access. While there is a report that the scholars in the developing country such as India may not have access to research articles as widely as those in the developed countries have (Gaule, 2009), no study has empirically shown that the open access helps the readers in the developing countries more than those in the developed countries. This study reports the heterogeneous effect of the open access, empirically showing that the research articles published in lower-tiered journals tend to get a boost in citations after open access and the readers in the developing countries tend to cite the research articles more after their being posted on SSRN.

III. Identification Strategy

Previous studies have identified three factors may cause the increases in citations for research articles with open access: 1) open 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 a difference-in-difference method to separate the effect of open 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 open access. In this setting, authors do not choose the time to post their articles for open access; instead, the organizations that the authors are affiliated with or the websites that host those articles choose the time to post articles for open 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 open 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 open 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 that 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 quality difference 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 open access.

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 open 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 open access.

For the most statistical analysis when the dependent variable was citation counts, 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 or Poisson Quasi-Maximum Likelihood (PQML) estimator. For some analyses on the

heterogeneous effect when the fraction was the dependent variable, I used a time-series ordinary least square (OLS) model.

IV. Data

A. Data Construction and Source

The data source I relied on for this study is SSRN, complemented with the Web of Science. 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 (Figure 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.

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 and the affiliations 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 open access. The estimate to which this error leads is, however, an underestimate, not an overestimate, of the effect of open 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 day spikes whenever an organization starts a new research working paper series or submits a large number of papers for the series. 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 open 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. 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.

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 close access 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 and affiliation 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, counts of non-self citations over time, publication sources, and the names and affiliations of authors and citing authors for both treated and control articles and posting dates on SSRN for treated articles.

B. Supply Side Analysis

For the analysis of the heterogeneous effect on the supply side, I divided the sample into two groups in various criteria. The first criterion was the Journal Impact Factor (JIF) published by Web of Science as of 2006. Some articles were published in the journals where the Journal Impact Factor was not known in 2006. Excluding those articles, I compiled 283 articles and they were grouped into the two: one is published in the journals above the median JIF and the other published in the journals below the median JIF. The two different groups were analyzed separately or with an interaction term in order to investigate the heterogeneous effect of the open access on the citations, which may differ depending on the prestige and thus the degree of the distribution of the journals where the articles were originally published. The second criterion was the quality of the institute with which the authors were affiliated. From the list used by Kim et al.(Kim et al., 2009), the top 25 institutes in the field of finance were identified. A binary variable was set to be 1 for the article published in the field of finance (and economics) and authored by at least one scholar affiliated with the top institute. The last criterion was the quality of the authors. Using the ranking of economists published by Repec as of 2011, I assigned a binary variable to be 1 for an economist if she is in the top 10% economists and 0 otherwise. I also selected the articles only published in the economics journals listed under the subject of economics categorized by Web of Science.

C. Demand Side Analysis

The readers or the citing scholars were also grouped in various ways. The first was the countries to which the citing authors' affiliations belonged. I collected the address of the authors of all the articles citing the treated and control articles. This analysis is to test a hypothesis that the open access will give the benefit to the readers in the developing countries, where the access to scholarly journals may be limited, more than those in the developed countries. Using the definition of the developed country by Wikipedia, I marked the country in the citing author's address as either developing or developed country. The fraction of the citing authors in all citing authors was estimated in two ways: one was to count the number of citing articles authored by scholars all from the developing countries and get the fraction of the number of those articles among the number of all the citing articles in each year; the second was to count the number of authors from the developing countries in each citing article, get the fraction of those authors in each citing article, and average the fraction over each year. The fractions in both estimates were dependent variables in this analysis, unlike the previous analyses, and a time-series OLS was used for the estimation of the SSRN effect.

The second criterion was the distance between the cited author and citing author. The research question was whether the SSRN-posting makes the distance closer or further apart. The distance was measured both in knowledge base and in geography. For the analysis on the knowledge distance, the articles published in journals of the three fields only, which are economics, finance, and law. The list of the journals under each field was obtained from Web of Science. Economics and finance were considered as one field as there are many journals cross-listed under the two subjects and law was considers as the other field. Only when the cited article was published in the journal in the fields of either economics/finance

or law, the article was selected for the analysis. Then the source of the citing articles for the selected articles was parsed and checked as either "within" or "across". The citing article was marked as "within" if it was published in the journal in the same field where the cited article was published and as "across" otherwise.

The geographic distance between the cited and the citing author was estimated only between the reprint authors or the first authors if the reprint author is not available. The articles were selected only when those authors' addresses were in the USA. The distance, averaged over each year, was regressed on the SSRN variables. A time-series OLS was used for this analysis, because the dependent variable was not a count variable as citations in the previous analysis.

V. Results and Discussion

A. Descriptive Statistics

The descriptive statistics on both treated and control are shown in Tables 1, 2, and 3. The numbers of SSRN articles and 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.

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 of 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.

