IT and Innovation: How Did the Internet Affect Firms' Reliance on Science?

(Preliminary draft)

Ruyu Chen

Stanford University

Abstract: This paper examines how the Internet facilitates the utilization of science in industrial innovation. I find that the Internet enables firms to discover "hidden gems" – commercializable yet under-recognized scientific findings that have been published by *early-career* scientists, in *less prestigious* journals, and/or with *fewer academic citations* but with *higher forward patent citations*. I compiled a database that contains 541,568 patent citations that refer to scientific papers; these citations came from patents applied between 1992 and 2000 by 3,651 public firm locations (firm sites in a given metropolitan statistical area). I then identified the staggered adoption of basic Internet at these firms. I show that access to the Internet at firm locations is associated with a 9.3% increase in the likelihood of citing scientific papers, and up to 13.2% increase in the likelihood of citing "hidden gem" papers. These findings suggest that IT reshapes the process that firms use to source knowledge in innovation. By reducing search costs, IT enables firms to access scientific knowledge that previously had been less visible – and to discover and capitalize on its commercial value.

Keywords: science, innovation, information technology, patent-to-paper citation

1. Introduction

Firms rely on academic scientific discoveries in their innovation process (Sorenson and Fleming, 2004; Arora et al., 2018; Bikard and Marx, 2020). Comparisons of patented inventions that cite science and those that do not have recently shown that patents directly citing the scientific literature have greater monetary value and novelty (Poege et al., 2019; Watzinger et al., 2021); an increased likelihood of renewal; and more forward patent citations (Ahmadpoor and Jones, 2017). Nevertheless, only a small fraction of scientific knowledge has been explored by firms, due to barriers in translating scientific discoveries into actual inventions (Ahmadpoor and Jones, 2017; Bikard, 2018; Marx and Hsu, 2021; Bikard and Marx, 2020). In addition, the explosion of knowledge in the digital era makes it increasingly difficult to identify the right piece of scientific work one seeks (Jones, 2009; Alberts, 2010). The high costs for firms to search, evaluate, and absorb the scientific literature (Nelson, 1959; 1982), and the associated costs of communication within the research teams involved in the innovation process may have slowed down the creation of science-based inventions. Information technology (IT) has the potential to reduce these costs and, thus, to bridge science and innovation.

Information technology is likely to increase firms' ability to find and source relevant science in the pursuit of innovation. However, it is less clear how IT influences the process. Does IT reinforce patterns of bias in existing access? Or does IT instead help to democratize the process? As a result of bounded rationality, firms might take notice of work largely undertaken by superstars, who publish in highly ranked journals and have more academic citations? Or, with keyword searches replacing manual browsing of curated journals, firms may instead come across scientific work that was more obscure prior to the advent of the quicker searching made possible by such technologies.

Existing literature in strategy, information systems, and economics provides competing predictions to these research questions. On the one hand, IT might amplify the Matthew effect of cumulative advantage by shining a spotlight on articles by "superstars" or work that is largely published in top journals. Bikard and Marx (2020) show that firms are more likely to use scientific papers with higher status – that is, those that published in hubs[1], have more academic citations, and those that are published in journals of higher impact factors. This finding could suggest that highly recognized scientific papers are more useful to firms, and that such papers are therefore expected to create the majority of the value in the industrial innovation process (i.e., the Pareto Principle or the power law). New IT users, especially those who have had little prior experience in citing science, are more likely to pay more attention to the publications that have achieved greater recognition. This could lead to a concentration in the utilization of the well-known publications, resulting in "superstar effects" (Rosen, 1981) and a "winner-take-all" phenomenon (Frank and Cook, 2010). On the other hand, IT instead may have the opposite effect: disproportionately increasing the citations of papers that are harder to find, due to the technologies' ability to lower search costs and to provide quicker, easier access to relatively obscure, lower-status papers available in the online knowledge base. Such research includes papers written by early-career scientists, work published in relatively low-impact-factor journals, and publications that have generated relatively fewer academic citations. For example, using the Social Science Research Network as a natural experiment, Kim (2013) found similar "long-tail" effects of open access on paper citations, with articles in lower-ranked journals gaining a boost in forward paper citations.

[1] Hubs refer to "a geographic concentration of patenting by firms in a specialized technical field. (Bikard and Marx, 2020)

This paper takes a first step toward answering these questions by empirically estimating how Internet adoption has shaped firms' reliance on science in the innovation process. I examine how access to basic Internet at firms affects the likelihood of citing the scientific publications in patent inventions, and, in particular, the heterogeneity of patent citations linked to high- and low-status scientific discoveries. To do so, I construct three measures for the academic status of each cited publication: (1) the career stage of the paper's author(s), (2) the journal impact factor at the time of the article's publication, and (3) the citation count by other academic papers. I merge several sources of data: a private technology-adoption dataset, Harte Hanks Market Intelligence Computer Intelligence Technology database (CI data); the recently assembled Patent Citation to Science (PCS) dataset; data from PatentsView, and, for firms' R&D expenditure, the Compustat dataset. I then construct a unique sample that consists of a balanced panel of 3,651 publicly listed firm locations (firm sites within a metropolitan statistical area (MSA)) that have at least one patent application over the period from 1992 to 2000. The commercialization of the Internet in late 1995 means that this period offers an opportune window of time to measure the staggered adoption of the Internet at firms, and to capture early diffusion of this technology. I identify 541,568 patent citations, which refer to 167,882 scientific papers; these citations were referenced in 192,856 patent applications successfully granted to firms in the sample over the period.

With these data, I provide two sets of findings. First, Internet adoption is associated with an increase in using science and, in particular, in using the less recognized scientific discoveries in the academic literature. Articles by early-career scientists, articles published in less prestigious journals, and articles with fewer academic citations are consistently more likely to be cited by patent inventors in Internet-adopting firms. The percentage changes in the likelihood of citation range from 8.1% to 13.2%. By contrast, firms' access to the Internet has no significant impact on

patent citation that refer to notable papers – those published by senior scientist(s), those published in top-ranked journals, and/or those with a large number of academic citations.

Second, Internet adoption at firms is associated with an increase in citing commercially valuable papers that have more forward patent citations. I use the forward patent citation count to a focal paper as a proxy for a paper's commercial impact. This could not easily have been observed by the public during the sample period, and, thus, the inventors were unaware of the commercial value of any cited paper during the period in which they worked to gain a patent. The two sets of results suggest that IT enables firms to discover "hidden gems" – commercializable science that had been less recognized in the academic literature. These findings shed light on the link between science and innovation, and how this link could be reinforced by IT.

Several recent studies have explored characteristics that affect firms' attention allocation to scientific literature, for example, research funding sources (Fleming et al, 2019), institutional origin (Bikard, 2018), and location (Bikard and Marx, 2020). The use of IT aligns with these studies but also extends them. By leveraging IT adoption as a treatment that reduces the costs for firms to cite science, this paper contributes to the literature by theorizing and providing empirical evidence on why firms pay attention to low-status scientific literature. The findings can provide important strategic implications for firms to identify and exploit useful scientific knowledge and to gain competitive advantage (Bikard and Marx, 2020).

This paper makes several contributions. First, it contributes to the literature on IT in innovation. As demonstrated by prior literature, IT can play many roles in innovation, including providing the inputs for innovation (Kleis et al., 2012, Ravichandran et al., 2017), complementing other innovation inputs to achieve greater returns (Joshi et al., 2010; Saldanha et

al., 2017); or serving as an effective search tool that facilitates recombination of knowledge (Wu et al., 2019;2020; Lou and Wu, 2021; Saldanha et al., 2021). Several papers examine the link between IT and knowledge flows. For example, early IT such as the cooperative university network BITNET has been shown to have increased coauthored works among scientists, and to have facilitated knowledge flows within academia (Agrawal and Goldfarb, 2008; Ding et al., 2010); the Internet has been shown to foster R&D collaboration within firms (Forman and van Zeebroeck, 2012; 2019). While there is a rising awareness that knowledge spillovers from academia to industry affect innovation performance (Arora et al., 2018; Marx and Fuegi, 2020), no direct evidence has emerged to show how IT can affect such knowledge flows. This paper fills the gap by highlighting the role of the Internet in bridging science and industrial innovation.

Second, this paper contributes to the literature in search-based innovation. I examine how IT tools can increase the breadth of searches by facilitating firms' utilization of scientific knowledge in the product innovation process. This finding adds value to the discussion on search breadth and innovation performance (Scott and Brown, 1999; Brown and Duguid, 2000; Leiponen and Helfat, 2011). In addition, some recent studies focus on scientific knowledge as one important component in invention portfolios (Ahmadpoor and Jones, 2017; Marx and Fuegi, 2020). One unsolved question from this line of the literature concerns which part of science firms build upon from a large knowledge stock. Empirical evidence partially unpacked this question by showing that government-funded research (Fleming et al., 2019), university research (Bikard, 2018), and research produced in geographic hubs (Bikard and Marx, 2020) are more likely to be cited by inventing firms. I contribute to this literature by providing empirical evidence of how IT can increase firms' utilization of scientific knowledge in innovation.

Third, I speak to the literature on the "long-tail" effect of the Internet in consumer goods, and provide different theoretical foundations to explain such long-tail effects of the Internet on scientific knowledge exploration. The long-tail literature in consumer goods suggest that IT leads to an increase in sales of niche products and to greater product variety due to lower search costs, more efficient inventory, and a better match between the demand side and supply side (Bakos, 1997; Anderson, 2006; Brynjolfsson et al., 2006, 2011; Fleder and Hosanagar, 2009; Peltier and Moreau, 2012). Scientific knowledge, as an intermediate input in firms' innovation function, is fundamentally different from consumer goods. Using the Social Science Research Network as an experiment, Kim (2013) found long-tail effects of open access on paper citations, with lower-ranked articles receiving a boost in forward paper citations after posting on SSRN. I contribute to this literature by theorizing and empirically demonstrating why IT can disproportionately increase the citations of scientific papers that are harder to find.

The rest of this paper is organized as follows: In Section 2, I provide the conceptual background of using the extent of scientific citations in patents as a measure of scientific input into innovation; and I formulate the hypothesis on how the Internet shapes firms' reliance on science. Section 3 describes the main data sources. Section 4 presents the empirical setting of how the Internet affects the utilization of science in patenting activities. Section 5 discusses the empirical results. Section 6 concludes.

2. Theoretical Framework

In this section, I formulate my hypothesis on how the adoption of the Internet at firms can lead to an increase in utilizing low-status scientific discoveries in innovation. I begin by describing science as an input in patent inventions, and I then analyze the frictions for firms to search, evaluate, and utilize science. I provide historical background to show how these frictions have

changed over time, and I then I theorize how and why IT can increase the utilization of low-status papers. I then predict the commercial impact of cited papers discovered by the Internet.

2.1 Scientific knowledge as an input in firm R&D

A wealth of research demonstrates the ability to exploit external knowledge as a critical component in the innovative activities that firms take to improve performance (Cohen and Levinthal, 1990; Laursen and Salter, 2006; Berchicci, 2012). Firms rely on a variety of channels to search for innovative opportunities (Scott and Brown, 1999; Laursen and Salter, 2006; Leiponen and Helfat, 2009), and each channel involves institutional norms, habits, and rules (Brown and Duguid, 2000). Among the different knowledge sources, scientific research from universities, research centers, and the private sector has proven to be a fundamental resource that plays an increasing role in industrial innovation. Science as a "map" can lead firm inventors directly to useful combinations (Fleming and Sorenson, 2004). More specifically, science enables firm inventors to generate hypotheses worth exploring, rule out paths that lack promise, provide tools to speed development, suggest techniques to aid laboratory or statistical work, and create basic pieces of scientific knowledge for recombination (Mokyr, 2005; Murray and Stern, 2007).

