The Commercialization of Publicly Funded Science: How Licensing Federal Laboratory Inventions Affects Knowledge Spillovers

Gabriel A. Chan*

DRAFT: November 12, 2014

Abstract

The U.S. federal government invests $126 billion per year in research and development (R&D), 40% of which is allocated to R&D centers it exclusively funds. For over thirty years national policy has required inventions discovered in federally funded R&D centers to be transferred to the private sector to diffuse knowledge and to promote private sector follow-on innovation, but there is limited empirical evidence for whether these policies have worked. I quantify the effect of technology transfer on innovation spillovers in the context of patent licensing at the U.S. National Laboratories using data on over 800 licensed patents since 2000. I demonstrate that licensing increases the annual citation rate to a patent by 20-37%, beginning two years after a license agreement is executed. I find that over 80% of follow-on innovation after a patent is licensed occurs outside of the licensing firm, indicating that knowledge from licensing diffuses broadly. These estimates rely on a novel matching algorithm based on statistical classification of the text of patent abstracts. I explore possible mechanisms for the effect of licensing on knowledge diffusion by examining the quality of patents that cite licensed patents and rule out the possibility of a strong strategic patenting effect. These results demonstrate that transactions over formal intellectual property enhance the benefits of publicly funded R&D in the "market for ideas."

Keywords: innovation, spillovers, technology transfer, patent licensing, intellectual property, natural language processing, patent citations, matching

* Ph.D. Candidate in Public Policy, Harvard University, Cambridge, MA (gabe_chan@hksphd.harvard.edu). I would like to thank Laura Diaz-Anadon, Joe Aldy, F.M. Scherer, Bill Clark, Richard Freeman, Venky Narayanamurti, Arthur Spirling, Alberto Abadie, Sam Liss, and Maryanne Fenerjian for their advice at various stages of this project. I also thank seminar participants of the Harvard Environmental Economics Lunch, the Harvard Energy Technology Innovation Policy and Consortium for Energy Policy Seminar, the Harvard Science, Technology and Public Policy Seminar, the Harvard Sustainability Science Fellows Seminar, the Global TechMining Conference, the Atlanta Conference on Science and Innovation Policy, and the Technology Transfer Society Conference. I thank Elsie Quaite-Randall, Juliet Hart, Cheryl Fragiadakis, Cheryl Cejka, Matthew Love, Walter Copan, Kimberley Elcess, Jessica Sosenko, Kathryn Klos, Alexandra Andrego, John Lucas, Matthew Ringer, and Sarah Jane Maxted at (or formerly at) the U.S. National Laboratories and U.S. Department of Energy for sharing data and providing helpful comments. I acknowledge support from the Sustainability Science Program and the Science, Technology, and Public Policy Program at the Harvard Kennedy School. All errors are my own.

1. Introduction

The federal government's $126 billion annual investment in research and development (R&D) supports nearly one-third of all R&D in the United States[1] (National Science Board, 2014). R&D is an economically important government function because inappropriable positive spillovers arise when research discoveries serve as the foundation for follow-on innovation (Nelson, 1959; Arrow, 1962)[2]. However, realizing the full social value of R&D spillovers, particularly spillovers from publicly sponsored R&D, requires complementary investment in downstream development and commercialization of follow-on innovations (Scotchmer, 1991; Green and Scotchmer, 1995). With this understanding, various policy instruments have been adopted over the past three decades to drive private investment towards commercializing federally funded inventions. These policies allow government sponsored inventors to transfer title of their invention to a private firm for exclusive use, aligning firms' profit incentive with investment in commercializing publicly funded technology. Yet because other firms and researchers cannot utilize a technology once it is exclusively transferred, there is a concern that a single firm holding the right to utilize a technology may slow the creation of innovation spillovers (Murray and Stern, 2007; Williams, 2013; Galasso and Schankerman, 2013).

In this paper, I provide the first large-sample empirical evidence for the relationship between transferring federally funded inventions to the private sector and

[1] In 2011, a total of $424 billion was spent on R&D in all sectors. Of this total, $126 billion (30%) was provided by the federal government, and of the total federal government expenditure, $49 billion (39%) was performed directly by the federal government or by an FFRDC. For comparison, total R&D funding spent in universities was $63 billion in 2011 ($39 billion of university R&D was provided by the federal government). See Figure 1 for the full time series of these ratios from 1953 – 2011. The figure shows that while the share of Federal R&D in total R&D has been declining since the mid 1960's, the share of intramural and FFRDC R&D in the federal R&D portfolio has been relatively constant. (National Science Board, 2014)

[2] Most simply, R&D can be thought of as the process of knowledge creation. Knowledge is a public good, which implies that markets will undersupply the optimal level of R&D as long as created knowledge induces spillovers to other firms that cannot be appropriated by the firm conducting R&D. The government also has a role in conducting R&D in fields where it has market power, such as those of strategic national priority, e.g. national defense and space exploration (Cohen and Noll, 1996).

the subsequent rate of innovation spillovers. Contrary to the concern that exclusively transferred technologies are "held up" within a single firm, I find that technology transfer has a statistically significant and meaningfully large positive effect on the rate of spillovers. This finding is consistent with two socially beneficial effects of technology transfer: (1) information about a technology's value is revealed when a firm declares its willingness to invest in the technology's development (Drivas et al., 2014) and (2) knowledge about how to better produce or utilize a technology is created as a firm gains commercialization experience, sometimes referred to as "learning-by-doing." I further demonstrate that the large majority of spillovers induced by technology transfer diffuse beyond the licensing firm and that these spillover inventions are more likely to be valuable inventions themselves.

In this paper I utilize a novel dataset of patents and licensing agreements originating from five U.S. Department of Energy (DOE) FFRDCs. This dataset includes nearly 3,000 utility patents filed from 2000 to 2012, of which over 800 have been licensed. I follow the literature on technological spillovers by using the forward citations to a patent as a measure of induced follow-on innovation (Jaffe et al., 1993, 2000). My empirical strategy utilizes variation in the timing of license agreements with respect to a patent's age and a novel strategy to match licensed and unlicensed patents. I propose a new procedure for identifying patent matches that utilizes a machine learning algorithm for automated reading of patent abstracts to identify patents of similar technological scope. My proposed approach offers distinct methodological advantages relative to current approaches in the literature that rely on coarse classifications assigned by patent examiners. I demonstrate the advantages of my approach and show that matching based on classified patent abstracts is less susceptible to omitted variable bias.