B. Average Effect

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 posting year of all the SSRN articles is set to be zero. 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 open access tend to be of higher quality (e.g., Davis and Fromerth, 2007; Kurtz, Eichhorn, Accomazzi, Grant, Demleitner, Henneken and Murray, 2005). In this setting, the authors did 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 post, 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, Eichhorn, Accomazzi, Grant,

Demleitner, Henneken and Murray, 2005; Moed, 2007; Metcalfe, 2005). The first difference in citation between the treated and the control articles before posting on SSRN can, therefore, indicate the inherent quality difference between the treated and the control. The second 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 and the quality difference between the treated and control articles is measured by the first difference, because the timing of posting was not chosen by the authors.

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):

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

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 open access offered by SSRN. Among the three potential factors to increase citations for articles with open access identified by previous researchers, which are open 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 open 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 open access. It is, however, possible that what SSRN provides is not a passive open 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 heterogeneous effect by the open access that the articles previously published in the low-tiered journal tend to get the most of the boost, as discussed later, suggests that the increased citations are likely due to the open 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:

(2) $Cite_{igt} = f(\varepsilon_{igt};\ c_i + \lambda_t + \delta_{t-pubyear} + \beta_3(SSRN\ \times After\_Posting)_{it})$

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.

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 open access on the diffusion of scholarly ideas is statistically significant as predicted by the theory.

C. Heterogeneous Effect on Supply Side

The benefit of the open access may differ across the authors or the articles posted on SSRN. Because the articles of this study were all published in a journal before being posted on SSRN, the prestige, often measured by the Journal Impact Factor, of the journals where the articles were published was known. The Journal Impact Factor changes with year and JIF in the year of 2006 was used for this analysis. The median of the Journal Impact Factor of the journals where the sample articles were published was 1.92. Approximately 100 articles were published in a journal, of which Journal Impact Factor (JIF) was not available in the year of 2006. Excluding those articles, the articles were grouped into two: one is published in the journals above 1.92 of Journal Impact Factor and the other published in the journals below 1.92 of Journal Impact Factor. As shown in Table 6, the overall effect of the open access on the citation becomes statistically insignificant due to a smaller size of samples (6-1). The effect of open access on

the articles published in the below-median JIF journals is, however, much stronger and statistically significant (6-2 and 6-3). To illustrate the difference in the effect between the low-tiered journal and the high-tiered journal, an interaction term was included in the model 6-4 and the SSRN-posting effect was much smaller in the articles published in the journals above median JIFs. The result suggests that the open access provides an additional outlet for the articles published in less distributed journals rather than further promotes the articles that were already well-known. This result suggests that the increased citations upon the posting on SSRN is likely to be due to open access not promotion following the posting, because the research articles already published in the top-tiered journals would have been promoted more than, at least equal to, those published in the low-tiered journals if any.

It can be still the case, however, that the articles published in low-tiered journals by high-profile authors receive a disproportionate boost from the SSRN-posting. This question is answered in the table 7, 8, and 9. Among the articles published in the field of finance, the authors affiliated with top institutes seem to have a much larger boost in the citations from the SSRN-posting (7-2). When the articles were selected when they were published in a combined list of journals in the economics and finance, the effect of top institute becomes statistically insignificant (8-2). The SSRN tends to host working paper series from the top institutes and the model 8-2 shows that the SSRN effect becomes insignificant when the top institute variable was included, suggesting that the quality differentials between SSRN and non-SSRN articles can be captured by that the author is affiliated with top institutes or not. A similar effect is observed for the high profile researchers. Instead of categorizing the institutes as in Table 7 and 8, I ranked researchers themselves and divided them into a group of top 10% as ranked by Repec as of 2011 and the rest. The top 10% economists do not seem to have a boost from the SSRN-posting (9-2 and 9-3). They receive a higher citation

count regardless of the SSRN. There is, however, a limitation of this analysis: the ranking of the economists change with time. I used the ranking from the year 2011, while the publication years and posting years articles in this analysis range from 1990s to 2006.

D. Heterogeneous Effect on Demand Side

On the demand or the readers, the benefit of the open access may not be the same. While it is known that the scholars in the developing countries have a limited access to the academic journals compared to those in the developed countries (Gaule, 2009), the differential effect of the open access advantage across countries of different income has not been reported yet. I collected the addresses of the citing authors and identified the countries where they are located. If all authors of a citing article are affiliated with organizations in the developing countries, the citing article is marked as one from the developing countries. If at least one author of a citing article is affiliated with organizations in the developed countries, the citing article is marked as one from the developed countries. After marking all of the citing articles either as one from the developing countries or as one from the developed countries, the ratio of those from the developing countries to all the citing articles each year was estimated. Similarly to the regression of (1), the ratio was regressed on *SSRN* and *SSRN x After_Posting*. In this analysis, the dependent variable is no longer a count variable but a fraction and a time-series OLS was used. As shown in Table 10, the ratio increases upon the SSRN-posting. The prestige of the journals does not appear to make a difference (10-2). In the next analysis, the fraction of the authors from the developing countries in all the authors in each citing article was estimated. For example, a citing article was authored by two scholars; one from developing country and the other from developed country. Then the fraction is 0.5. This fraction was obtained for each