2.2 The Internet reduces search and transfer costs of scientific knowledge

Historically, the search for scientific knowledge has proved costly to business. For decades, information-retrieval systems designed in the 1960s were used by reference librarians, patent attorneys, and other specialized professionals trained to search the collections of documents. The advent of the Internet, which became available to the business sector beginning in late 1995[2],

[2] The arrival of the World Wide Web in the early 1990s fundamentally changed the way that knowledge was stored, but also brought complex challenges in search, such as ranking, multiple meanings, and problems related to the constantly changing nature of information and knowledge (Easley and Kleinberg, 2010). However, the Internet had not been privatized at that time. Due to institutional and legal constraints, most commercial firms had no access to

exposed the high search costs for firms to access academic science. At that time, firms had two major acquisition channels: physical libraries and subscriptions to print journals. Firm inventors would visit a local library to find relevant materials on the bookshelves, or to consult with the librarians, who would use a special retrieval language to conduct keyword search in the database on (localized) computers. A search would include digitized contents of printed publications and the print books and journals stored on the bookshelves of the libraries[3]. Subscriptions to academic journals were available for a few firms with high R&D efforts. These firms generally had subscriptions to only a limited selection of journals, and those journals were available almost exclusively in print, with some journals beginning to be digitized[4]. These two channels suggest the high search costs for firms seeking to source science in their work.

In addition to the costs of searches, the absence of advanced IT technologies created high communication barriers for collaboration with universities, and for participation in academic activities (e.g., conferences), resulting in limited involvement of firms with academia (Agrawal and Henderson, 2002; Murray, 2002; Bikard and Marx, 2020). This further impeded knowledge flows from academia to industry. Such significant barriers made it difficult for research to capture the level of inventor attention required to foster the transfer of scientific knowledge into actual inventions (Koput, 1997; Ocasio, 1997).

---

the Internet until late 1995 (Greenstein, 2015). Only in very rare circumstances (such as for research purposes) did commercial firms have access to the Internet prior to 1995.

[3] The digitization of journal articles in the United States can be traced back as far as the 1960s. The National Library of Medicine started to digitize selected articles in the field of medicine with an index ("Index Medicus"), and to provide access to the content of this index via the MEDLINE system, which was based on specialized telecommunications systems. Project Gutenberg, founded in 1971, was the first digital library that focused on digitizing full texts of books. These digitized articles were relatively easy to find in the databases of localized computers, though the querying system is different from a present-day web search. (https://www.nlm.nih.gov/medline/medline_history.html and telephone interview)

[4] For example, the JSTOR, which stands for Journal Storage, was founded in 1955, becoming one the early digital libraries. By the end of 1996 it had collected and digitized 16 journals (including the American Economic Journal, Ecology, Econometrica, and The Journal of Modern History). By the end of 2000 JSTOR had digitized 224 journals. Source: author's interview with JSTOR Content Team.

Thus, the arrival of the Internet brought two disruptive and fundamental changes to the ways in which firms source science. First, connecting to the Internet reduces the costs of knowledge search and lowers the access barriers to academic science. Individual inventors can now conduct direct keyword searches, browse the Web , and access bibliographic information of the relevant scientific articles they find. The Internet also improves knowledge transfer. Inventors can evaluate and absorb the scientific knowledge from selected articles at a lower cost (Alavi and Leidner, 2001). Therefore, the two mechanisms indicate that the use of the Internet increases firms' reliance on science. This implies the baseline prediction:

H1: Other things equal, Internet adoption will be associated with an increase in the likelihood of citing scientific papers in patent inventions.

2.3 The Internet increases utilization of low-status academic scientific papers

Since at least the work of March (1991), organizational scholars have distinguished between "exploitation" and "exploration" in strategic formulation. A key to establishing sustainable competitive advantage is not merely harvesting legacy investments (exploitation) but coming up with new inventions (exploration) that can eventually be brought to market as new products and services. Evidence has shown that IT can facilitate knowledge exploration. In some cases, IT can facilitate firms' exploration of knowledge about areas of expertise that are unfamiliar to the firm, or of information from geographically remote areas, leading to novel recombinations of ideas (Offsey, 1997; Uzzi el al., 2013; Forman and van Zeebroeck, 2018; Wu et al., 2019; Zheng and Wang, 2020). In other cases, IT can reinforce individuals' exploitation of knowledge within a familiar base (Rosenblat and Mobius, 2004; Van Alstyne and Brynjolfsson, 1996, 2005). Yet little has been studied about how IT affects utilization based on the status of knowledge – that is,

whether IT primarily leads firms to use well-known knowledge, or to increase the use of under-recognized knowledge. In this section I theorize that the Internet can disproportionately increase firms' utilization to *low-status* scientific publications, and I offer insights into why this is the case.

The use of the Internet and the knowledge explosion associated with the Internet revolution have tremendously increased the number of scientific papers easily available to firms with the online knowledge base. This increase is disproportionately larger for papers that are harder to access in a physical library. Such lower-status papers include those that are published by early-career scientists, those published in lower-ranking journals, and those with fewer academic citations. As a result, access to these "long-tail" papers increased to a disproportionately greater degree than access to the most prominent papers.

The technological advancement in Web searches associated with the Internet also increases firms' attention allocation to long-tail papers. Web searches enable inventors to perform nonspecific searches by typing a single keyword or a few different keywords to find the digitized content of any scientific papers on popular digital platforms. In the 1990s, such platforms included Yahoo!, Excite and AltaVista, for example. Such searching on the World Wide Web (Berners-Lee et al., 1994) refers to "the process of discovering pages that are relevant to a given query" (Kleinberg and Easley, 2010). The ability to conduct Web searches fundamentally improved upon the traditional, offline database-querying systems that had been used in physical libraries[5]. For example, if an inventor types two keywords that are from two very distant areas, the most relevant content is likely to surface via higher rankings generated by

[5] See the discussion on *relevance, popularity*, and distillation of broad search topics through the discovery of "authoritative" information sources on the Web.

Web page-ranking algorithms[6]. In this way, less recognized papers in the academic literature are more likely to capture firms' attention on the Internet through a more effective page ranking than through a library database query. In addition, behavioral changes associated with Web searches can also increase the demand for articles that are harder to find. For example, individual inventors can try different combinations of relevant keywords in multiple ways as part of the search process. This was not possible through traditional library searches. An improved combination of keywords increases the visibility and accessibility of the most desired and most relevant scientific discoveries, which were harder to find prior to the advent of such search capacities (Pant and Srinivasan, 2010).

From the inventors' learning perspective, the Internet reduces the uncertainty in knowledge exploration to less-known discoveries. Information overload and the high cost to find papers make it impossible for any firm to search the entire knowledge space for commercializable science. As a result, firm inventors usually rely on the status of journals or scientists to source science. For example, in the process of knowledge exploration, an article published in "Nature" or "Science" is easier to find and considered to be less uncertain, with more predictable learning outcomes, compared to papers in lower-ranked journals. The availability of Internet tools greatly reduces knowledge search and transfer costs, and thus enables firms to explore these long-tail scientific discoveries with reduced uncertainty. In this way, the Internet increases firms' demand for long-tail scientific discoveries.

With the discussions above, the second hypothesis is:

[6] The most influential two algorithms include (1) Hypertext-Induced Topic Search or HITS by Jon Kleinberg, and (2) PageRank developed in 1998 by Google's founders Sergey Brin and Larry Page.

H2: Other things equal, Internet adoption will be associated with a larger increase in the likelihood that patents cite papers with low academic status, defined as those papers by early-career scientists, those published in lower-ranked journals, and/or those with fewer academic citations.

2.4 The Internet increases utilization of scientific publications with high commercial impact

Innovation processes involve searching for new ideas that have commercial potential (Laursen and Salter, 2006). Yet the commercial value of published papers can be under-recognized in the academic literature due to various reasons. This section discusses how IT enables firms to discover commercializable science.

Papers with high creatively and novelty. Highly creative and novel work resulting from unique or atypical combinations of prior knowledge (Uzzi et al., 2013) can bring valuable scientific inspirations to firms. Yet some of the papers that contain this type of scientific work may be harder to find, and they may be under-recognized in the academic literature. First, a fresh concept may take a long time to assimilate because few conventional narratives, languages, or cultures fit it (Wang et al., 2016; Cetina, 2009). Second, the academic community may take a long time to recognize the validity of highly novel ideas, and, in turn, to promote their value(van Raan, 2004; Bornmann and Daniel, 2008). Third, highly creative work may face a high risk of failing the peer-review process (Mobley et al., 1992; Estes and Ward, 2002; Foster et al., 2015). For these reasons, a commercially useful, novel work may end up being published in journals that have lower impact. As discussed in Section 2.3, these papers may be harder to find without effective IT tools.

Similarity, creative and novel work may not be identified through *forward paper citation count*. Prior literature has shown that paper citation counts and novelty can sometimes have negative correlations. Wagner et al. (2019) found that the highly cited international, collaborative articles are less novel and with fewer atypical combinations of conventionality. Papers with high levels of forward academic citations reflect an audience effect and the preferential attachment based on reputation, in which authors from more countries get access to the article. Though little evidence is available on the reverse relationship (between forward paper citation counts and a novel paper), the findings of previous research suggests that many factors other than the usefulness of a paper can account for the forward academic citations. As a result, one would expect that inventive firms with access to Internet tools would also have greater access to novel papers, even if these papers have relatively fewer academic citations. Using a similar rationale, I discuss three additional sets of publications of high commercial value that can be discovered by Internet searching.

Papers with a narrow and specific topic. Academic papers with a narrow topic that is not targeted to top scientific journals can be a useful resource to firms. These papers provide useful details that can, for example, aid laboratory or statistical work, and suggest effective tools to speed development (Mokyr, 2005; Murray and Stern, 2007).

Conference papers. Conference papers and non-peer-reviewed publications can be another useful source for emerging knowledge. In certain fields, academic literature start circulating via conferences prior to the actual publication, and given that knowledge and science is moving forward rapidly, this could be a place to stay abreast of the research frontier, which is always moving. For example, the appearance of "digital science" in the case of complex innovation requires firms to quickly absorb the knowledge to "measure, analyze, and model

chemical compounds, diseases, and human biology" (Dougherty and Dunne, 2012). Academic conferences could be one of the best sources for such emerging knowledge.

Other under-recognized works in the academic literature. In general, the sociology of science is a communication and exchange of research findings and results; publication serves to generate professional recognition and esteem, promotion, advancement, and funding for future research for the authors (Fox, 1983). Therefore, the institutional design of publication suggests many other potential cases in which a work of high commercial value may be under-recognized in the academic literature. Merton (1968) pointed out that young cohorts of scientists with limited reputation and less cumulative advantage are more likely to be undervalued in the peer-review process. Thus, the publications by such early-career researchers are another potential set of research that may be commercially valuable but may at the same time be under-recognized in terms of their academic value. I formalize the content of these discussions into the third hypothesis:

H3: Other things equal, Internet adoption will be associated with a larger increase in the likelihood that patents cite papers with higher commercial impact.

3. Data and key variables

I use a variety of data sources to identify the effects of Internet adoption on patent citations among commercial firms. I match data on firm IT adoption from a private technology-adoption dataset (CI data) to data in the PatentsView patent dataset and in the newly assembled Patent Citation to Science (PCS) dataset (Marx and Fuegi, 2020). I also obtain information from Compustat on firm R&D expenditures as additional control. I estimate the model from 1992 to 2000, a period of time that captures the early diffusion of the Internet technology.