Utilizing a difference-in-differences regression design, I estimate that licensing increases the rate at which a patent accumulates citations by 0.22 – 0.34 citations per year above a pre-licensing average of 0.71 citations per year. This amounts to a 31 – 48%

increase in follow-on invention induced by licensing. This estimate of licensing-induced innovation is concentrated in a period two to eight years after licensing, as I am not able to accurately measure effects beyond eight years. Alone, this result does not speak to the social benefits of technology transfer since follow-on innovation could be isolated within a licensing firm. Therefore, I examine follow-on inventions stratified by patent assignee and find that in fact, more than 75% of follow-on inventions induced by licensing occur outside of the licensing firm. Taken together, I conclude that licensing a federally funded invention enhances rather than detracts from the realization of the full social value of publicly funded R&D.

Several studies in the literature have examined innovation spillovers at the patent-level (see Section 4 for examples); however, only very recently has the literature examined the effect of formal technology transfer arrangements on innovation spillovers. In the most relevant study to this one, Drivas et al. (2014) examine the effect of patent licenses on subsequent citations with a different empirical approach and in a different institutional context, patents managed by the University of California, but find results of the same sign and of marginally greater magnitude to what I find.

Forty percent of federal R&D is conducted directly by the federal government or by federally funded research and development centers (FFRDCs), resulting in over 1,000 newly issued patents per year and a stock of over 4,000 invention license agreements (National Science Board, 2014). However, despite these totals and multiple policy reforms over the past three decades to enhance technology transfer at FFRDCs, there is limited evidence for the relationship between formal transactions over intellectual property and innovation spillovers from federally sponsored R&D outside of the university context[3]. There are two notable exceptions. Jaffe and Lerner (2001) study the effect of policy reforms and management practices on patenting and technology transfer

[3] For comparison, 50% of federal R&D is conducted by universities and colleges (National Science Board, 2014), yet patenting in the university context, covered under the Bayh-Dole Act, has received significant attention in the innovation policy literature.

activities at FFRDCs; however, the authors were limited by data availability and only examined aggregate patterns without specifically studying knowledge diffusion outcomes. Additionally, in a more narrowly focused study, Adams et al. (2003) utilize two surveys on FFRDCs from the late 1990s to conclude that cooperative research and development agreements (CRADAs) between an FFRDC and an industrial lab induce greater patenting by the industrial lab. My contribution in this paper is to provide the first estimates of the spillover effects of transferring non-university federally funded inventions to the private sector, providing a relevant input into the ongoing policy process of reforming technology transfer policies at FFRDCs[4].

Beyond the direct results of this paper, this paper also makes two contributions to the broader literature. First, the method that I develop to construct matched control patents based on the text of patent abstracts offers a new tool to reduce omitted variable bias and model dependence in the large literature in innovation economics that relies on matching patents[5]. Because it is too costly for the researcher to read each document, existing studies treat the text of patents as unobservable noise. However, this is clearly an unreasonable assumption made for convenience, as the primary purpose of patents is to disclose novel information. Second, I provide some of the first empirical evidence in the literature on the role of formal intellectual property transactions in enhancing knowledge diffusion in the "market for ideas." While representative data on patent licenses between two private actors is not readily available due to secrecy concerns, I

[4] In July 2014, the U.S House of Representatives passed H.R. 5120, the Department of Energy Laboratory Modernization and Technology Transfer Act of 2014, which would reduce certain bureaucratic requirements for technology transfer while giving new authority to adopt best practice technology transfer activities from other agencies at the Department of Energy FFRDCs. As of October 2014, the bill is in committee in the Senate. (113th Congress, 2014) In a parallel effort, the White House initiated a process in October 2011 under presidential memorandum to accelerate technology transfer at all FFRDCs by requiring federal agencies to develop five-year plans to improve technology transfer goals and metrics, streamline technology transfer processes, and develop regional commercialization partnerships (The White House, 2011). Finally, the President's Fiscal Year 2015 budget request included several relevant new initiatives, described in the President's Management Agenda as priorities for "accelerating and institutionalizing lab-to-market practices," which reflect "the Administration's commitment to accelerating and improving the transfer of the results of Federally-funded research to the commercial marketplace" (The White House, 2014).

[5] A few example studies that match patents are Jaffe et al. (1993), Thompson and Fox-Kean (2005), and Singh and Agrawal (2011).

was able to collect scrubbed information for license agreements that involved a public sector partner, making this a unique empirical opportunity to contribute to this literature which has thus far relied heavily on theoretical or sectorally-limited evidence (Arora et al., 2004; Hellmann, 2007).

The rest of this paper is organized as follows: Section 2 describes technology transfer at the U.S. National Labs and lessons to be learned from university technology transfer; Section 3 describes the data I use in my analysis; Section 4 details the empirical framework of the paper, including a description of the matching algorithm and text-based classification of patents I develop; Section 5 presents the results of my empirical work; and Section 6 concludes by putting my results in the context of the broader literature.

2. The U.S. National Labs and Technology Transfer

The U.S. National Lab system emerged from the facilities created under the Manhattan Project to build the atomic bomb during World War II. Following the War, the newly created Atomic Energy Commission (AEC) assumed responsibility for the Labs and used its new authority to expand the mission of the Labs to cover fundamental scientific research in nuclear sciences. Following the 1973 Arab Oil Embargo, new legislation and executive actions expanded the mission of the AEC to cover non-nuclear forms of energy R&D, doubled total federal investment in energy R&D, and later shifted institutional management of the Labs from the AEC to the short-lived Energy Research and Development Administration (ERDA). In 1977, the cabinet-level Department of Energy was established to replace ERDA (as well as several other energy-related federal organizations), and assumed responsibility for managing the Labs[6]. Today, the National Lab system includes seventeen labs with a combined $13 billion R&D budget (FY 2011),

[6] For a general history of the Labs see Westwick (2003) and for a detailed history of each of the seventeen Labs see DOE Office of Scientific and Technical Information (2014b).