citing article and averaged over each year. The result of the regression, when the fraction was the dependent variable, shows that, upon SSRN-posting, the fraction increases (11-1). This increase is mostly driven by the low-tiered journals (11-2, 11-3, and 11-4), because when the articles are divided into two groups, one published in journals with the above-mean JIF and the other published in journals with the below-mean JIF, the ratio only in the articles published in the below-mean JIF increases upon the SSRN-posting (11-2). The result suggests that the scholars or citing authors that have a limited access to the articles published in journals distributed less widely get an access to those articles upon the SSRN-posting.

The open access may promote the citation across the field more than the citation within the field. The scholars may get access to the journals in their own field whether they are available in their affiliated organizations or not through other channels such as contacting the authors and their peers. If this is the case, the open access may promote the citation across the field. To see this effect, the samples were first reduced only to those published in the journals in the field of economics, finance, and law. The reduced samples were regressed as before and compared to the all samples (12-1 for all samples and 12-2 for reduced samples). The SSRN-posting effect is insignificant in both samples. The next analysis was to make the citations only by the articles in the same field per year as the dependent variable (12-3). The result does not show the SSRN-effect. The next model (12-4) uses a time-series OLS with the fraction of within-field citing to all citing articles being the dependent variable. Again, the SSRN-effect does not appear to affect the fraction.

The spillover of knowledge over geographic distance has been studied by many researchers (i.e., Griffith et al., 2011;Abramovsky and Simpson, 2011). I investigate whether the open access changes the distance between the citing and the cited authors. As shown in Table 13, the geographic distance between the

citing and cited authors does not change upon the SSRN-posting. By the time that these articles are posted on SSRN, they were published already for 4 years by the design of this study. In addition, the samples included in this analysis were only those with both citing and cited authors being located in the USA. It is not, therefore, surprising that the distance is not affected by the SSRN-posting.

E. Other Potential Factors

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 open access, there are other potential effects associated with posting on SSRN, except the early exposure and the selection bias that this study controlled for. The first is a low search cost associated with SSRN. 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 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 open access *per se*, but due to low search cost. This argument would be applicable to unpublished SSRN articles that are not available in other widely used 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. The 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 open to any user while the Web of Science is only available to subscribing individuals or the users affiliated with subscribing organizations. The second is a marketing or promotion of the research articles by the SSRN. SSRN not only provides a passive outlet to post research articles but send a personalized email to its users for newly posted articles. The increased citation may be due to this new promoting effect by SSRN not by the open access. As shown in Table 6, the boost in the citation is observed only in the articles published in the low-tiered journals. If the promotion had increased the citation, the effect would be even more pronounced for the articles published in the high-tiered journals. Therefore, the wider access to the articles is a likely cause for the boost.

VI. Conclusion

The main contribution of this study is to report the heterogeneous effect of the open access as well as the causal relationship between open access and citations. In theory, open access to ideas should help their diffusion, and research articles with open access should receive more citations, a proxy for diffusion. However, previous empirical studies have not been able to separate the effect of open access from selection bias and have reported inconsistent findings from no or negative effect to an over-300% increase compared to citations of close-access articles. By using a natural experiment that estimates the effect of open access separate from that of confounding factors, this study identifies the effect of open access to research articles on citation counts and separately estimates the quality differentials. The effect is not as large as some previous studies have reported. It is less than 20%. When the sample size was reduced, the effect became

statistically insignificant. The selection bias or the quality differential explained the most of the increased citations of the open access articles.

However, the effect was not heterogeneous across the authors and readers. The articles previously published in low-tiered journals drove the effect while the articles published in high-tiered journals did not get a boost in citations from the open access. After the free posting of the articles, their citing authors tend to become more from developing countries than they were before the free posting. The distance in both knowledge and geography between the citing and cited authors doesn't seem to change with the free posting of the articles.

REFERENCES

Abramovsky, Laura and Simpson, Helen. "Geographic Proximity and Firm-University Innovation Linkages: Evidence from Great Britain." *Journal of Economic Geography*, 2011, *11*(6), pp. 949-77.

Allison, P.D. and Waterman, R.P. "Fixed-Effects Negative Binomial Regression Models." *Sociological Methodology*, 2002, *32*(1), pp. 247-65.

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

Brynjolfsson, E.; Hu, Y.J. and Simester, D. "Goodbye Pareto Principle, Hello Long Tail: The Effect of Search Costs on the Concentration of Product Sales." *Management Science*, 2011, *57*(8), pp. 1373-86.