3.1 Data

3.1.1 Patent citations to science data

My main dependent variable, the extent of firms' reliance on science, is measured by the patent citation to scientific papers using the newly assembled, large scale, open-source Patent Citation to Science dataset (PCS) (see Marx and Fuegi (2020), available on relianceonscience.org). It extracts the references to scientific articles on both the front-page and body-text of patent documents granted by U.S. Patent and Trademark Office and European Patent Office. Based on a combination of machine-learning and heuristic-based rules, the algorithm enables the unformatted body-text citations to be identified, and it achieves a high precision rate[7]. While the front-page citation serves as a legal purpose to disclose prior art, the body-text citations are not legally binding, and, thus, they are believed to be a more accurate proxy for the actual scientific inspiration for inventors. They are also more diverse temporally, geographically, and topically than the front-page citations (Marx and Fuegi, 2020).

The PCS dataset links each paper to Microsoft Academic Graph (MAG), which collects over 160 million papers published since 1800. Therefore, a variety of detailed information for each cited paper can be identified, including journal name and impact factor; author name and affiliation; and fields of study[8]. I rely on these paper and author characteristics to define the status of cited papers in the academic literature.

References to non-patent literature and references to prior patents are the two primary forms of "prior art" in patent inventions. While references to prior patents have been studied

[7] A third-party assessment shows that the algorithm can capture up to 93% of patent citations to science with an accuracy rate of 99% or higher.

[8] The journal impact factor is calculated for all journals in MAG. In this paper I used six fields of study, taken from the Organization of Economic Co-operation and Development (OECD), and mapped from MAG subjects. The PCS data also provide 39 OECD subfields, Web of Science fields, and more than 200,000 fields automatically extracted from the papers.

intensively in innovation literature for decades (e.g., Trajtenberg, 1990; Jaffe et al., 1993; Hall et al., 2001; 2005; Fleming et al., 2007), very little research has studied the citations from patents to non-patent literature due to the costly data construction (Fleming and Sorenson, 2004; Katila and Ahuja, 2002; Gittelman and Kogut, 2003; Fleming et al., 2019). Moreover, these studies include only front-page citations, and they are mostly limited to a single industry or to a small number of firms. The PCS dataset enables me to conduct multiple-industry analysis with comprehensive citations from both front pages and the main texts.

3.1.2 IT data

I rely on the Harte-Hanks Market Intelligence Computer Intelligence Technology database (hereafter, CI database) to measure Internet adoption. The CI data contain information on establishment characteristics, such as the installations of IT software and hardware, the number of employees at site, and industrial classifications. As one of the most comprehensive sources of micro-level IT investment, this dataset has been used by many researchers to study the adoption and economic implications of IT investments (Bloom et al., 2012; Bresnahan et al., 2002; Bresnahan and Greenstein ,1996; Forman et al., 2005, 2012; Nagle, 2019).

I focus on multiple industries within the manufacturing sector (North American Industry Classification System (NAICS) codes 31-33) because the patenting purpose and activities are more similar within this sector than in others. I exclude smaller establishments with fewer than 100 employees to prevent the potential measurement error as demonstrated in prior literature that used establishment-level CI data (e.g., Forman et al. 2005, 2008, 2012). My sample period extends from 1992 to 2000 to capture the early diffusion of commercial Internet, which started in the second half of 1995. Due to data constraints, I use every other year data, and I set the adoption rate to be zero in 1992 and 1994.

The unit of analysis in this paper is a firm-MSA year. A firm-MSA represents an aggregation of establishments in a focal MSA that belong to the same public firm. It can be regarded as a "plant" of the firm. In the sample, over half of firms contain multiple establishments in a given MSA. I merge the CI data to the Compustat data using the crosswalk between the CI establishment identifier and the Compustat public-firm identifier, the Global Company Key (GVKEY), a unique, six-digit code assigned to each company. In many cases, a firm has establishments in several locations. I use the MSA as the geographical unit, and I match each establishment to an MSA using its county FIPS code. I aggregate all the establishments that belong to the same firm in a focal MSA into a larger unit (i.e., firm MSA). This larger unit can better account for the commuting patterns of the inventors.

3.1.3 Patent data

As the PCS dataset only contains patents that have at least one scientific citation, I use the PatentsView dataset to identify all patent applications by my sample firms that were filed at and U.S. Patents and Trademarks Office (USPTO) over the period from 1992 to 2000, and were successfully granted eventually. The crosswalk between firm GVKEY and patent identification is provided by Autor et al. (2020). I also divided the patents into six different technological classifications using the definition from Hall, Jaffe, and Trajtenberg (2001) (HJT category): Chemical; Computers & Communication; Drugs & Medical; Electrical & Electronic, Mechanical; and Other. Because each patent reports the county FIPS code of all inventors, I aggregate the inventor addresses at the MSA level to merge to the firm-MSA data.

3.1.4 Other data sources

I also collect data from other sources to control key factors that may affect firms' patent-to-science citations. I collect the R&D expenditures at firms from the Compustat data. To deal

with the missing values, I adopt a similar approach to Hall and Oriani (2006) and Simeth and Cincera (2016), assuming a growth rate of R\&D stock, and simulating the missing observations for firms that reported in selected years. For the manufacturing sector I use a growth rate of 6\%, a proxy for the average R\&D expenditure growth rate calculated from reported firms. The details on recovering missing R\&D data are described in Appendix C. Because each observation in Compustat data is at public-firm level, I normalize the per-location spending using the number of firm-MSA locations in the CI data.

3.2 Key Variables

With the four datasets, my final sample contains 541,568 patent citations to 167,882 unique scientific papers. These patent applications were successfully submitted from 3,651 large, public-firm MSAs between 1992 and 2000. Because I use every-other-year data, the number of observations is 18, 225. The main dependent variable and independent variables are described as follows:

3.2.1 Dependent variable

Incidence of patent citation to science. The interest is to understand the implications of IT adoption on firms' citation to science to scientific papers in innovation process. I use the incidence of patent to science citation at firm-MSA-year as the key dependent variable. Overall, over 28.1% observations in the entire sample report a patent citation to at least one scientific paper (Table 1). Specifically, the main contribution of this paper is to analyze the heterogeneity between patent citation to high status papers and low status papers. Thus, I divide the 167,882 cited papers into high status and low status groups according to their authors' career stages, forward paper citations, journal impact factors, and forward patent citations. The incidences of citation to these different groups of papers are reported in Table 1.

3.2.2 Independent variable

Internet adoption. I consider the establishment to have adopted basic Internet if it reports one of the followings in the CI data: (1) an Internet Service Provider (ISP); (2) internal intranet based on the TCP/IP protocol (Transmission Control Protocol/Internet Protocol); (3) TCP/IP based email; or (4) having used the Internet for research purposes. These investments are basic Internet access, which was technologically mature and required little complementary investment and adaptation in the business process by organizations. Therefore, it allows me to focus on the short-term changes of firms' propensity to cite science in innovation in response to the adoption of new technology tools. In aggregating the establishments into firm-MSAs, I took the maximum value and consider the adoption to be true if at least one of the establishments has connected to the Internet. Among the sample firm-MSAs, none had adopted the Internet in 1992 and 1994, 30% had adopted by 1996, and up to 93% had adopted by 2000.

Other controls include firm-MSA characteristics, such as number of employees, log of patent applications in current period, log of R&D expenditures, dummy for existence of different HJT category patents; as well as local characteristics, such as log of patent applications in the focal MSA. Table 1 provides a summary statistic of the key variables.

4. Empirical Framework

4.1. Effects of Internet adoption on patent to paper citation

I exploit the variations in patent-to-science citations in firms with and without Internet in period before and after the adoption to compare how basic Internet adoption influences firm citing science in their innovation process. The unit of analysis is a firm-MSA-year. I focus on a linear model with fixed effects in the baseline analysis to document the underlying relationships

between internet adoption and incidence of scientific citation. The baseline difference in differences framework is as follows:

$$ScienceCitation_{ijt} = \alpha_0 + \alpha_1 X_{ijt} + \alpha_2 Z_{jt} + \beta Internet_{ijt} + \sum_{h=1}^{6} \mathrm{D}(Patent_h) + \mu_{ij} + \tau_t + \varepsilon_{ijt}, \quad (1)$$

where $ScienceCitation_{ijt}$ is the outcome variable such as dummy for whether firm i from MSA j has cited at least one scientific paper in all its patent applications in year t. $Internet_{ijt}$ is a dummy that equals one if firm i from MSA j has adopted internet by year t. X_{ijt} is a vector of time-varying controls at the firm-MSA level, including log number of patents applications in the current period, log number of employees, log of R&D expenditure, etc. Z_{jt} is a vector of time-varying local characteristics, including the log number of patent applications at the MSA level to control for the innovation capability in local area, log population and log GDP per capita to control for economic factors that may affect firm innovation. $Patent_h$ is a dummy for whether the firm has at least one patent application that belongs to HJT tech category (1-6) in year t. μ_{ij} and τ_t are firm-MSA fixed effects and time fixed effects, respectively, and ε_{ijt} is an idiosyncratic error term.

4.2. Event studies

I conduct an event study to show the parallel trends between Internet adopting firms and their peer firms who have no access to the Internet for pre-adopting periods, as the following:

$$ScienceCitation_{ijt} = \alpha_0 + \sum_{k=-3}^{2} \beta_k \, 1(PeriodSinceInternet_{ijt}^{k} = k) + \alpha_1 X_{ijt} + \alpha_2 Z_{jt} + \mu_{ij} + \tau_t + \varepsilon_{ijt} \quad (2)$$

The event window ranges from three periods before adoption (i.e., $k = -3, -2, -1$) and three periods after the adoption of the Internet (i.e., $k = 0, 1, 2$) for any given firm-MSA. The control

variables include time fixed effects, firm-MSA fixed effects, and time-varying firm characteristics and local controls. The baseline is the period before Internet adoption (i.e., $k = -1$).

5. Results

I first establish a relationship between Internet adoption at firms and the incidence of patent citation to science. Then I examine the heterogeneities of cited scientific papers using four paper characteristics: career stage of all authors and the first author, journal impact factor when the paper was published, forward academic citations, and forward patent citations. As discussed in section 2, the first three are used to measure the status of cited paper in the academic literature, and the last one is a measure for its commercial impact. Finally, I explore robustness with respect to sample and specification.

5.1 Baseline Results

My baseline regression presents a strong and positive correlation between Internet adoption and firms' usage of science. The results are reported in Table 2. Column (1) shows that Internet adoption is associated with a 2.6 percent point increase in the likelihood of citing scientific papers in patent applications, statistically significant at the 5% level. Comparing the average patent citation rate to science of 28.1% in the sample, this translates to a 9.3 percentage increase in the likelihood of observing a patent to paper citation. To control for the effects of the Internret on patenting propensity, I limit the sample into regular patenting firms, i.e., firm-MSAs that have patent applications in at least four periods between 1992 and 2000. The results remain

robust and are reported in Column (2). Column (3) presents the sub-sample regression results where I exclude the top 2% largest sized firm-MSAs (with above 9,000 employees), because the largest firms may be insulated in ways due to the outlier size. I further demonstrate robustness by excluding the top 5%, 10%, and top 25% sized firm-MSAs.

Figure 1a plots the coefficients estimates \(\beta_{k}\) for the event study in equation (2). The trends between Internet adopting firm-MSAs and their non-Internet adopting peers are similar prior to adoption. The coefficients of the event times \(k = -3, -2, -1\) are small and insignificant. In contrast, the coefficients increase sharply after the adoption of Internet, and remain consistent in the three post adoption periods. Figure 1b shows the event study result without the 2$\%$ sized firm-MSAs (sample in Column (3) of Table 2). By excluding the largest samples, the pre-trend becomes closer to zero.

5.2 Heterogeneity analysis by status of cited papers in the academic literature

In this session, I divide the cited papers into high status group and low status group in the academic literature according to three key features, i.e., career stage of scientist(s), journal impact factor, and forward academic citations. I rank all those papers in the baseline; I also conduct robustness checks where I rank the papers within each OECD field of studies to account for the differences across fields.