97% of which is provided by the federal government[7]. For comparison, total R&D expenditure at U.S. universities and colleges in 2011 was $63 billion. (National Science Board, 2014)

The seventeen National Labs are heterogeneous. First, they are broadly geographically distributed across the country, as shown in Figure 2. Some are located near urban centers in close proximity to large research universities while others are in remote locations (by design as they were originally secretive nuclear research facilities). Second, the management structure of the Labs varies. While each Lab is owned by the Federal government and must report to DOE, only one of the seventeen Labs, the National Energy Technology Laboratory, is actually operated by DOE. Seven of the Labs are currently operated by a university, four are operated by a non-profit R&D company, and the remaining five are operated by an industrial corporation. The Labs also differ in the breadth of their research focus and thus vary in their technology transfer activities. Table 1, presents summary statistics for R&D expenditure, patenting activity, and licensing activity for the seventeen Labs disaggregated by their operator type. For a deeper discussion of the implications of different Lab operator types and other Lab management issues, see Jaffe and Lerner (2001), Logar et al. (2014), and Stepp et al. (2013).

Under the 1980 Stevenson-Wydler Act (P.L. 96-480) and subsequent reforms, all FFRDCS, the National Labs included, are legislatively required to transfer inventions to the private sector[8]. As part of this mission, the FFRDCs have each established technology transfer offices and appropriate a minimum of 0.5% of their R&D budget

[7] The R&D budget of the seventeen DOE labs constitute 75% of all U.S. expenditures at FFRDCs. The remaining 25% of FFRDC R&D is conducted by labs under the responsibility of other Federal agencies or organizations. Prominent examples and their sponsoring agencies are the Jet Propulsion Laboratory (NASA), Lincoln Laboratory (DOD), the National Center for Atmospheric Research (NSF), and the National Cancer Institute (HHS).

[8] For a discussion of the U.S. policy architecture for National Lab and FFRDC technology transfer, see Bozeman (2000), Jaffe and Lerner (2001), Margolis and Kammen (1999), Cannady (2013), and Federal Laboratory Consortium for Technology Transfer (2011) .

towards technology transfer, which can include several mechanisms of cooperation with the private sector (e.g. cooperative R&D agreements or "CRADAs", leasing user facilities—such as bio-refineries and cyclotrons, spin-out company formation, and patent licensing). Figure 3 compares the patenting and licensing activity enabled by Stevenson-Wydler across each federal agency.

The reforms that followed Stevenson-Wydler slowly changed the way the federal government used intellectual property (IP) to protect government-sponsored inventions. In order to facilitate the transfer of the rights to develop a publicly funded invention effectively, government lab technology transfer offices were given the mandate to quickly and thoroughly apply for IP protection so that eventual licensees could be guaranteed clear rights to utilize the technology in new product development. The long-understood tradeoff with greater IP protection is that while new inventions are disclosed publicly, access to utilizing new inventions is made exclusive to the right holder. The effect is greater incentive to develop new technologies through the lure of monopoly profits at the societal expense of slowed diffusion of protected technologies, also raising the issue of equitable access to the fruits of innovation. With technology transfer of government inventions to a commercial partner, benefits of publicly sponsored innovations accrue back to the public in the form of access to new technologies and services developed by the commercial partner. Yet, there is a second important channel through which the public benefits from inventions discovered in the organizations it funds: Namely, due to the cumulative nature of innovation (Merton, 1973; Rosenberg, 1982), the introduction of new technologies leads to inspiration for follow-on inventions. Thus, complete evaluation of a policy that affects an innovation system must account for its effect on spurring further inventions (referred to as "spillovers"), a form of positive externality (Scotchmer, 1991).While greater IP protection slows the rate of knowledge diffusion *ceretis paribus*, using IP protection to leverage additional private investment in commercializing technologies that spurs follow-on innovation is a countervailing force.

2.1. Lessons from University Technology Transfer

The effect of greater IP protection in the context of university-sponsored research has been thoroughly studied in the context of the 1980 Bayh-Dole Act (Henderson et al., 1998; Mowery et al., 2001, 2002; Hausman, 2010; Wright et al., 2014), but the effect of greater IP protection of directly funded government innovation has been given less scholarly attention. Some of the concerns raised in the context of university research also apply for government innovation. Dasgupta and David (1994) summarize one of the most prominent concerns of greater IP protection in these contexts. They argue that promoting greater "industrial tranferrability" of basic research findings may induce short-run benefits through better utilization of existing scientific knowledge but could also have dynamic costs if these activities erode future development of new scientific knowledge. This erosion can occur if researchers are required to divert their effort towards technology transfer activities, for example by dedicating time towards the difficult task of transferring the tacit knowledge that enables the utilization of transferred technology (Arora, 1995; Arora et al., 2004). An additional dimension of technology transfer of university or government inventions is its effect on the strength of existing IP. Transferring title of an invention to the private sector increases the likelihood that a patent will be litigated for infringement. FFRDCs, like universities, may be unenthusiastic about pursuing costly litigation to enforce their patent rights (Rooksby, 2013), but once title is transferred to a private actor, it becomes in the licensee's interest to pursue litigation for infringement. Finally, there is an ethical concern that inventions discovered with public funds, once transferred to a single private actor, create benefits inequitably to licensees rather than the general public.

2.2. National Lab Patent Licenses

This section briefly describes the general features of National Lab patent licensing agreements. For more detail and a typical sample license agreement see DOE

Technology Transfer Working Group (2013). It is difficult to generalize National Lab license agreements, as each license is negotiated individually with particular idiosyncratic terms. Patent license agreements are typically structured to incentivize the licensee to develop the technology (e.g. with a performance diligence requirement delineating milestone targets for technology development) while returning a share of profits from commercializing the technology back to the lab. A National Lab license agreement typically includes terms for a license issuance fee due when a license is executed, patent cost reimbursement, a minimum annual royalty, and a running royalty equal to a fixed percentage of sales. License agreements can be terminated by the licensee, typically at any point, or by the Lab if diligence requirements or royalty obligations are not met by the licensee. Finally, the U.S. government retains a "march-in" right to re-license an already licensed patent or to use a licensed patent discovered in a National Lab for purposes in the national interest. (DOE Technology Transfer Working Group, 2013; LBNL, Innovation and Partnerships Office, 2014; PNNL, Technology Transfer, 2014)

In my dataset, 49% of licensed patents are licensed on an exclusive basis, meaning the Lab agrees to not license the patent to any other interested licensee. The remaining licenses are nearly all partially exclusive, either for a particular field of use[9] or for a certain geographic region. While these non-exclusive licenses could lead to a single patent being licensed multiple times, it is very rare for a single patent to be licensed non-exclusively in such a way that licensees compete for the same market share. One additional distinction between exclusive and non-exclusive licenses is that the right to sublicense a patent to another firm is usually provided for in exclusive license

[9] An example of a patent that was non-exclusively licensed for two separate fields of use is patent 6,507,309, "Interrogation of an object for dimensional and topographical information" developed at the Pacific Northwest National Lab. This patent was licensed to a firm to develop millimeter wave body scanners for exclusive application in the fields of aviation, prison, building and border crossing security (these scanners are used extensively in U.S. airports). The same patent was later licensed to another company for a distinct field of use to create body measurements for custom-fit clothing. (Turner, 2004)

agreements but not non-exclusive licenses. For the empirical section of this paper, I utilize the first date a patent is licensed to construct the main independent variable.