Davis, P.M. and Fromerth, M.J. "Does the Arxiv Lead to Higher Citations and Reduced Publisher Downloads for Mathematics Articles?" *Scientometrics*, 2007, *71*(2), pp. 203-15.

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

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

Evans, J.A. "Electronic Publication and the Narrowing of Science and Scholarship." *science*, 2008, *321*(5887), pp. 395-99.

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

Furman, J. and Stern, S. "Climbing Atop the Shoulders of Giants: The Impact of Institutions on Cumulative Knowledge Production." *American Economic*

Review, 2011, *101*(5), pp. 1933-63.
Gaule, Patrick. "Access to Scientific Literature in India." *Journal of the American Society for Information Science & Technology*, 2009, *60*(12), pp. 2548-53.
Griffith, Rachel; Lee, Sokbae and Van Reenen, John. "Is Distance Dying at Last? Falling Home Bias in Fixed-Effects Models of Patent Citations." *Quantitative economics*, 2011, *2*(2), pp. 211-49.
Guimaraes, P. "The Fixed Effects Negative Binomial Model Revisited." *Economics Letters*, 2008, *99*(1), pp. 63-66.
Harnad, S. and Brody, T. "Comparing the Impact of Open Access (Oa) Vs. Non-Oa Articles in the Same Journals." *D-lib Magazine*, 2004, *10*(6).
Hausman, J.; Hall, B.H. and Grtltches, Z. "Econometric Models for Count Data with an Application to the Patents-R&D Relationship." *Econometrica*, 1984, *52*(4), pp. 909-38.
Hilbe, Joseph M. *Negative Binomial Regression*. Cambridge: Cambridge University Press, 2007.
Kim, E.H.; Morse, A. and Zingales, L. "Are Elite Universities Losing Their Competitive Edge?" *Journal of Financial Economics*, 2009, *93*(3), pp. 353-81.
Kurtz, M.J.; Eichhorn, G.; Accomazzi, A.; Grant, C.; Demleitner, M.; Henneken, E. and Murray, S.S. "The Effect of Use and Access on Citations." *Information Processing & Management*, 2005, *41*(6), pp. 1395-402.
Kurtz, M.J. and Henneken, E.A. "Open Access Does Not Increase Citations for Research Articles from the Astrophysical Journal." *Arxiv preprint arXiv:0709.0896*, 2007.
Lawrence, S. "Online or Invisible." *Nature*, 2001, *411*(6837), pp. 521.
McCabe, M.J. and Snyder, C.M. "Did Online Access to Journals Change the Economics Literature?" *Social Science Research Network (SSRN), January*, 2011, *23*.
____. "Identifying the Effect of Open Access on Citations Using a Panel of Science Journals." 2013.
Metcalfe, T.S. "The Rise and Citation Impact of Astro-Ph in Major Journals." *Arxiv preprint astro-ph/0503519*, 2005.
Moed, H.F. "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*, 2007, *58*(13), pp. 2047-54.
Norris, M. "The Citation Advantage of Open Access Articles." 2008.
Schwarz, G.J. and Kennicutt, R.C. "Demographic and Citation Trends in Astrophysical Journal Papers and Preprints." *Arxiv preprint astro-ph/0411275*, 2004.
Wooldridge, JM. "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, 2007, *19*, pp. 2009.