5.2.1 Career stage of scientist

The first measurement is constructed using author-level information. I obtained the career-long publication records of all authors in the cited articles from Microsoft Academic Graph. If a paper is published in the first three years of an author's publication record, then I

define it as an early-career work of this scientist. I construct two sets of early-career papers: when all authors are in their early career, and if the first author is in his/her early-career.

The regression result of the effects of Internet adoption on firms' patent citation to early-career work is presented in Table 3. Column (1) shows that Internet adoption at firms is associated with a positive 2.5 percent point increase in the likelihood of citing young scientists' work (significant at 5% level), which translates into a 10.4 percentage increase compared to a 24% citation rate. Column (2) presents a stronger effect of Internet adoption when only the first author is in his/her early career stage, with a 3.1 percent point increase and a 12.1 percent increase in the likelihood of patent citation by Internet adopting firms, significant at 1% level. Column (3) and (4) explore the effect of Internet adoption on patent citation to papers with all authors being senior scientist(s), and the first author being senior, respectively. The results are not economically or statistically significant.

Figure 3 plots the coefficients estimates for patent citation to papers written by early-stage scientist(s) and senior scientist(s), respectively. The trends between Internet adopting firm-MSAs and their non-Internet adopting peers are similar prior to adoption in both groups. The coefficients of the three pre-adopting periods are small and not significant. In post-adoption periods, the coefficients increase sharply for patent citation to papers by early-career scientist(s) and remain consistent. However, the effects of patent citation to papers by senior scientist(s) experience no significant changes in post adoption periods.

5.2.2 Journal impact factor

The second measurement, the journal impact factor (JIF) is one important index to measure the academic impact of a journal. It is calculated by dividing the number of paper citations in a given journal by the total number of articles published in the previous two years. I

rank all sample papers based on the JIF at the time of publication. I use top 25%, top 50%, bottom 25% and bottom 50% ranked papers to represent scientific works of different levels of academic impact, respectively. The results are reported in Table 4.Column (1) shows Internet adoption at firms has a 3.0 percentage point effect on the incidence of patent citation to the bottom 25% papers, which is significant at the 1% level. This translates into a 13.2% increase in the likelihood of citing a lowest ranked paper. Column (2) shows that the correlation between Internet adoption and patent citation to the bottom 50% papers are slightly smaller in terms of percentage point (2.6) and percentage change (9.6%), and is significant at the 5% level. In contrast, columns (3) and (4) show that the effect of Internet adoption on patent citations to academically more impactful papers are much less significant, in particular, negative for the top 25% papers. The coefficients of Internet adoption are neither statistically nor economically significant.

Figure 4 shows that the incidence of patent citation to scientific papers in bottom 25% impact factor journals increases sharply after the adoption of Internet, and that increase last for three periods after the Internet adoption. In contrast, the incidence of patent citation to papers in top journals show no apparent trend- it decreases slightly in the first two periods after Internet adoption and shows a small increase three years after adoption. For both groups, there are no pre-trends before access to the Internet.

5.2.3 Forward academic citations

The third measure for a paper's academic impact is the forward academic citation count. I sum up the number of citations received by each paper till the end of 2019 and create quartiles. Column (1) and (2) in Table 5 show that Internet adoption has no statistically or economically significant effect on patent citation to papers with top 25% or top 10% academic citation count.

In contrast, column (3) suggests that patent citation to the bottom 39% papers with zero academic citation count is positively correlated with Internet adoption. These papers experience a 2.0 percentage point increase in the likelihood to be cited by Internet adopting firms (which translates into an 8.1% increase compared with a citation rate of 24.56%, and statistically significant at the 10% level). The results suggest that scientific articles with few academic citations, which have little impact in academia, turn out to be a useful source of knowledge to innovative firms if the access barrier to these discoveries is mitigated or removed by advanced IT tools such as the Internet. The event study analysis as plotted in Figure 5 show no pretrends before the adoption.

5.3 Heterogeneity analysis by the commercial impact of cited papers

5.3.1 Forward patent citations

I use forward patent citations received by each paper as a proxy for the commercial impact of any cited papers. The underlying hypothesis is that the Internet enables firms to find useful scientific discoveries that contribute to the innovation process. These papers may receive more forward patent citations by other firms, and in return, they have higher revealed commercial value. To check this hypothesis, I use the entire PCS dataset to sum up the total forward patent citation count received by each cited paper till the end of 2018. As illustrated in Figure 2, the sample papers (all published before 2000) have continuously been cited by USPTO patents until the latest year in the PCS dataset.

To check the hypothesis that Internet adoption is associated with an increase in patent citation to papers with higher commercial value, I divide all papers into quartiles based on their total patent citations. Then I construct the firm-MSA level measure as the dependent variable: incidence of patent citation to papers in the top quartile(s) and in the bottom quartile(s). The

results are reported in Table 6. As shown in Column (1) and (2), Internet adoption is associated with a positive and significant increase in the likelihood of citing papers with top patent citation counts. For the top 10% papers, the percent point is 1.4, which translates into a 12.2% increase and statistically significant at the 5% level. The top 25% papers experience a similar increase in the likelihood of being cited by Internet adopting firms, with a 2.0 percent point and a 10.7% increase, significant at the 5% level. In contrast, the effect of Internet adoption on citing papers with fewer patent citations is not statistically different from zero. The results suggest that commercially valuable papers are more likely to be discovered by firms with the Internet.

5.4 Robustness checks

5.4.1 Placebo Test

I conduct a placebo test using randomly assigned Internet adoption years for all firm-MSAs. The null hypothesis is that Internet adopters and non-Internet adopters should not be significantly different using placebo adopting years. I random assign an Internet adoption year between 1996 and 2000 for each firm-MSA in the sample and conduct the estimate of equation (1) for 500 times using false adopting years. The results are plotted in Figure 6. The placebo effects center around zero, and the observed effect size (9.3%) lies outside of the 99% confidence interval of the distribution of coefficients from 500 placebo tests. The results suggest that a false adoption time of the Internet is not associated with an increase in the likelihood of patent citation to science in Internet adopting firm-MSAs.

5.4.2 Journal Impact Factor by OECD fields of study

The heterogeneity across field of studies may affect forward paper citations and journal impact factor. I account for the cross-field difference by ranking the sample papers within its

OECD field of studies[9]. I use the six categories, i.e., natural science, engineering and technology, medical and health sciences, agricultural science, social science, and humanities. As shown in Figure A2, natural science and medical and health sciences journals have higher impact factors on average, and the factors are increasing slightly across time. Table A2 reports the heterogeneity analysis by journal impact factor with adjustments to OECD fields, and the results remain robust to the baseline as in Table 4.

5.4.3 Controlling for supply of scientific knowledge

The internet revolution is associated with a knowledge explosion, which can increase the *supply* of scientific knowledge and affect the propensity to cite science at firms. To control for the supply of knowledge, I refine my analysis to a subset of sample papers that were published before 1989. The underlying assumption is that firms' exploitation to older knowledge base has little correlation with the knowledge boom during the Internet age. The regression results of Internet adoption on the likelihood of citing papers published before 1989 remain robust to the main result, as is reported in Column (1) of Table A1. These papers experience a 2.0 percent point increase in the likelihood to be cited by an Internet adopting firm (which translates into an 8.5% increase and is statistically significant at the 5% level).

In addition, I use data mining to collect the digitization data of each paper that is available on the PubMed platform. PubMed is a search engine that provides online access to papers in the MEDLINE database on life sciences and biomedical topics. The National Library of Medicine started to create the digital code for selected papers since the 1970s[10]. As a result, these papers are more accessible in the offline database in the local library using the retrieval language by the librarians. Column (2) in Table A1 shows that these more accessible old papers

[9] The definition of OECD fields is defined at: http://www.oecd.org/science/inno/38235147.pdf.
[10] author's interview with the U.S. National Library of Medicine, History of Medicine Division

with digital codes have no significant correlation with the Internet adoption. In contrast, column (3) reports the regression results for all other old papers published before 1989 that have no MEDLINE digital codes, and thus might be harder to find in offline databases. These less accessible papers experience a boost in the likelihood to be cited by Internet adopting firms with a 2.3 percentage point increase (10% percent increase) and statistically significant at 5% level. Figure A1 shows that the distributions of the two set of papers with and without digital codes are consistent. However, the results can also be driven from a field specific effect, as the PubMed papers are limited to life sciences literature.

6. Conclusion and discussions

6.1 Conclusion

This paper explores how the Internet affect firms' exploration of high-status and low-status scientific publications in patent-invention process. It suggests that the Internet enables firms to discover "hidden gems" – commercializable yet under-recognized scientific findings. These are findings that come from *early-career* scientists, from work published in *less prestigious* journals, and from papers with *fewer academic citations* but have *higher forward patent citations*. These findings suggest that IT reshapes the process of how firms source knowledge for innovation. By reducing search costs, IT enables firms to disproportionately increase the reliance on scientific knowledge that previously had tended to receive less attention. The results shed light on how IT reinforces the link between science and innovation.

6.2 Discussion

6.2.1 Long-term effects vs. short-term effects

This paper focuses on the short-run impact of the Internet on firms' utilization of scientific knowledge in innovation processes. In the long run, complementary changes in process innovation within organizations can lead to a change in the product-innovation process in ways that involve a higher reliance on science. For example, in the case of complex innovation, such as the discovery of new drugs, digital science can transform innovation by integrate new knowledge into innovation processes. Here, digital science refers not only to digitized scientific articles online, but also new ways to "measure, analyze, and model chemical compounds, diseases, and human biology" (Dougherty and Dunne, 2012). Access to the Internet increases the availability and creation of digital science at firms. It facilitates learning about emerging digital knowledge through a wide variety of publications, including scientific articles and academic blogs. In addition, the use of IT at organizations can be associated with an emerging boundary spanning. This can reinforce the use of digital knowledge in complex innovation, which usually involves cooperation among workers from different divisions (Levina and Vaast, 2005).

References

Ahmadpoor, M. and Jones, B.F., 2017. The dual frontier: Patented inventions and prior scientific advance. Science, 357(6351), pp.583-587.

Alberts, B. 2010. Is the frontier really endless? Science 330(6011):1587.

Agrawal, A. and Goldfarb, A., 2008. Restructuring research: Communication costs and the democratization of university innovation. *American Economic Review,* 98(4), pp.1578-90.

Arora, A., Belenzon, S., Marx, M. and Shvadron, D., 2021. (When) Does Patent Protection Spur Cumulative Research Within Firms? (No. w28880). National Bureau of Economic Research.

Arora, A., Belenzon, S. and Suh, J., 2021. Science and the Market for Technology (No. w28534). National Bureau of Economic Research.

Autor, D., Dorn, D., Hanson, G.H., Pisano, G. and Shu, P. 2020. "Foreign Competition and Domestic Innovation: Evidence from US Patents." American Economic Review: Insights, 2 (3): 357-74.

Bar-Isaac, H., Caruana, G. and Cuñat, V., 2012. Search, design, and market structure. *American Economic Review*, 102(2), pp.1140-60.

Bush, V. Science the Endless Frontier, A Report to the President (U.S. Government Printing Office, Washington, DC, July 1945).

Bikard, M., 2018. Made in academia: The effect of institutional origin on inventors' attention to science. *Organization Science, 29(5),* pp.818-836.

Bikard, M. and Marx, M., 2020. Bridging academia and industry: How geographic hubs connect university science and corporate technology. *Management Science,* 66(8), pp.3425-3443.

Brynjolfsson, E., Hu, Y. and Smith, M.D., 2010. Research commentary—long tails vs. superstars: The effect of information technology on product variety and sales concentration patterns. *Information Systems Research,* 21(4), pp.736-747.

Brynjolfsson, E., Hu, Y. and Simester, D., 2011. Goodbye pareto principle, hello long tail: The effect of search costs on the concentration of product sales. *Management Science,* 57(8), pp.1373-1386.

Corredoira, R., Goldfarb, B., Sampson, R.C. and Shi, Y., 2021. The Changing Nature of Firm R&D: Short-termism & Technological Influence of US Firms.