In selecting licensees, federal policy requires Labs to give preference to small business licensees and licensees whose production activities are located domestically. In addition, Labs typically must justify their choice of licensee as following from a fair process, although whether or how this is enforced is unclear. In general though, licensing opportunities are advertised on DOE and Lab websites.

Markets for licensing agreements are highly frictional due to the large information asymmetries inherent in transacting over technologies. Despite the existence of patents, which disclose the primary functioning of the technology, nearly all technologies also require additional tacit knowledge possessed by the inventors to be maximally useful (Arora et al., 2004). Because this tacit knowledge is, by definition, not codified and because Lab inventors are typically not aware of the business challenges and technology needs of firms, potential licensees and Lab technology owners face large information asymmetries. This suggests that the role of technology transfer officers is important in finding the suitable matches between available Lab technologies and licensees (Hellmann, 2007). This also helps explain the long tail in the lag between when the Labs file patents and when these patents are eventually licensed (See Figure 4).

3. Data

In this study, I utilize data on 2,796 utility patents filed between January 1, 2000 and December 31, 2012 and developed at, or in partnership with, five of the seventeen National Labs (Brookhaven National Laboratory, Sandia National Laboratory, Lawrence Berkeley National Laboratory, Pacific Northwest National Laboratory, and the National Energy Technology Laboratory)[10]. These five Labs include at least one Lab in each of the

[10] I contacted the technology transfer offices at the other large National Labs engaged in applied R&D, but was not able to procure sufficient data from these labs to include in my analysis. To account for possible

four management structure categories (government operated, university operated, non-profit operated, and industry operated). Table 1 displays summary statistics of R&D expenditure, patenting, and licensing activity for the sample of five Labs I utilize in this study compared to the full set Labs. The five Labs in my sample are representative of the seventeen Labs in terms of the rate of patenting per R&D expenditure, but they licensed a slightly higher fraction of their patents (23% instead of 18%) while their average patent license brought in slightly less in terms of royalties ($20,057 instead of $27,484). For the empirical section of the paper, I only utilize data for the five Labs I have patent-level data on licensing.

Data for National Lab patents comes from two DOE databases, the U.S. Energy Innovation Portal, maintained by DOE's Office of Energy Efficiency & Renewable Energy (2014) and DOepatents, maintained by DOE's Office of Scientific and Technical Information (2014a). For each patent, I collect detailed patent-level covariates from two patent databases, the U.S. Patent and Trademark Office's (USPTO) Full-Text and Image Database (U.S. Patent and Trademark Office, 2014) and Google Patents (Google, 2014). These databases allow me to observe the DOE contract number of the R&D agreement the patent was developed under, the name of the inventors and initial assignee, the application, grant dates, priority date, the U.S. and international technology classification, and the text of the patent abstract and claims.

For the five Labs in this study, I obtained comprehensive records of patent license agreements from each Lab's technology transfer office. Of the 2,796 patents in the full dataset, I observe that 877 were licensed between January 1, 2000 and December 31, 2012[11]. For each licensed patent, I observe the date on which the licensing agreement went into effect and whether the license was issued exclusively or non-exclusively.

selection issues, I include lab fixed effects in all specifications that do not already include (co-linear) patent fixed effects.

[11] Some of the patents in the full dataset may have been licensed prior to January 1, 2000 or after December 31, 2012. However, this is not problematic for my analysis as I limit comparisons to patents filed in similar

Table 2 presents descriptive statistics for all patents and licensed patents in my dataset aggregated by observations at the patent-level (2,796 patent observations, of which 877 are licensed) and the patent-year-level (27,402 patent-year observations, 9,852 of which are for patents that are licensed between 2000 and 2012).

4. Estimating a Citation-Based Model of Knowledge Diffusion

In this paper, I assess the public returns to patent licensing, as measured by the differential citation rates of licensed patents relative to non-licensed patents. To reduce selection bias that may arise from the non-random assignment of licensing status to patents in different technological areas, I carefully match patents based on their pre-license citation trajectory and a novel measurement of their technological scope derived from the text of their abstracts. I then compare citation rates after one of the matched patents is licensed using a difference-in-differences framework.

The matching method that I propose in this sectional is novel, but the way in which I estimate and measure spillovers draws heavily on the literature that has examined innovation spillovers at the patent-level. For example, Galasso and Schankerman (2013) examine the effect of patent invalidation on subsequent citations, Singh and Agrawal (2011) ask whether firms develop follow-on innovation through hiring already successful inventors, and Jaffe et al. (1993) study how geographic proximity between an initial and follow-on inventor affect subsequent citations. As I show below, my proposed matching method has several important advantages for reducing bias and improving the precision of estimates, making it a useful tool to reevaluate some of this earlier literature.

time windows and the full database covers all unlicensed patents in the time period over which I have licensing data.

Measuring knowledge diffusion using technology-level data is difficult due to a lack of available data at a granular level. Patents, however, have proven to be a useful source of data for measuring knowledge diffusion because they include detailed information about the antecedent inventions on which a patent builds on. The citations included in a patent also play a legal role by demarcating prior art and thereby limiting the claims of a patent with respect to previous patents, and therefore citations are a noisy by still useful measure of knowledge diffusion. Patent citations are included on the front page of a patent document and are added by inventors, legal counsel, or patent examiners. For a detailed discussion of the role and significance of patent citations in the context of economic research, see Jaffe and Trajtenberg (2002).