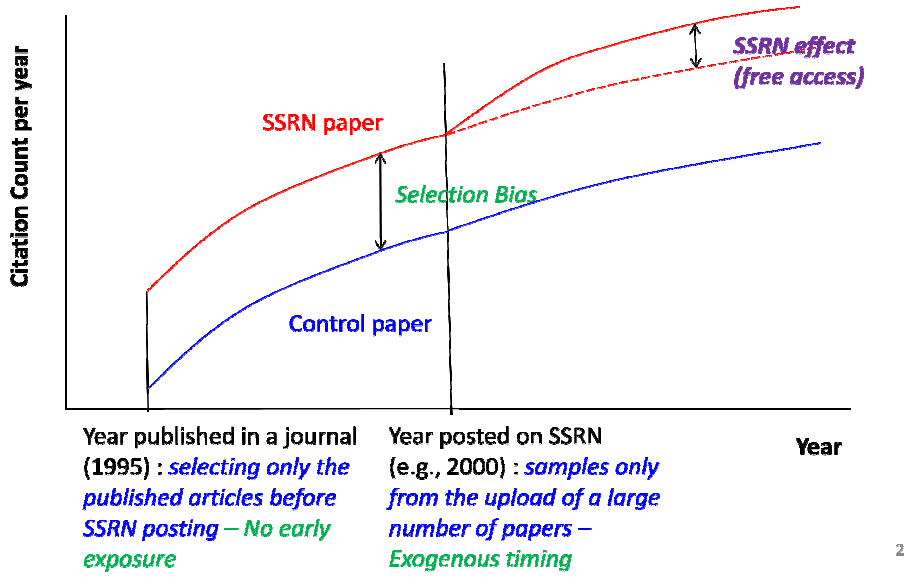

FIGURE 1. IDENTIFICATION STRATEGY

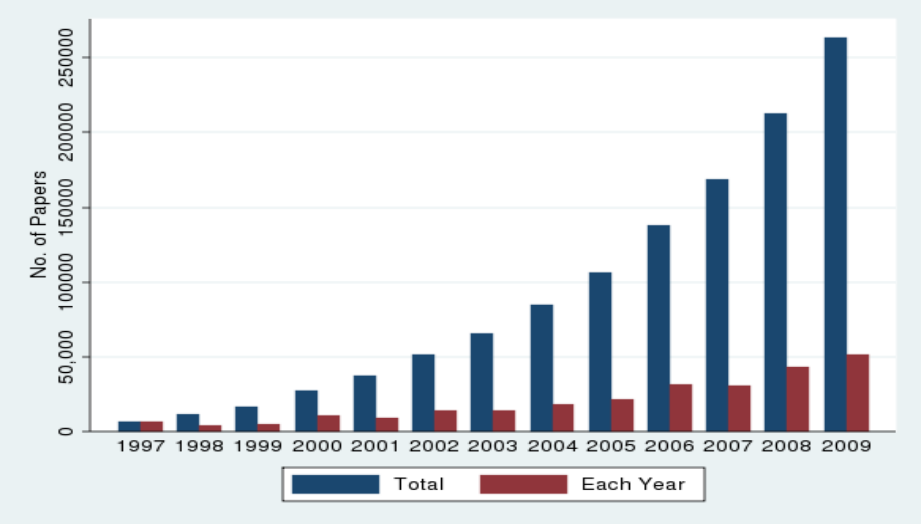

FIGURE 2. NUMBER OF PAPERS POSTED ON SSRN

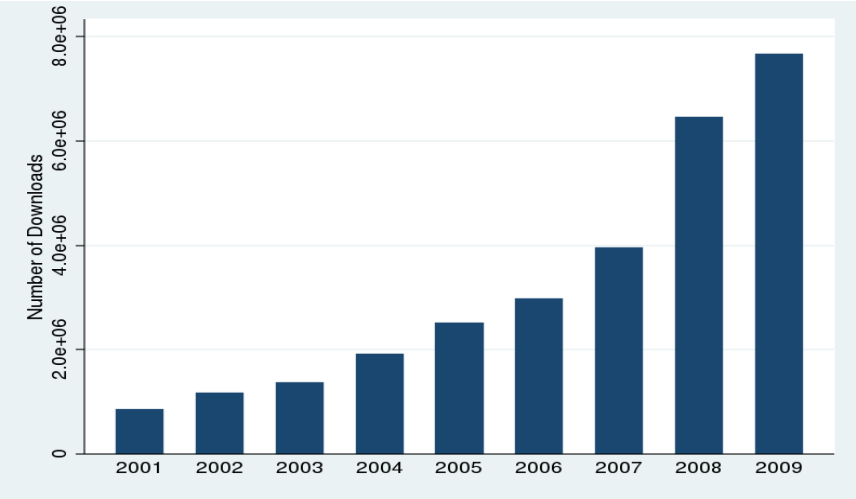

FIGURE 3. NUMBER OF DOWNLOADS.