Cohen, W.M. and Levinthal, D.A., 1989. Innovation and learning: the two faces of R & D. *The economic journal,* 99(397), pp.569-596.

Cohen, W.M. and Levinthal, D.A., 1990. Absorptive capacity: A new perspective on learning and innovation. *Administrative science quarterly*, pp.128-152.

Ding, W.W., Levin, S.G., Stephan, P.E. and Winkler, A.E., 2010. The impact of information technology on academic scientists' productivity and collaboration patterns. *Management Science,* 56(9), pp.1439-1461.

Easley, D. and Kleinberg, J., 2010. Networks, crowds, and markets (Vol. 8). *Cambridge: Cambridge university press*.

Fleming L. 2001. Recombinant uncertainty in technological search. *Management Science 47(1): - 117*.

Fleming, L., O. Sorenson. 2004. Science as a map in technological search. *Strategic Management J. 25 909-928*

Fleming, L., Greene, H., Li, G., Marx, M., & Yao, D. (2019*)*. Government-funded research increasingly fuels innovation.*Science, 364(6446), 1139–1141*.

Forman, C. and Zeebroeck, N.V., 2012. From wires to partners: How the Internet has fostered R&D collaborations within firms. *Management science, 58(8), pp.1549-1568*.

Forman, C. and van Zeebroeck, N., 2019. Digital technology adoption and knowledge flows within firms: Can the Internet overcome geographic and technological distance?. *Research policy, 48(8), p.103697*.

Gittelman, M. and Kogut, B., 2003. Does good science lead to valuable knowledge? Biotechnology firms and the evolutionary logic of citation patterns. *Management Science,* 49(4), pp.366-382.

Janger, J. and Nowotny, K., 2016. Job choice in academia. *Research Policy*, 45(8), pp.1672-1683.

Jones, B.F., 2009. The burden of knowledge and the "death of the renaissance man": Is innovation getting harder?. *The Review of Economic Studies,* 76(1), pp.283-317.

Laursen, K. and Salter, A., 2006. Open for innovation: the role of openness in explaining innovation performance among UK manufacturing firms. *Strategic management journal*, 27(2), pp.131-150.

Laursen, K. and Salter, A.J., 2014. The paradox of openness: Appropriability, external search and collaboration. *Research policy*, 43(5), pp.867-878.

Leiponen, A. and Helfat, C.E., 2010. Innovation objectives, knowledge sources, and the benefits of breadth. *Strategic management journal,* 31(2), pp.224-236.

Levin, S.G. and Stephan, P.E., 1991. Research productivity over the life cycle: Evidence for academic scientists. *The American economic review,* pp.114-132.

Marx, M. and Hsu, D.H., 2021. Revisiting the Entrepreneurial Commercialization of Academic Science: Evidence from "Twin" Discoveries. *Management Science*.

Marx, M. and Fuegi, A., 2020. Reliance on science: Worldwide front-page patent citations to scientific articles. Strategic Management Journal, 41(9), pp.1572-1594.

Nagaraj, A and Reimers, I., 2021. Digitization and the Demand for Physical Works: Evidence from the Google Books Project. Available at SSRN: https://ssrn.com/abstract=3339524 or http://dx.doi.org/10.2139/ssrn.3339524

Hall, B.H., Jaffe, A.B. and Trajtenberg, M., 2001. The NBER patent citation data file: Lessons, insights and methodological tools.

Koput, W. K., 1997. A Chaotic Model of Innovative Search: Some Answers, Many Questions. Organization Science 8 (5) 528-542 https://doi.org/10.1287/orsc.8.5.528

Roach, M. and Cohen, W.M., 2013. Lens or prism? Patent citations as a measure of knowledge flows from public research. Management Science, 59(2), pp.504-525.

Rosenkopf L, Nerkar A. 2001. Beyond local search: boundary-spanning, exploration, and impact in the optical disk industry. Strategic Management Journal 22(4): 287-306.

Van Alstyne, M. and Brynjolfsson, E., 2005. Global village or cyber-balkans? Modeling and measuring the integration of electronic communities. Management Science, 51(6), pp.851-868.

Watzinger, M., Krieger, J.L. and Schnitzer, M., 2021. Standing on the Shoulders of Science. Harvard Business School Working Paper 21-128

Zhang, M.B., 2019. Labor-technology substitution: Implications for asset pricing. The Journal of Finance, 74(4), pp.1793-1839.

Zheng, Y. and Wang, Q., 2020. Shadow of the great firewall: The impact of Google blockade on innovation in China. Strategic Management Journal, 41(12), pp.2234-2260

Table 1. Summary statistics of key variables

Variable	Mean	Std. Dev.	Min	Max
Internet adoption	0.397	0.489	0	1
Incidence of patent citation to all scientific papers	0.281	0.449	0	1
papers in the bottom 25% impact factor journals	0.227	0.419	0	1
papers in the bottom 50% impact factor journals	0.270	0.444	0	1
papers in the top 25% impact factor journals	0.111	0.314	0	1
papers in the top 50% impact factor journals	0.172	0.378	0	1
papers all authors at the first 3 years of career-long publication	0.240	0.427	0	1
papers with the first author at first 3 years of career-long publication	0.256	0.436	0	1
papers with all authors above 5 years of career-long publication	0.152	0.359	0	1
papers with the first author above 5 years of career-long publication	0.199	0.399	0	1
Log number of employees	6.307	1.084	4.605	12.613
Log of patent applications in current period	1.074	1.295	0	7.492
Log of patent applications in local MSA	6.232	1.631	0	9.828
Log of R&D expenditures	3.912	2.819	0	9.758
Dummy for existence of chemical patent	0.221	0.415	0	1
Dummy for existence of computers and communications (C&C) patent	0.167	0.373	0	1
Dummy for existence of drugs and medical (D&M) patent	0.083	0.277	0	1
Dummy for existence of electrical and electronics (E&E) patent	0.233	0.422	0	1
Dummy for existence of mechanical patent	0.259	0.438	0	1
Dummy for existence of other patent	0.260	0.438	0	1
Instrument variables				
Year of change to Price cap regulation x after 1996 dummy	56.628	46.295	0	99

Number of observations is 18,255.

Table 2. Baseline DID estimation results of Internet adoption on patent to science citations

VARIABLES	(1) Entire sample	(2) Regular patentors	(3) W/O top 2% largest firm-MSAs
Internet adoption	0.026**	0.041**	0.029**
	(0.011)	(0.019)	(0.011)
log number of employees	0.003	-0.001	0.005
	(0.012)	(0.016)	(0.011)
log of patent applications in current period	0.183***	0.166***	0.188***
	(0.009)	(0.009)	(0.007)
log of patent applications in local MSA	-0.012	-0.039	-0.012
	(0.015)	(0.033)	(0.015)
log of R&D expenditures	0.003	-0.000	0.008
	(0.003)	(0.005)	(0.010)
Observations	18,255	7,630	18,200
R-squared	0.724	0.674	0.721
Year FE	YES	YES	YES
Firm-MSA FE	YES	YES	YES
HJT patent category dummies	YES	YES	YES
Mean of DV	0.281	0.552	0.2795
Percentage change	9.3%	7.4%	10.3%

Notes: The dependent variable is the incidence of a patent citation to scientific papers in firm-MSA, and the independent variable is Internet adoption (access to basic internet) at firm-MSA. All regressions include a constant term, firm-MSA fixed effects, time dummies, and dummies for six HJT patent category indicators.
* Significant at the 10% level. ** Significant at the 5% level. *** Significant at the 1% level.

Table 3. Heterogenous effects of firm internet adoption on patent citation to scientific papers by career stage of scientists

	(1)	(2)	(3)	(4)
	Young scientists' paper		Senior scientists' paper	
DV: incidence of patent citation to papers	All authors at first 3 years of career-long publication	First author at first 3 years of career-long publication	All authors above 5 years of career-long publication	First authors above 5 years of career-long publication
Internet adoption	**0.025****	**0.031*****	0.013	-0.004
	(0.011)	**(0.011)**	(0.010)	(0.008)
Log of number of employees	**-0.004**	**-0.004**	0.003	-0.011
	(0.011)	(0.011)	(0.009)	(0.010)
log of patents in current period	0.177***	0.181***	0.148***	0.130***
	(0.009)	(0.009)	(0.010)	(0.011)
log of patents in local MSA	-0.005	-0.010	-0.003	0.007
	(0.012)	(0.013)	(0.013)	(0.011)
log of R&D spending	0.002	0.003	0.005**	0.004*
	(0.002)	(0.002)	(0.002)	(0.002)
Observations	18,255	18,255	18,255	18,255
R-squared	0.704	0.716	0.698	0.671
Year FE	YES	YES	YES	YES
Firm-MSA FE	YES	YES	YES	YES
HJT patent category dummies	YES	YES	YES	YES
Mean of DV	0.240	0.256	0.199	0.152
Percentage change	10.4%	12.1%	6.5%	-2.6%

Notes: the dependent variable is patent citation to early-career scientists' papers. A paper is defined as early-career work if the publication year is within the first 3 years of this author's first publication. I identify every author's first publication year using all papers published before 2001 from Microsoft Academic Graph.

Table 4. Heterogenous effects of Internet adoption on patent citation to scientific papers by journal impact factor

	(1)	(2)	(3)	(4)
			Entire sample	
DV: incidence of patent citation to papers	Citation to bottom 25% papers	Citation to bottom 50% papers	Citation to top 25% papers	Citation to top 50% papers
Internet adoption	**0.030*** **	**0.026** **	-0.008	0.011
	(0.010)	**(0.011)**	(0.007)	(0.007)
Log of number of employees	0.002	0.004	0.000	-0.009
	(0.011)	(0.011)	(0.006)	(0.007)
log of patents in current period	0.162***	0.182***	0.093***	0.131***
	(0.010)	(0.010)	(0.010)	(0.009)
log of patents in local MSA	0.012	-0.004	0.008	-0.015
	(0.012)	(0.013)	(0.009)	(0.011)
log of R&D spending	0.002	0.002	0.008***	0.008***
	(0.003)	(0.003)	(0.002)	(0.002)
Constant	-0.083	0.008	-0.075	0.136*
	(0.098)	(0.107)	(0.072)	(0.080)
Observations	18,255	18,255	18,255	18,255
R-squared	0.696	0.714	0.705	0.726
Year FE	YES	YES	YES	YES
Firm-MSA FE	YES	YES	YES	YES
HJT patent category dummies	YES	YES	YES	YES
Mean of DV	0.227	0.270	0.111	0.172
Percentage change	13.2%	9.6%	-7.2%	6.4%

Notes: the dependent variable is patent citation to papers in lower ranked journals. The ranking is based on the Journal Impact Factor from publication year, and the quartiles are based on the 167,882 cited papers in estimation sample.