4.1. Matching

My analysis relies on comparing patents within and across technology application areas. It is well known that the USPTO's patent classification system poorly measures what researchers seek to use the classification as a proxy for (Scherer, 1982). Existing patent classification schemes are not well suited to this task because (1) in the USPTO, patents are not classified by their potential areas of application, but instead are classified by their technical characteristics (Hirabayashi, 2003) (2) the level of granularity in patent classifications is inconsistent across technology areas, and further, within a single class there may be substantial heterogeneity across patents (Thompson and Fox-Kean, 2005), (3) classifications are continuously revised over time, (4) classification relies on idiosyncratic decisions by patent examiners and exploratory analysis reveals that very similar technologies (even pairs of patents that are co-licensed by the same specialized company) are not consistently classified[12]. In addition,

[12] One striking example of the idiosyncratic nature of USPTO classifications is shown by the example of the Combustion Controls and Diagnostics Sensors technology developed at the National Energy Technology Laboratory in the early 2000s. This technology involved an initial patent, 6,429,020 titled "Flashback detection sensor for lean premix fuel nozzles," which was filed in June 2000 and granted in August 2002. This patent was followed up with a continuation in part patent, 6,887,069 titled "Real-time combustion controls and diagnostics sensors (CCADS)" which was filed in September 2001 and granted in May 2005.

inventions, particularly high-value breakthrough inventions, very often involve the combination of technologies from distinct fields (Weitzman, 1998; Fleming, 2001; Arthur, 2009), and thus may span multiple USPTO categories. However, while patents can be placed in multiple categories, they must also have a single declared primary category. In most empirical studies, the single primary category is used for analytic traction. In addition, because of finite sample size, the coarse nature of patent classifications implies that studies that rely on matching have to discard a large number of patents due to the lack of suitable matches. All matching approaches applied to patents will result in a substantial fraction of discarded observations, as by definition, each patent must be "novel," but coarse measures like the USPTO classification make identifying more similar patents difficult.

Fortunately, patent documents do contain a plethora of information concerning the underlying innovation (that is explicitly what they are designed to do). Patents often contain pages of text and figures, and therefore, but in previous research, actual reading of the text of patents has proved too time-consuming and too substantively demanding for social science researchers to manually read each document and then determine which are most similar for conducting statistical analysis with large sample sizes[13]. Therefore, studies in the innovation literature that utilize patent data have not widely incorporated

Demonstrating the similarity of these two patents, they shared three inventors, were jointly licensed by Woodward Industrial Controls in December 2001, and were the subject of two CRADA agreements between NETL and Woodward. The abstracts of the two patents are extremely similar: Patent 6,887,069 states it is "an apparatus for the monitoring of the combustion processes within a combustion system," and patent 6,429,020 is described as "a sensor for detecting the flame occurring during a flashback condition in the fuel nozzle of a lean premix combustion system." However, they were given two different primary classifications by the USPTO. The initial patent, 6,429,020, was given the primary classification 436/153. Class 436 is described as "a generic class for … process[es] which involve a chemical reaction for determining qualitatively or quantitatively the presence of a chemical element, compound or complex," and its 153 subclass specifies that "measurement of electrical or magnetic property or thermal conductivity … of an ionized gas." The continuation patent, 6,887,069, was classified not just in a different subclass but also in a different class, 431/12, which is described as "processes of combustion or combustion starting," and its subclass specifies "processes controlling the supply of fuel or air discharged into the combustion zone." Admittedly, the two patents do share three classes designated by at least one of them as a secondary class, but neither patent contains the subclass of the other's primary subclass.

[13] One notable exception is Scherer (1982) who, with a team of four engineering and chemistry students, read and classified 15,000 quasi-randomly selected patents.

the fundamental information in patents. This leads to greater omitted variable bias and model dependence as there will be potentially greater endogeneity concerns in these studies due to a lack of meaningful observables on which to account for selection bias.

Because of the issues of using USPTO classifications and recent development in automated content analysis in computer science, I am able to implement a more sophisticated patent classification algorithm. I classify the patents in my dataset using a machine learning algorithm based on the textual content of the patent abstracts. I use the Latent Dirichlet Allocation (LDA) algorithm, which uses a Bayesian model of word co-occurrence, to classify documents into endogenously defined technology topic areas (Blei et al., 2003; Blei and Lafferty, 2007; Blei, 2010).

4.1.1. Topic Modeling

Using text as a primary data source in a causal inference framework is not straight forward. In particular, it is difficult to discern the difference between changes in behavior and changes in the way people use language to *describe* a particular behavior. One advantage of using data from the abstracts of patents, is that innovators themselves tend to write patent abstracts (lawyers are typically more involved in writing the claims in the body of a patent); therefore, it is less likely that the abstracts contain strategically motivated language. In addition, because of the legal function patents play and because of the U.S. common law system (based on precedent), the use of language within patents may be more stable than other corpuses with long time series.

Text-based analysis of natural language has a strong legacy in computer science and statistics. (See for example, Mostellar and Wallace [1964] for an early text classification analysis in statistics.) More recently, advances in computational power, the growing acceptance of text-based data in social science research, and the digitization of text sources have led to a proliferation of text-based analyses in the social sciences (Hoberg and Phillips, 2010; Hopkins and King, 2010; Quinn et al., 2010; A few recent examples include: Alexopoulos, 2011; King et al., 2013; Grimmer and Stewart, 2013; Kaplan and Vakili, 2014). Text-based analysis relevant for social science research has developed several different approaches for classification of documents into similar (predefined or unknown) categories[14].

In this paper, I classify patents based on the text of their abstracts to account for otherwise-unobserved heterogeneity in the technological scope of an invention. Text-based classification requires the selection of the appropriate method. Because technologies arise through the recombination of existing technologies (Weitzman, 1998; Fleming, 2001; Arthur, 2009), they can often span the multiple technological categories of their predecessor technologies, making classification methods that allow for multi-category membership well-suited. Further, patents are highly technical documents, which gives multi-membership models greater power in modeling the facets of individual documents (Quinn et al., 2010). Finally, the complexity of patents greatly increases the cost of implementing a supervised method that would require human coding. Therefore, I have chosen the Latent Dirichlet Allocation (LDA) model (Blei et al., 2003; Blei, 2010), which is an unsupervised mixed-membership classification method utilizing a Bayesian machine learning algorithm. A notable disadvantage of unsupervised classification is that the resulting grouping are difficult to interpret since they are not pre-defined. However, my main objective in classification doesn't rely on directly interpreting classes of patents.