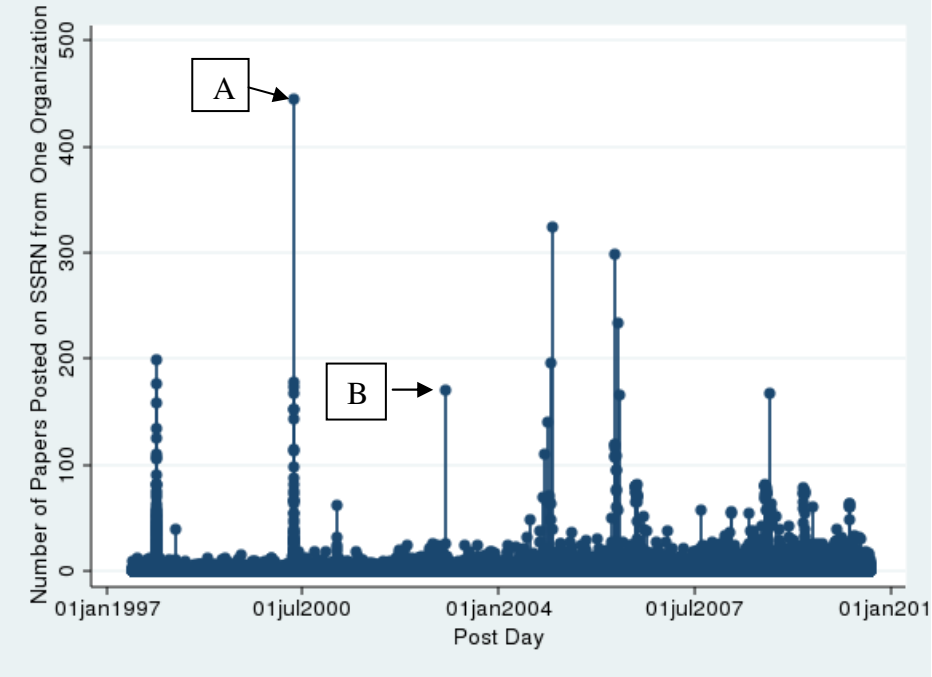

FIGURE 4. NUMBER OF NEW PAPERS SUBMITTED FROM ONE ORGANIZATION IN ONE DAY

Notes: 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.

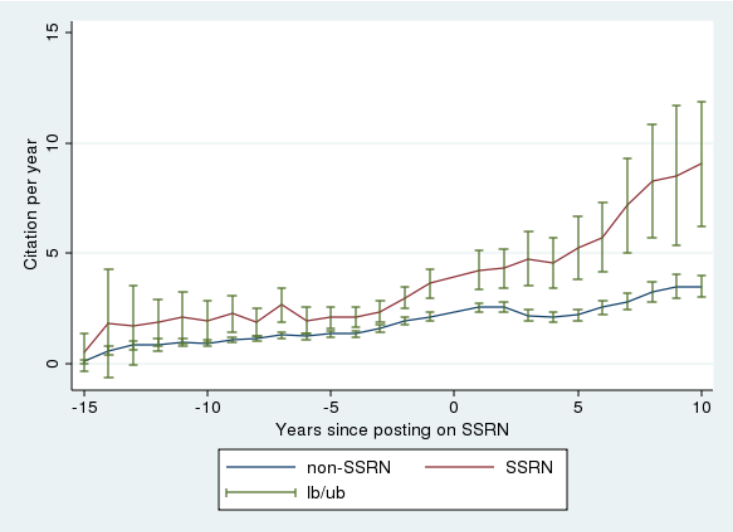

FIGURE 5. CITATION-AGE PROFILE

Notes: The error bar is one standard deviation.

TABLE 1—ARTICLE CHARACTERISTICS FOR SAMPLES USED IN THE LONGITUDINAL STUDY

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			
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			
Observations	3820			

Notes: Control articles were drawn from the same journal, volume, and issue where SSRN-articles were published.

TABLE 2–ARTICLE-YEAR CHARACTERISTICS BEFORE POSTING ON SSRN FOR SAMPLES USED IN THE LONGITUDINAL STUDY

	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			

Notes: Control articles were drawn from the same journal, volume, and issue where SSRN articles were published.

TABLE 3–ARTICLE-YEAR CHARACTERISTICS AFTER POSTING ON SSRN FOR SAMPLES USED IN THE LONGITUDINAL STUDY

	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			

Notes: Control articles were drawn from the same journal, volume, and issue where SSRN articles were published.

* Citation in all tables is non-self citation.

TABLE 4. VALUE OF OPEN ACCESS: SSRN EFFECT FOR LONGITUDINAL SAMPLES

	Conditional Fixed Effect Negative Binomial (4-1)	Conditional Fixed Effect Poisson (4-2)	Quasi-ML Poisson (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

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

*** Significant at the 1 percent level. ** Significant at the 5 percent level.* Significant at the 10 percent level.

TABLE 5–VALUE OF FREE ONLINE ACCESS WITH ARTICLE-FIXED EFFECT.

	(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

TABLE 6–THE HETEROGENEOUS EFFECT ON THE SUPPLY I: JOURNAL IMPACT FACTOR OF THE JOURNALS WHERE THE TREATED AND CONTROL SAMPLE ARTICLES WERE PUBLISHED BEFORE POSTING YEAR

DV=Citation per year	All articles (6-1)	Articles published only in the journals with JIF <= 1.9 (6-2)	Articles published only in the journals with JIF > 1.9 (6-3)	All articles (6-4)
Citation per year				
SSRN	0.393**	0.571*	0.217	0.599*
	(0.1485)	(0.2464)	(0.1278)	(0.2397)
	[1.482]	[1.770]	[1.243]	[1.820]
SSRN x Post_SSRN	0.047	0.238*	-0.181	0.184
	(0.0828)	(0.1056)	(0.1093)	(0.1074)
	[1.049]	[1.269]	[0.835]	[1.202]
SSRN x High_JIF				-0.412
				(0.2718)
				[0.663]
SSRN x Post_SSRN x High_JIF				-0.315*
				(0.1452)
				[0.730]
N of article-years	27617	13359	14208	28266
N of articles	2908	1620	1282	3018
N of SSRN article-year	3235	1973	1262	3235
N of SSRN articles	283	195	88	283
N of Journal/Vol/Is	264	187	77	283
N of Journal	132	112	21	146
Log-Likelihood	-77083	-30986	-44950	-77610