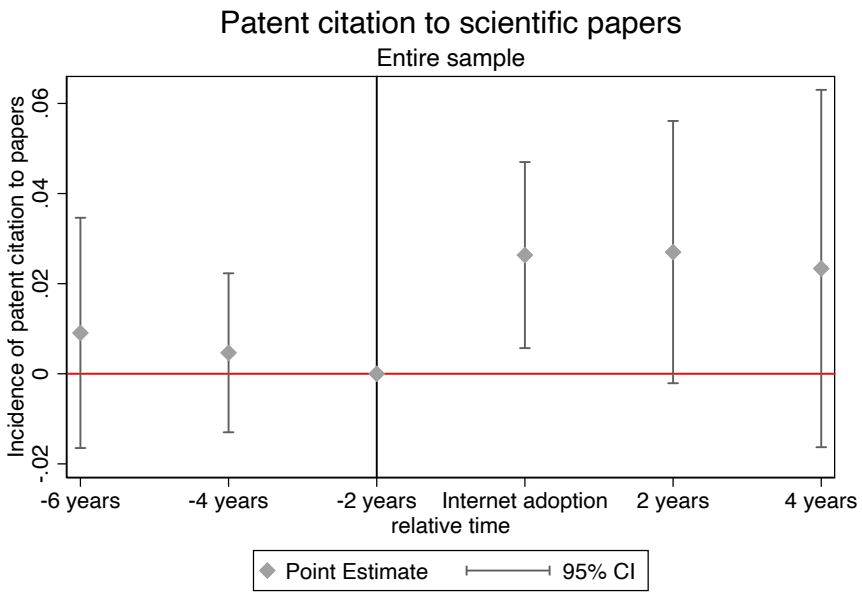
Table 5. Paper impact: Heterogenous effects of Internet adoption on patent citation to scientific papers by forward academic citations

Incidence of patent citation to papers with	(1) Top 10% paper impact by 2000	(2) Top 25% paper impact by 2000	(4) bottom 39% papers with zero citations by 2000
Internet adoption	-0.007	0.009	**0.020***
	(0.007)	(0.007)	**(0.011)**
log number of employees	-0.013*	-0.005	0.001
	(0.007)	(0.007)	(0.011)
log of patent applications in current period	0.114***	0.154***	0.188***
	(0.009)	(0.008)	(0.007)
log of patent applications in local MSA	0.006	-0.008	-0.012
	(0.009)	(0.009)	(0.014)
log of R&D expenditures	0.020***	0.009	0.006
	(0.007)	(0.008)	(0.009)
Constant	0.008	0.071	0.071
	(0.067)	(0.075)	(0.112)
Observations	18,255	18,255	18,255
R-squared	0.666	0.706	0.703
Year FE	YES	YES	YES
Firm-MSA FE	YES	YES	YES
HJT patent category dummies	YES	YES	YES
Mean DV	0.0949	0.1415	0.2456
Percentage change	-7.4%	6.4%	8.1%

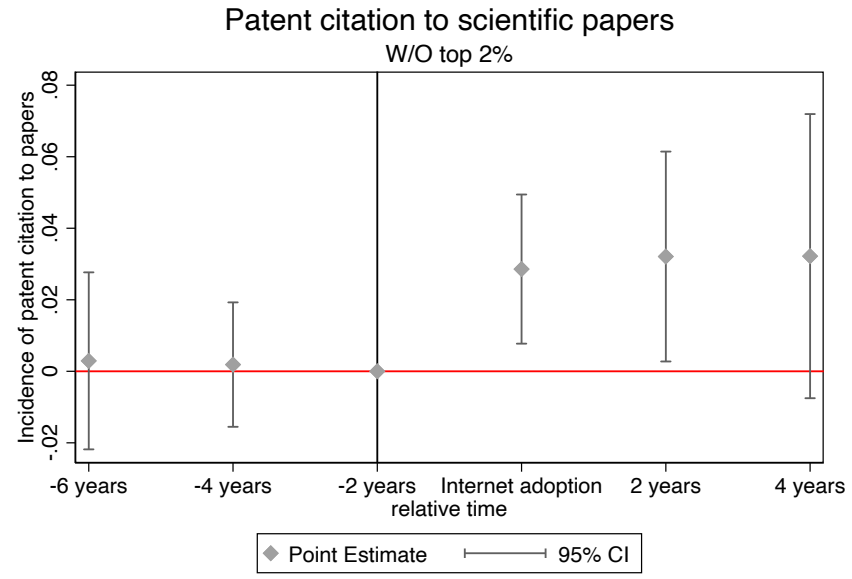
The bottom 39% of the unique 167,881 cited papers receive zero academic citations till 2000.

Table 6. Paper commercial impact: Heterogenous effects of Internet adoption on patent citation to scientific papers by paper commercial impact

Incidence of patent citation to papers with	(1) Top 10% forward patent citations by 2018	(2) Top 25% forward patent citations by 2018	(3) Bottom 25% forward patent citations by 2018
Internet adoption	0.018**	0.020**	0.002
	(0.007)	(0.009)	(0.009)
log number of employees	0.009	0.012	-0.001
	(0.008)	(0.010)	(0.009)
log of patent applications in current period	0.155***	0.164***	0.200***
	(0.008)	(0.009)	(0.009)
log of patent applications in local MSA	0.006	0.006	-0.011
	(0.010)	(0.010)	(0.013)
log of R&D expenditures	0.022***	0.015	0.002
	(0.007)	(0.009)	(0.008)
Constant	0.018**	0.020**	0.002
	(0.007)	(0.009)	(0.009)
Observations	18,255	18,255	18,255
R-squared	0.703	0.72	0.655
Year FE	YES	YES	YES
Firm-MSA FE	YES	YES	YES
HJT patent category dummies	YES	YES	YES
Mean of DV	0.148	0.187	0.190
Percentage change	12.2%	10.7%	1.1%



(a)



(b)

Figure 1. Event studies of Internet adoption on Patent to science citations

Note: This figure shows the event study results of Internet adoption on patent to scientific paper citations in the estimation sample. The dependent variable in this regression is the incidence of patent to scientific paper citations, and the independent variables include three pre-adoption indicators and two post-adoption indicators, firm-MSA fixed effects, dummies of HJT patent technology classifications, log employment size, log number of patent applications at firm-MSA, log of R&D expenditure, and log of patent applications in local MSA. Standard errors are clustered at MSA level.

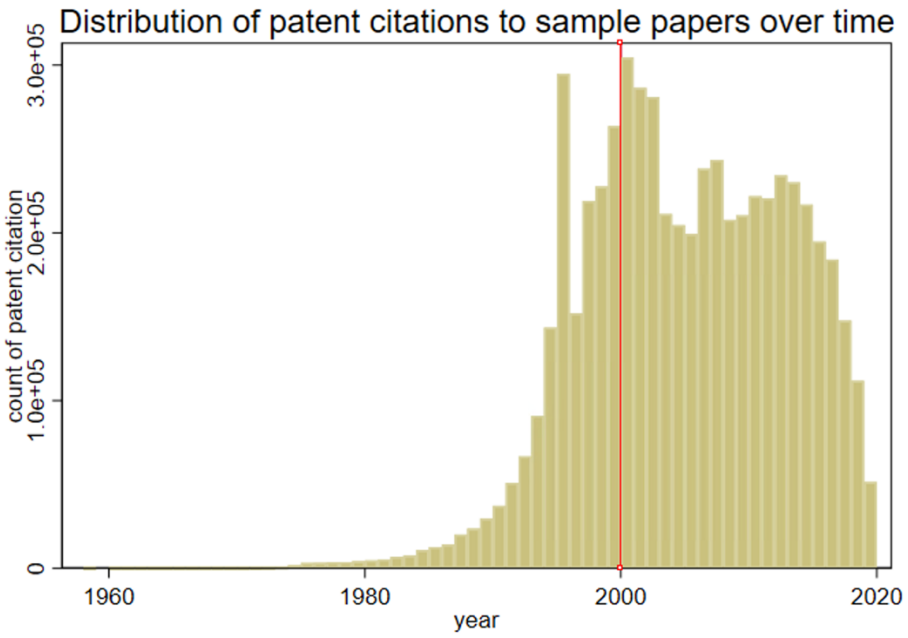


Figure 2. Distribution of patent citations to sample papers over time

Notes: The definition of paper commercial impact is citations from USPTO granted patents by the end of 2018. Each observation in this figure is a patent-paper citation received by sample papers in the entire PCS dataset.

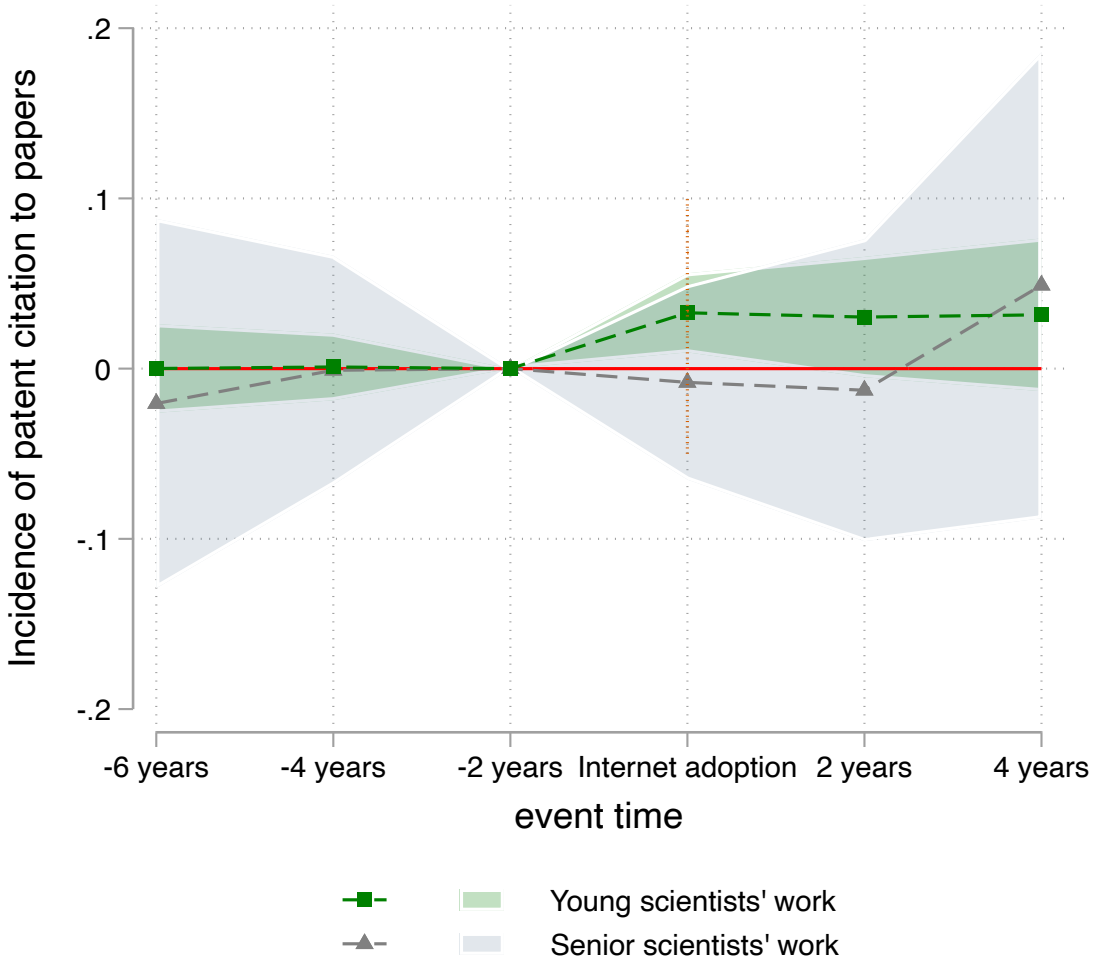


Figure 3. Event studies of internet adoption on patent citation to scientific papers by career stage of paper authors

Note: This figure shows the event study results of Internet adoption on patent to scientific paper citations by career stage of paper author. I regress the incidence of patent to citation to papers written by young scientist and by senior scientist, respectively. The independent variables include three pre-adoption indicators and three post-adoption indicators, firm-MSA fixed effects, dummies of HJT patent technology classifications, log employment size, log number of patent applications at firm-MSA, log of R&D expenditure, and log of patent applications in local MSA. Standard errors are clustered at MSA level.