Abstractly, the LDA model estimates two simultaneous types of probability distributions. First, for the corpus of documents as a whole, LDA estimates a specified number of latent "topics," representing the likelihood that words co-occur within a document. Each topic is a probability distribution over all words in the vocabulary, with

[14] Depending on the complexity of the classification task, text-based analysis can be supervised (meaning human input is required to "teach" an algorithm which documents belong in a category) or unsupervised (meaning the algorithm incorporates a method to "learn" how documents should be divided into categories). Quinn et al. (2010) and Grimmer and Stewart (2013) provide guidance for selecting appropriate classification methodologies in the context of political science research.

greater probability weight assigned to words that are more likely to occur when that topic appears in a document. LDA requires that the number of topics is pre-specified by the researcher; greater number of topics gives more granular information about the structure of a document. The second set of distributions LDA estimates are the probabilities that any one word in a document originates from one of the topics. For this paper, the relevant output of the LDA model is a document-level probability vector of topic frequencies. These frequencies provide a continuous measure of a patent's substantive content and can be used to assess the topical similarity of two different patents.

The hierarchical classification of LDA provides useful advantages in a causal inference framework relative to alternative approaches, such as word-frequency clustering (e.g. k-means), that directly classify documents based on word frequencies. First, LDA allows the researcher to control the granularity of the classification by choosing the number of topics, giving the researcher a principled method to determine the specificity of patent classification in a way that scales with the sample size. The frequency of topics also gives a chosen number of covariates to describe a patent, and thus can help separate meaningful words and combinations of words from meaningless words (whereas direct word frequencies give an intractable number of covariates, often more than the number of documents). Second, the hierarchical structure provided by the topic distributions allows for words to take on distinct meanings depending on the words they are likely to co-occur with. For example, the word "compound" is likely to have a distinct meaning mean it is used with words concerning chemical compounds compared to when it co-occurs with words that describe inventions with two compounded sub-components.

In Figure 5, I provide a visual representation of the LDA model applied to classifying the subset of patents in my dataset from the National Energy Technology Lab. In the example, I implement the LDA model with 25 topics, and show, for a single patent, the document-level distribution of topics, the ranking of most likely words to occur within the topic, and the classification of words within the document to individual topics.

While the LDA model has been applied to classifying patents before (see, for example, the recent work of Kaplan and Vakili [2014] and Venugopalan and Rai [2014]), to my knowledge, this is the first paper that incorporates LDA-based classification to account for otherwise-unobserved heterogeneity in a model of causal inference. The LDA model, including full mathematical representation, is described in detail in the Appendix.

4.1.2. Matching Implementation

The matching method I implement has several steps which seek to vastly improve the number of observable characteristics extracted from the underlying patents. There are several methodological choices required to implement a matching method with many types of covariates of different relative importance. I describe alternative methods for matching licensed to unlicensed patents and implement several alternative algorithms to assess robustness. These choices apply different combinations of exact matching, coarsened exact matching (CEM) (Iacus et al., 2012) and nearest neighbor matching on available covariates and estimated balancing scores derived from the text-based classification described in Section 4.1.1. In all approaches, I apply coarsened exact matching to filing date and average annual pre-licensing citations.

I consider three approaches to matching patents on their estimated topical structure from the LDA analysis: I consider matching patents based on a balancing score that is a function of the estimated topic proportions, direct Mahalanobis distance between estimated topics proportions, and the USPTO classification. For the methods that rely on the LDA classification, I estimate matches with a 25-topic and 50-topic model. This section refers to Table 8, where regression results described below, are estimated for different matching approaches.

In the balancing score approach, I match licensed and unlicensed patents based on a balancing score that capture the probability of a patent being licensed in a panel data setting. A filed patent can be licensed at any point and I have information on the timing of each patent's filing and license; therefore, the simple propensity score of being licensed does not fully utilize the information available to estimate licensing probability. To construct a more appropriate balancing score, I implement a Cox proportional hazard (Cox, 1984) model to estimate the (constant) hazard that a patent is licensed in any one year as a function of a patent's topical structure. The predicted hazard is a balancing score, analogous to propensity scores, the most commonly used balancing score in matching studies (Stange, 2011). I also include lab fixed effects to account for heterogeneity in lab ability to license. This allows me to utilize the information about the delay between a patent being filed and licensed as a key input for determining licensing probability. Matching on the estimated hazard ratios is assessed in specifications (1) and (2) in Table 8. The outcome of the hazard regression is predicted hazard ratios for each topic in the topic model (see Section Figure 6), which can be used to calculate expected hazards of being licensed for each patent.

The Mahalanobis distance matching approach simply takes the distance between the vector of estimated topics and finds nearest neighbor matches. I implement both the Cox proportional hazard model and the Mahalanobis distance matches on the panel of patents using the log of the estimated topic proportions from the topic model as the covariates. I use the log of estimated topic proportions to normalize the distribution of topic frequencies[15]. Matching on the Mahalanobis distance s assessed in specifications (3) and (4) in Table 8.

[15] The distribution of topic frequencies is strongly right-skewed, which is a desirable property of the LDA model as it demonstrates stronger differentiation of topics.

To reduce the noise from small document-level estimated topic proportions, I also apply a calipers approach that sets all estimated topic proportions below the 90th percentile of topic proportions to zero. Therefore, these approaches utilize only the estimated topic proportions that describe a substantial proportion of the patents. I perform CEM on a binary indicator for the patent having an above-90th percentile estimated proportion of each topic and then resolve CEM matches with nearest neighbor matching based on the estimated hazard (specificaions (5) and (6) in Table 8) and with Mahalanobis distance (specifications (7) and (8) in Table 8).

In the USPTO classification approach, I match patents exactly on their primary assigned class, ignoring secondary classes and subclasses to increase the number of matches. Still, while matching for topical scope using the hazard ratio or distance metrics preserves more than 9,000 patent-year observations, matching just on the primary USPTO classification reduces the number of observations to below 3,000 patent-year observations due to the lack of suitable controls. This again highlights the strength of the LDA classification approach in allowing for the granularity of topic matches to be defined relative to the sample size. Within PTO classes, I choose one-to-one matches randomly (specification (9) in Table 8) and also according to nearest neighbor matches in terms of the estimated hazard (specifications (10) and (11) in Table 8) and Mahalanobis distance (specifications (12) and (13) in Table 8).