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

*** Significant at the 1 percent level. ** Significant at the 5 percent level.* Significant at the 10 percent level.

TABLE 7. THE HETEROGENEOUS EFFECT ON THE SUPPLY II: CITATION OF SSRN-ARTICLES AUTHORED BY AT LEAST ONE SCHOLAR FROM TOP INSTITUTES (THE FIELD OF FINANCE ONLY)

	(7-1)	(7-2)
Citation per year		
SSRN	0.208	0.822*
	(0.1137)	(0.3469)
	[1.232]	[2.275]
SSRN x Post_SSRN	0.297*	-0.392*
	(0.1216)	(0.1830)
	[1.346]	[0.676]
Top Institute		0.590*
		(0.2480)
		[1.804]
SSRN x Top Institute		-0.772*
		(0.3686)
		[0.462]
SSRN x Post_SSRN x Top Institute		0.755***
		(0.2261)
		[2.128]
N of article-years	3284	3284
N of articles	299	299
N of SSRN article-year	438	438
N of SSRN articles	33	33
N of Journal/Vol/Is	31	31
N of Journal	9	9
Log-Likelihood	-9911	-9628

Notes: Exponentiated forms of coefficients (or Incidence-Rate Ratios) are reported in bracket

*** Significant at the 1 percent level. ** Significant at the 5 percent level.* Significant at the 10 percent level.

TABLE 8. THE HETEROGENEOUS EFFECT ON THE SUPPLY III: CITATION OF SSRN-ARTICLES AUTHORED BY AT LEAST ONE SCHOLAR FROM TOP INSTITUTES (THE FIELD OF ECONOMICS AND FINANCE)

DV= Citation per year	(8-1)	(8-2)
Citation per year		
SSRN	0.317**	0.213
	(0.1157)	(0.1903)
	[1.373]	[1.237]
SSRN x Post_SSRN	-0.041	-0.024
	(0.0901)	(0.1424)
	[0.960]	[0.977]
Top Institute		0.403***
		(0.0775)
		[1.497]
SSRN x Top Institute		0.024
		(0.2304)
		[1.025]
SSRN x Post_SSRN x Top Institute		-0.023
		(0.1648)
		[0.977]
N of article-years	22496	22496
N of articles	2255	2255
N of SSRN article-year	2409	2409
N of SSRN articles	194	194
N of Journal/Vol/Is	175	175
Log-Likelihood	-63889	-63007

Notes: Exponentiated forms of coefficients (or Incidence-Rate Ratios) are reported in bracket

*** Significant at the 1 percent level. ** Significant at the 5 percent level.* Significant at the 10 percent level.

TABLE 9. THE HETEROGENEOUS EFFECT ON THE SUPPLY IV: HIGH-PROFILE RESEARCHERS

	(9-1)	(9-2)	(9-3)
Citation per year			
SSRN	0.339**	0.130	0.125
	(0.1304)	(0.1364)	(0.1363)
	[1.404]	[1.138]	[1.133]
SSRN x Post_SSRN	-0.111	-0.189	-0.189
	(0.0955)	(0.1261)	(0.1261)
	[0.895]	[0.828]	[0.828]
High-Profile Researcher		0.343***	0.335***
		(0.0947)	(0.0954)
		[1.409]	[1.398]
High-Profile Researcher x SSRN		0.160	0.395
		(0.2022)	(0.2572)
		[1.174]	[1.485]
High-Profile Researcher x SSRN x Post_SSRN		0.072	0.118
		(0.1449)	(0.1737)
		[1.075]	[1.125]
High-Profile Researcher x SSRN x High-JIF			-0.386
			(0.2783)
			[0.680]
High-Profile Researcher x SSRN x Post_SSRN x High-JIF			-0.117
			(0.1815)
			[0.890]
N of article-years	20133	20133	20133
N of articles	2036	2036	2036
N of SSRN article-year	2158	2158	2158
N of SSRN articles	173	173	173
N of Journal/Vol/Is	155	155	155
N of Journal	64	62	64
Log-Likelihood	-55967	-55092	-54907

TABLE 10-THE HETEROGENEOUS EFFECT ON THE DEMAND SIDE I: THE FRACTION OF THE CITING ARTICLES AUTHORED BY SCHOLARS ALL FROM THE DEVELOPING COUNTRIES EACH YEAR

DV=The fraction of citing articles authored by scholars all from developing countries per year	(10-1)	(10-2)
SSRN	-0.009	-0.006
	(0.0054)	(0.0079)
SSRN x Post_SSRN	0.016*	0.025*
	(0.0076)	(0.0105)
SSRN x High_JIF		-0.007
		(0.0103)
SSRN x Post_SSRN x High_JIF		-0.023
		(0.0134)
Constant	0.652	0.651
	(0.3527)	(0.3524)
N of article-years	27617	27617
N of articles	2908	2908
N of SSRN article-year	3235	3235
N of SSRN articles	283	283
N of Journal/Vol/Is	264	264
N of Journal	131	133
Log-Likelihood	6648	6652

*** Significant at the 1 percent level. ** Significant at the 5 percent level.* Significant at the 10 percent level.

TABLE 11–THE HETEROGENEOUS EFFECT ON THE DEMAND SIDE OF CITATION II: THE FRACTION OF THE CITING AUTHORS FROM DEVELOPING COUNTRIES IN EACH CITING ARTICLE, AVERAGED OVER ALL CITING ARTICLES PER YEAR.