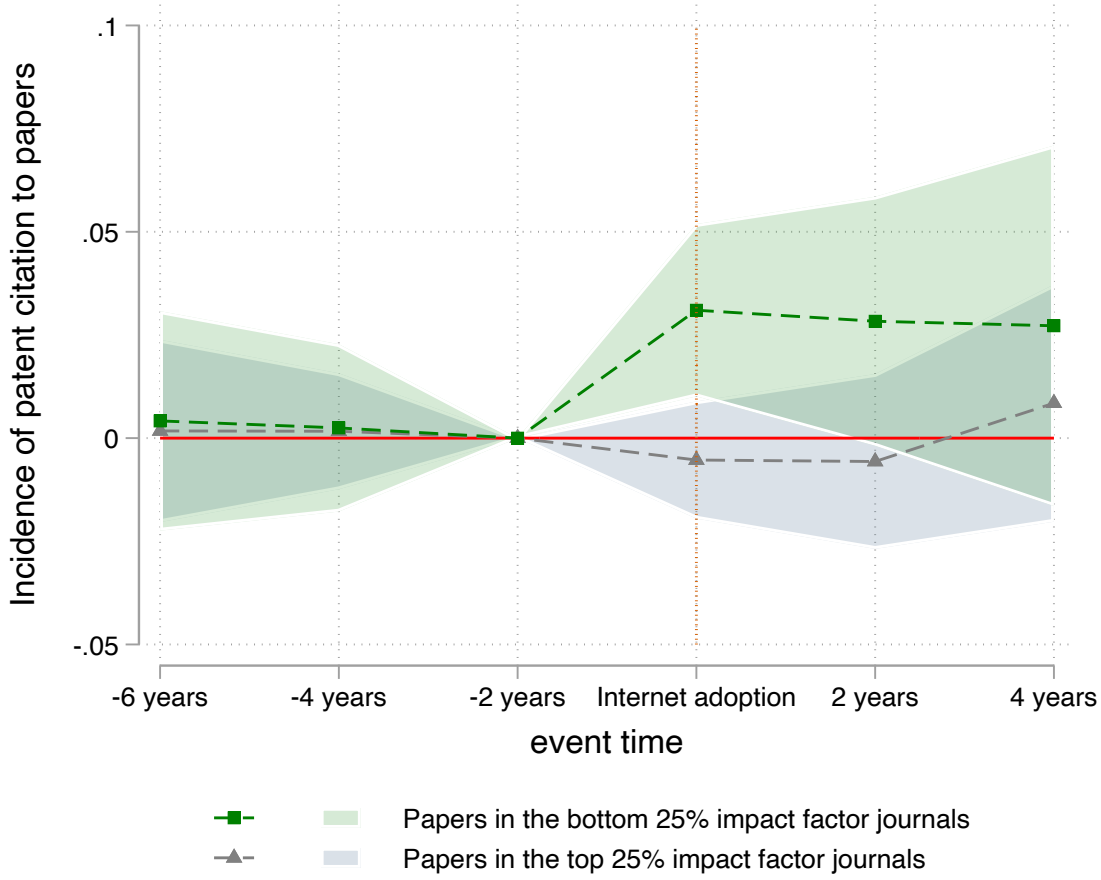


Figure 4. Event studies of internet adoption on patent citation to scientific papers by journal impact factor

Note: This figure shows the event study results of Internet adoption on patent to scientific paper citations by journal impact of cited papers. I regress the incidence of patent to citation to papers in the bottom 25% impact factor journals and patent citation to papers in the top 25% impact factor journals, respectively. The independent variables include three pre-adoption indicators and three post-adoption indicators, firm-MSA fixed effects, dummies of HJT patent technology classifications, log employment size, log number of patent applications at firm-MSA, log of R&D expenditure, and log of patent applications in local MSA. Standard errors are clustered at MSA level.

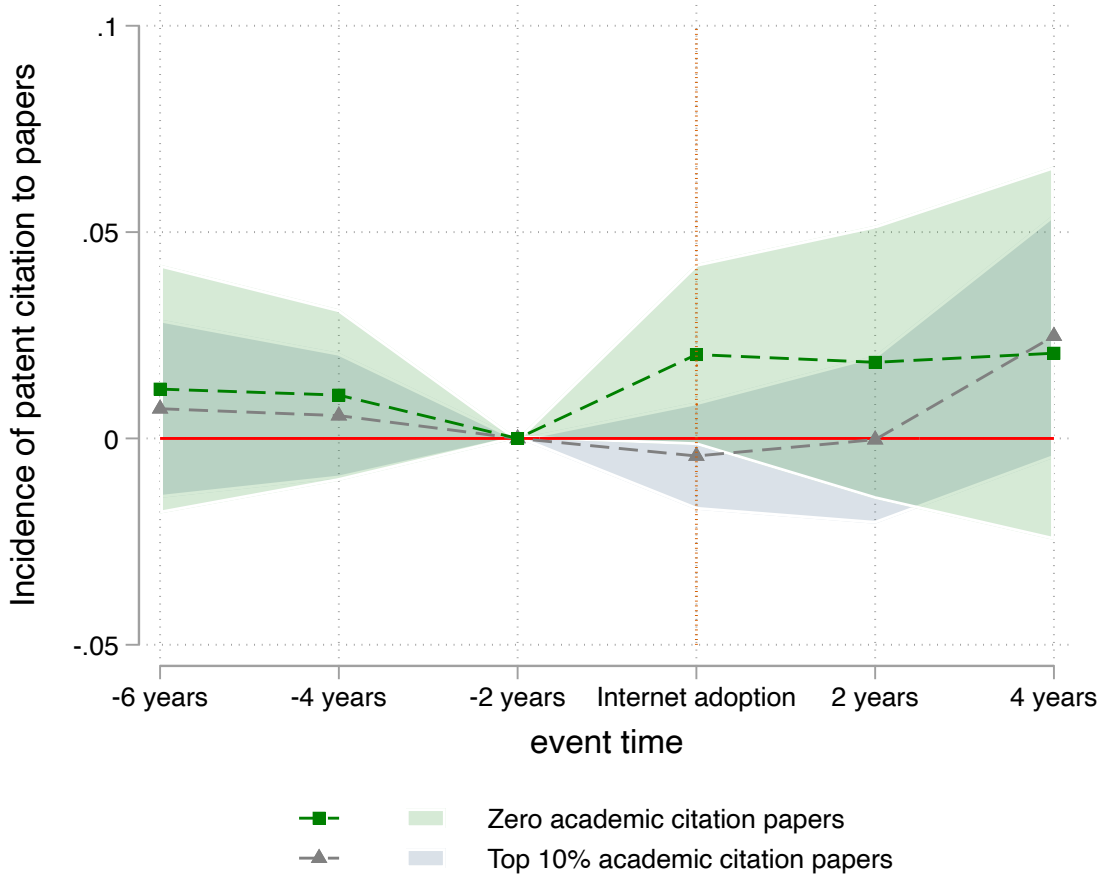


Figure 5. Event studies of internet adoption on patent citation to scientific papers by academic citations of papers

Note: This figure shows the event study results of Internet adoption on patent to scientific paper citations by forward academic citation of cited papers. I regress the incidence of patent to citation to papers with zero academic citations and with top 10% academic citations, respectively. The independent variables include three pre-adoption indicators and three post-adoption indicators, firm-MSA fixed effects, dummies of HJT patent technology classifications, log employment size, log number of patent applications at firm-MSA, log of R&D expenditure, and log of patent applications in local MSA. Standard errors are clustered at MSA level.

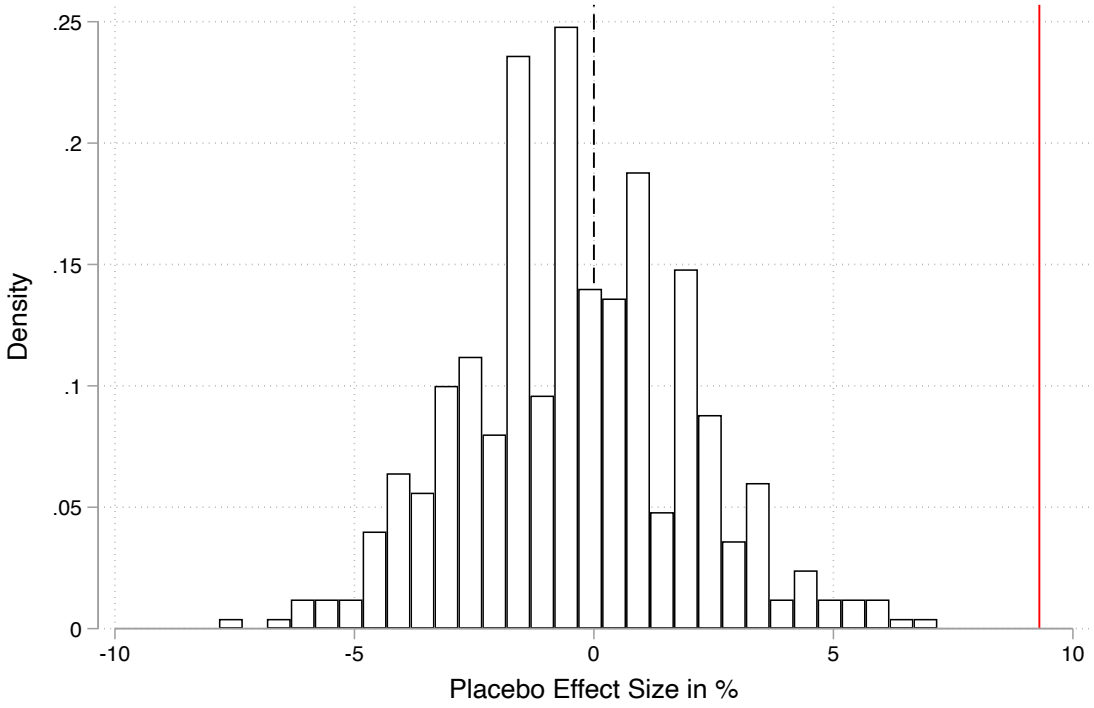


Figure 6. Placebo Test with Random Internet Adoption Years

Notes: This figure plots the results of a placebo test that randomly assign Internet adoption year. I conduct 500 estimates and draw the distribution of the 500 placebo effect sizes (each effect size is calculated by compare the coefficient to the average rate of patent citation to science). The red line represents the observed effect size, which lines outside of the 99% confidence interval of the distribution of the coefficients.

Appendix A. Additional Robustness Checks

Table A1. Timing falsification test for citation to low-status papers

	(5)	(6)	(3)	(4)	(7)	(8)	(9)	(10)
	First author young		bottom 25% journal impact factor		Zero academic citation		Top 25% forward patent citations	
DV: Patent citation to paper…								
Internet adoption	0.032***	0.033***	0.031***	0.031***	0.020*	0.019*	0.020*	0.019*
	(0.012)	(0.012)	(0.010)	(0.010)	(0.011)	(0.011)	(0.011)	(0.011)
Internet adoption in future two years	0.000		-0.002		-0.010		-0.010	
	(0.010)		(0.011)		(0.010)		(0.010)	
Internet adoption in future four years	0.007		0.003		0.004		0.004	
	(0.009)		(0.009)		(0.010)		(0.010)	
Internet adoption in future two or four years		0.001		-0.001		-0.002		-0.002
		(0.009)		(0.010)		(0.010)		(0.010)
log number of employees	-0.003	-0.003	0.003	0.003	0.001	0.001	0.001	0.001
	(0.010)	(0.010)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)
log of patent applications in current period	0.195***	0.195***	0.194***	0.194***	0.188***	0.188***	0.188***	0.188***
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
log of patent applications in local MSA	-0.013	-0.013	0.008	0.008	-0.012	-0.012	-0.012	-0.012
	(0.013)	(0.013)	(0.012)	(0.012)	(0.014)	(0.014)	(0.014)	(0.014)
log of R&D expenditures	0.014	0.014	0.004	0.004	0.006	0.006	0.006	0.006
	(0.010)	(0.010)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
Observations	18,255	18,255	18,255	18,255	18,255	18,255	18,255	18,255
R-squared	0.713	0.713	0.693	0.693	0.703	0.703	0.723	0.723
Year FE	YES	YES	YES	YES	YES	YES	YES	YES
HJT patent category dummies	YES	YES	YES	YES	YES	YES	YES	YES
Firm-location FE	YES	YES	YES	YES	YES	YES	YES	YES