In what I call the "preferred matched sample" I apply CEM matching on patent filing year and pre-license average annual citations. I then use the 50-topic LDA model and match patents one-to-one based the topic-dependent estimated hazard of being licensed.

A simple assessment of the matches is shown in Table 3, which gives naïve difference-in-difference estimates based on conditional means, suing the matching to control for pre-treatment heterogeneity in observables. Assessment of balance after matching under the preferred specification is shown in Table 4, which shows that filing year, grant year, priority year, and grant delay are all statistically significant for licensed and unlicensed patents in the full sample, but balanced in the preferred matched sample.

4.1.3. Evaluating Matching Covariates with Rosenbaum Bounds

Unbiasedness of matching estimators for causal inference depends on the conditional independence condition, which requires that after conditioning on the observable covariates used in matching, treatment assignment is independent of potential outcomes, or "as good as randomly assigned." The fundamental identification concern with matching is therefore whether there are unobserved factors that affect treatment assignment and bias causal estimates. An unobserved factor will bias a causal estimate only if it is correlated with the outcome variable and with treatment assignment. With this motivation, Rosenbaum (2002, 2005), proposes assessing sensitivity to the conditional independence assumption by calculating the maximum value of the quantity Γ that would make a causal estimate of interest no long statistically significant, where Γ is the largest difference in odds of treatment for two units with the same value of observed covariates. In effect, Γ can be thought of as a measure of the maximum co-variation in unobservable covariates and treatment assignment that preserves the conclusion of a hypothesis test based on a matching estimator. By assessing sensitivity of a matching estimator to Γ, overall sensitivity to all unobservables can be assessed, thus providing a useful assessment of the conditional independence assumption.

I compare several matching approaches using Rosenbaum Bounds to assess sensitivity of matching to violations of the conditional independence assumption. I do this on a subset of my dataset that provides a more straightforward comparison based on binary treatment and a single dependent variable. The objective of this analysis is to compare matching approaches that control for the technological scope of patents to account for non-random "assignment" to being licensed. For this section, I only examine licensed patents that were licensed in the same year that they were filed. I further restrict this subsample to patents with PTO classifications in which there is at least one licensed and one unlicensed patent (for fair comparison across methods). The final size of this subsample contains 140 licensed patents, for which I attempt to find a match in a pool of 1,131 unlicensed "control" patents. I search for non-licensed patent matches using seven approaches and calculate the average change in citations for licensed patents due to licensing, which can be thought of as an average treatment effect on the treated (ATT). The approaches control for technological scope of patents in different ways, and thus are biased to different degrees depending on how well they capture the unobservable factors driving whether a patent is licensed. The key question I want to understand with this analysis is whether capturing the technological scope of a patent using the topic modelling approach is less susceptible to omitted variable bias relative to approaches that use the examiner-assigned primary PTO classification (as previous papers in the literature have used).

For each of the seven approaches, I present the average treatment effect (on the treated) and its standard error, and the Rosenbaum Bound that would make the average treatment effect not statistically different from zero. It should be noted that the estimates presented in this section are not meant to be directly compared to the estimates in the remainder of the paper as these estimates do not take advantage of the panel nature of the data (and therefore do not include important control strategies that I utilize elsewhere, such as patent fixed effects).

Table 5 summarizes the results from the analysis of Rosenbaum Bounds. Without matching, this subsample of the data contains licensed patents that receive 0.71 (S.E. = 0.11) additional citations per year compared to unlicensed patents (which receive 0.46 citations per year). In the simplest approach, I match licensed and unlicensed patents based on a propensity score model with two sets of fixed effects for the patent's filing year and its originating Lab. With just this simple matching, the estimated difference in citations is 0.24 (S.E. = 0.36). However, this effect is not statistically significant (so there is no relevant Rosenbaum Bound). Next, I add onto this propensity score model by including fixed effects for each primary PTO classification. This increases the treatment effect estimate to 0.62 (S.E. = 0.16). The associated Rosenbaum Effect for this matching procedure is 1.34, which can be interpreted as follows: if two patents have the same probability of being licensed based only on their filing year, originating Lab, and PTO classification, unobservable factors can cause at most a 1.34 factor difference in the odds of being licensed before the estimated difference in citations is no longer statistically different from zero at the 5% level (for a two-sided test).

I compare this result to two approaches that use the topic modeling approach described earlier in this section: once using a 25 topic model and once using a 50 topic model. I include the log of the estimated patent-level topic proportions[16] along with the same three baseline fixed effects in the propensity score model. I find similar average treatment effects with these two approaches, 0.58 (S.E. = 0.18) and 0.63 (S.E. 0.17), respectively. Most importantly, I find larger values of Γ in the analysis of Rosenbaum Bounds: 1.64 and 1.45, respectively. A direct comparison of values of Γ is possible because the effect estimates from the topic modelling matches are either smaller or approximately the same as the PTO matches. In both cases, the topic models perform better than the PTO matching, and the 25 topic model performs better than the 50 topic model.

Finally, I implement two additional matching approaches that combine the PTO classification with the two topic model outputs. In these models I find treatment effect

[16] Taking the log regularizes the topic proportions and makes the distribution of topic proportions across documents close to Normal.

estimates of 0.55 (S.E. = 0.18) and 0.69 (S.E. = 0.20) with corresponding values of Γ 1.20 and 1.59.

The results from this sensitivity analysis highlight several considerations for matching design. First, matching with the topic model with 25 and 50 topics performs better than with the PTO classification. The matching procedure with 25 topics appears to be the most robust to confounders. Second, the 25 topic model performs better than the 50 topic model, but when combined with the PTO classification, this is reversed. This finding exemplifies the bias-variance tradeoff in matching (Black and Smith, 2004). Third, while the topic model performs better than the PTO classification, it is not clear whether combining the two approaches together is preferable, again because of the bias-variance tradeoff. Finally, it is important to note that the matching procedure relying on the topic models has a potentially large advantage over the PTO classification matching due to the condition of common support. Matching on the PTO classification requires dropping all patents in PTO classes in which all patents in the sample were licensed or all were unlicensed (in these cases the PTO class perfectly predicts licensing). This could be a potentially large loss of data (in this example, I dropped eight licensed patents out of 148 that met the other criteria for the subsample, but I was fortunate to have a large control group to pull from). The topic modeling matching approach allows the "distance" between any two patents of different technological scope to be compared, and therefore significantly reduces matching issues that arise from the common support condition.