DV= Fraction of the developing countries of the citing authors' affiliations	(11-1) All articles	(11-2) All articles published in journals of JIF <= 1.92	(11-3) All articles published in journals of JIF > 1.92	(11-4) All articles
SSRN	-0.009	-0.012	-0.003	-0.015*
	(0.0050)	(0.0074)	(0.0057)	(0.0074)
SSRN x Post_SSRN	0.017**	0.025**	0.002	0.029**
	(0.0066)	(0.0097)	(0.0072)	(0.0094)
SSRN x High_JIF				0.014
				(0.0093)
SSRN x Post_SSRN x High_JIF				-0.031**
				(0.0116)
Constant	-0.171	-0.367	0.093	-0.169
	(0.1858)	(0.3324)	(0.1119)	(0.1861)
N of article-years	27073	12949	14075	27073
N of articles	2798	1550	1243	2798
N of SSRN article-year	3200	1951	1249	3200
N of SSRN articles	272	185	87	272
N of Journal/Vol/Is	247	174	73	244
N of Journal	126	102	18	123
Log-Likelihood	8421	2802	6002	8424

TABLE 12–THE HETEROGENEOUS EFFECT ON THE DEMAND SIDE OF CITATION III: CITATIONS WITHIN OR ACROSS DISCIPLINES FOR THE CITED ARTICLES PUBLISHED IN THE FIELD OF LAW, ECONOMICS, OR FINANCE.

	All citations per year (All articles) (12-1)	All citations per year (Articles only in the field of economics, finance, or law) (12-2)	Citations only by the articles in the same field per year (12-3)	Fraction of within-field citing (12-4)
SSRN	0.393**	0.307**	0.355**	0.017
	(0.1485)	(0.1119)	(0.1138)	(0.0151)
SSRN x Post_SSRN	0.047	-0.033	-0.002	0.020
	(0.0828)	(0.0878)	(0.0918)	(0.0132)
Constant				2.154***
				(0.3843)
N of article-years	27617	23316	23316	23316
N of articles	2908	2430	2430	2430
N of SSRN article-year	3235	2596	2596	2596
N of SSRN articles	283	225	225	225
N of Journal/Vol/Is	264	206	206	206
N of Journal	131	90	90	91
Log-Likelihood	-77083	-65029	-54945	-6925

Notes: Economics and finance are considered one field; law is the other field. If the cited article was published in the journals of economics/finance or law and the citing article was published in the journals in the same field, respectively, the citation was defined as a citation within the field.

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

*** Significant at the 1 percent level. ** Significant at the 5 percent level.* Significant at the 10 percent level.

TABLE 13–THE HETEROGENEOUS IMPACT ON THE DEMAND SIDE OF CITATION IV: GEOGRAPHIC DISTANCE BETWEEN THE REPRINT OR THE FIRST AUTHOR OF CITED ARTICLES AND THE REPRINT OR THE FIRST AUTHOR OF THEIR CITING ARTICLES, AVERAGED OVER EACH YEAR

	Geographic distance between cited and citing author (All articles) (13-1)	Geographic distance between cited and citing author (Articles published in journals with JIF<=1.92) (13-2)	Geographic distance between cited and citing author (Articles published in journals with JIF>1.92) (13-3)	Geographic distance between cited and citing author (All articles) (13-4)
SSRN	-234.5*	-359.8*	-58.6	-404.2**
	(99.2)	(140.2)	(137.4)	(138.4)
SSRN x Post_SSRN	105.5	116.6	15.2	210.5
	(119.3)	(176.1)	(156.7)	(172.0)
SSRN x High_JIF				366.8
				(194.0)
SSRN x Post_SSRN x High_JIF				-238.5
				(230.6)
Constant	3973.8	4081.8	1892.7	3871.4
	(4126.2)	(7002.9)	(4110.8)	(4075.6)
N of article-years	17253	7078	10175	17497
N of articles	1681	759	922	1725
N of SSRN article-year	2451	1429	1022	2451
N of SSRN articles	200	125	75	200
N of Journal/Vol/Is	153	89	59	149
N of Journal	70	50	16	74
Log-Likelihood	-161239	-66413	-94768	-163529

*** Significant at the 1 percent level. ** Significant at the 5 percent level.* Significant at the 10 percent level.