Table A2. Effects of Internet adoption on patent citation to a subset of old papers

Incidence of patent citation to	(1) All papers published before 1989	(2) Papers published before 1989 with a PubMed digital code (easily accessible in the library before the Internet age)	(3) All other old papers (1) excluding (2)
Internet adoption	0.020**	-0.004	0.023**
	(0.010)	(0.006)	(0.009)
log number of employees	0.028***	0.012***	0.027***
	(0.006)	(0.004)	(0.006)
log of patent applications in current period	-0.016**	0.003	-0.015**
	(0.006)	(0.004)	(0.006)
log of patent applications in local MSA	0.010	-0.010	0.013
	(0.013)	(0.008)	(0.013)
log of R&D expenditures	0.026***	0.010**	0.031***
	(0.008)	(0.005)	(0.008)
Observations	18,255	18,255	18,255
Year FE	YES	YES	YES
Firm-MSA FE	YES	YES	YES
HJT patent category dummies	YES	YES	YES
R-squared	0.679	0.697	0.703
Mean DV	0.236	0.073	0.229
Percentage change	8.5%	-5.5%	10.0%

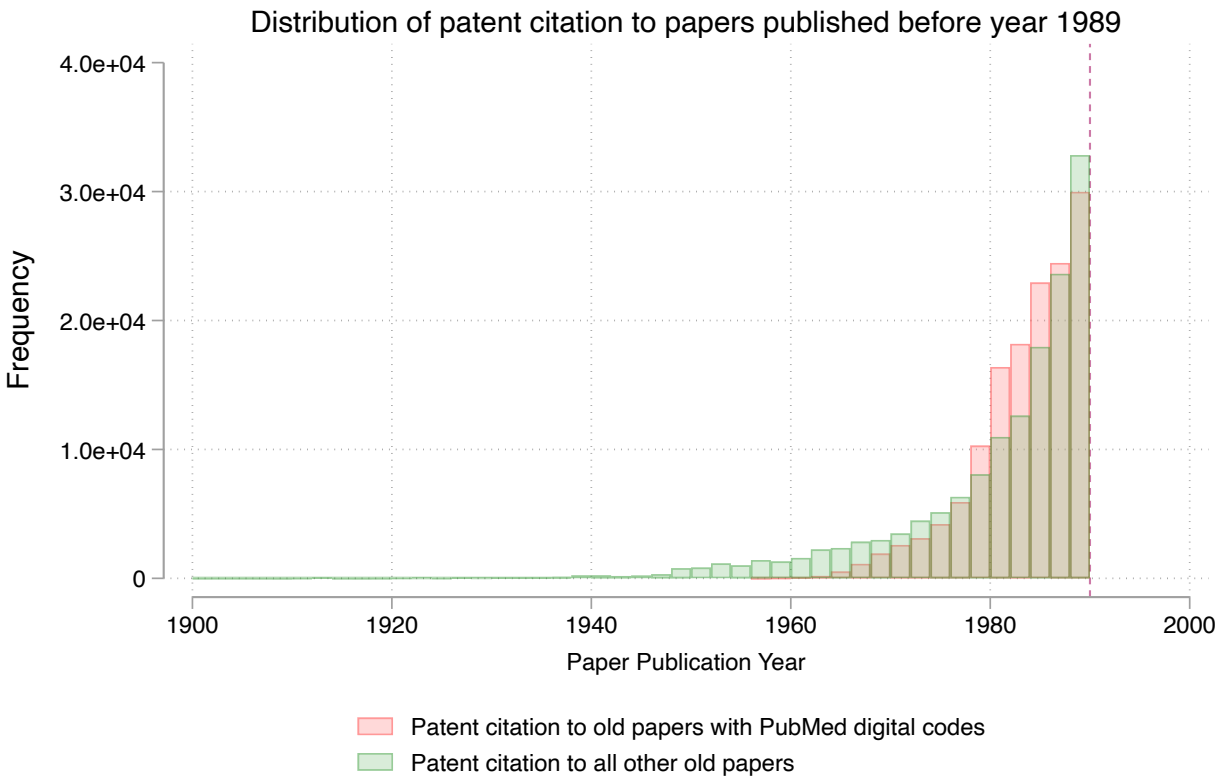


Figure A1. Distribution of patent citations to papers published before year 1989

Notes: each observation is a patent-paper citation link. The red shadow represents the citation to papers that have PubMed digital codes. by the time of citation (Column (2) in Table A1), and the green shadow shows the citation to all other old papers (Colum (3) in Table A2).

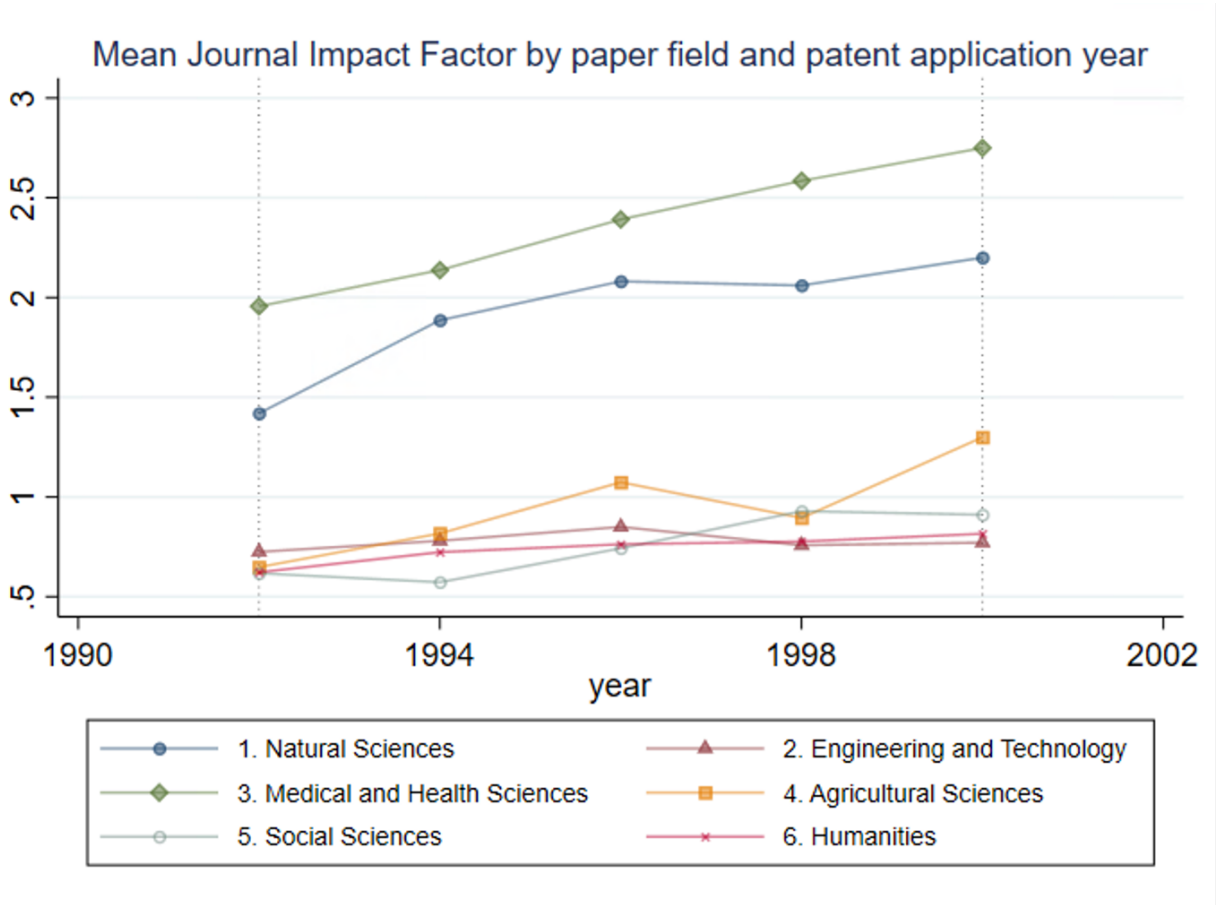


Figure A2. Distribution of Journal Impact Factor of sample papers by OECD six fields of studies

Table A3. Effects of Internet on patent to science citation by journal impact factor (adjusted by OECD paper field)

	(1)	(2)	(4)	(5)
Journal Impact Factor (JIF) Ranked by OECD field	Top 10%	Top 25%	Bottom 10%	Bottom 25%
Internet adoption	-0.006	0.001	0.031***	0.028***
	(0.007)	(0.007)	(0.009)	(0.010)
log number of employees	-0.002	0.003	0.007	0.007
	(0.007)	(0.006)	(0.011)	(0.011)
log of patent applications in current period	0.088***	0.112***	0.161***	0.170***
	(0.010)	(0.010)	(0.010)	(0.010)
log of patent applications in local MSA	0.017*	-0.012	0.010	0.006
	(0.009)	(0.009)	(0.012)	(0.013)
log of R&D expenditures	0.025***	0.030***	0.004	0.002
	(0.008)	(0.008)	(0.008)	(0.009)
Observations	18,255	18,255	18,255	18,255
Year FE	YES	YES	YES	YES
Firm-MSA FE	YES	YES	YES	YES
HJT patent category dummies	YES	YES	YES	YES
R-squared	0.679	0.705	0.697	0.703
Mean DV	0.0859	0.1321	0.2239	0.2366
Percentage change	-7.0%	0.8%	13.8%	11.8%

Appendix B

Examples of "hidden gems" – less recognized papers with high commercial value

			Academic impact			Commercial impact
Year	Journal Name	Paper Title	First author is early-career scientist	Journal impact factor at the time of paper publication	Forward paper citations by the end of 2000	Forward patent citations by the end of 2000
1969	Advances in Enzymology - and Related Areas of Molecular Biology	Solid-phase peptide synthesis	Yes	0	2	742
1983	DNA	In Vitro Deletional Mutagenesis for Bacterial Production of the 20,000-Dalton Form of Human Pituitary Growth Hormone	Yes	0	22	419
1987	Current protocols in molecular biology	Growing lambda-derived vectors	Yes	0	5	473
1991	PCR Methods Appl	Capture PCR: efficient amplification of DNA fragments adjacent to a known sequence in human and YAC DNA	Yes	0	3	985

Appendix C. Recovering missing R&D

In this paper, I assume a growth rate of R&D stock and simulate the missing observations for firms who only reported in selected years. I use a growth rate of 6%11 as a proxy for the manufacturing sector, which is the average growth rate calculated from reported firms, as shown in Table B1. My approach is similar to Simeth & Cincera (2016) and Hall and Oriani (2006).

Table B1. Observations of R&D expenditure and growth rate

Year	Obs. of R&D reporting firm-MSAs (percentage)	log of R&D	biennial growth rate
1992	2,709 (74%)	2.25	
1994	2,771 (76%)	2.38	0.059
1996	2,865 (78%)	2.57	0.079
1998	2,880 (79%)	2.72	0.059
2000	2,769 (76%)	2.81	0.033
Total	13,994 (77%)	Average biennial growth rate in R&D	0.058

Table B2. Frequency of missing/reporting annual R&D for firm-MSAs in the estimation sample from Compustats

Frequencies reporting R&D	Number of firm-MSAs	Percent	Note
0	541	14.82	Treat as zero
1	118	3.23	Treat as zero? Or below
2	133	3.64	Assume the growth to be 6% in manufacturing sector; Use base-year value if reported; if not, then use the next period value.
3	212	5.81	
4	261	7.15	
5	2,386	65.35	Use the original value from Compustats
Total	3,651 firm-MSAs	100	

[11] This is a biennial growth rate as I use every two-year data in the sample.