4.1.4. Remaining Selection Bias Concerns

The fundamental challenge in estimating the causal effect of licensing is that patents are selected for licensing by interested agents. This can be problematic for causal identification if patents that are licensed are also cited more, for example because they are of greater value (see Table 3). A more nuanced version of selection bias can occur if patents that are available for licensing are actually licensed at a time when they would be of greater value, for example when complementary discoveries are made elsewhere in the economy. This second form of selection bias shows that controlling for pre-licensing citations, a proxy for patent value, alone may not solve the selection issue if there are secular trends in technology-area value. A third and related issue is that of simultaneity, patent licensing may occur in anticipation of a technology becoming more useful. In this case, patents may appear to be cited more after licensing but they would have had greater citations even without licensing.

Matching reduces pre-treatment imbalance, importantly bias from secular technology trends. However, all matching approaches rely on an identification assumption that all relevant variables determining selection (in this case, being licensed) are observable and controlled for to mitigate selection bias (Heckman and Navarro-Lozano, 2004; Imbens, 2004). Without random assignment of patents to licensing status, it is impossible to rule out selection bias concerns. Nevertheless, the matching approaches I present in this paper make an important contribution by providing a method to incorporate a large number of relevant characteristics of patents described by the text of the patent documents, previously treated as unobservable. Introducing these additional covariates reduces bias relative to other approaches that ignore the text of the patents as long as the text of the patent abstracts contain information relevant to whether a patent is licensed that is not incorporated elsewhere. This almost certainly has to be the case in this context if we believe licensees carefully select the patents that they license, so my approach reduces selection bias concerns relative to simpler approaches to matching on observables.

In addition to matching, I utilize patent-level fixed effects in my preferred specifications which focuses the identifying assumption on the timing of licensing relative to unobservable factors that differentially affect a patent over time. Patent fixed effects control for all unobservable time-invariant characteristics of patents, such as the patent's inherent value, so the most concerning omitted variable bias must stem from factors that simultaneously affect the timing of licensing and cause citations to licensed patents to relatively increase. Examples include a shock to a patent's value that isn't picked up by the topic model that affects license probability and citation rate, such as the discovery of a complementary technology. I feel confident that this is not likely to be a strong factor as there are huge frictions in the market to license these technologies. From qualitative interviews with technology transfer offices, Lab employees describe the huge effort and time lags required to market technologies to the right companies.

4.2. Regression Framework

I use patent citations to infer the effect that a patent has had on knowledge diffusion; when a patent is cited, I interpret this as the patent in question having spurred a subsequent invention (Jaffe et al., 1993; Jaffe and Trajtenberg, 1996, 1999). In the baseline specifications, the key dependent variable of interest is the annual citations that a patent receives. In each specification, I estimate the causal effect of a patent being licensed on the rate of citations it receives. In this section, I present a set of regression models that build up from simple cross-section and time-series regressions to full difference-in-difference regressions. The specifications that I implement are inspired by previous work that has estimated the causal effect of events on patent or paper-level citation rates (e.g. inventors changing firms, filed patents being eventually granted, and contested patents being ruled invalid) (Murray and Stern, 2007; Singh and Agrawal, 2011; Furman and Stern, 2011; Galasso and Schankerman, 2013; Drivas et al., 2014). In the specifications I discuss below, I also apply the matching approaches described in Section 4.1 to preprocess the data to reduce selection bias and model dependence.

The simplest models to quantify the effect of a patent being licensed takes either 1) a cross-section of licensed and unlicensed patents and compares citation rates or 2) a time series of licensed patents and compares citation rates before and after licensing. These two models suffer from selection bias as they do not account for the differential quality of licensed versus unlicensed patents nor possible secular trends driving both licensing and citations. Nevertheless, I present these models for comparison.

The cross-section regression is estimated with the equation

$$CITES_{i,t} = f\big(\psi_L\, EVER\_LICENSED_i + \delta_{i,t} + \gamma_t + \varepsilon_{i,t}\big). \tag{1}$$

In this equation, $CITES_{i,t}$ are citations received by patent i in year t, $EVER\_LICENSED_i$ is a dummy variable equal to 1 if patent i is licensed and 0 if the patent is not licensed, $\delta_{i,t}$ are fixed effects for a patent's age (based on the filing year of the patent), and γ_t are fixed effects for the citing year. ψ_L is the coefficient of interest as it estimates the difference in citation rate for licensed patents relative to unlicensed patents. This regression is estimated using data on the post-licensing period only, using the post-period for unlicensed patents defined by their matched pair that was licensed.

The time series regression is estimated with the equation

$$CITES_{i,t} = f\big(\psi_{LP}\, LICENSED_{i,t} + \delta_{i,t} + \gamma_t + \varepsilon_{i,t}\big). \tag{2}$$

$LICENSED_{i,t}$ is a dummy variable that takes the value 1 if patent i in year t has been licensed and is 0 otherwise. This equation is estimated only for patents that are eventually licensed. ψ_{LP} is the coefficient of interest as it represents the change in citation rate for a patent that is licensed relative to its citation rate before it is licensed.

The cross-section and time series regressions are biased due to selection. I estimate these models for a sample of patents where licensed patents are matched to unlicensed patents based on their pre-licensing characteristics. This helps to partially correct for bias by balancing observed omitted variables. However, bias still clearly remains due to unobserved patent-level factors. A difference-in-difference approach more rigorously estimates causal effects by accounting for systematic time-invariant differences between licensed and unlicensed patents and patent-invariant differences between patents of different ages. The basic difference-in-difference regression is estimated with the equation

$$CITES_{i,t} = f\big(\psi_L\, EVER\_LICENSED_i + \psi_{LP}\, LICENSED_{i,t} + \psi_P\, POST_{i,t} + \delta_{i,t} + \gamma_t + \varepsilon_{i,t}\big). \tag{3}$$

$POST_{i,t}$ is a dummy variable equal to 1 if a licensed patent has been licensed by year t or if the matched licensed patent for a never licensed patent has been licensed by year t. This approach makes a strong assumption that matching finds comparable pairs, as contrasted with the more lenient assumption that matching achieves covariate balance overall between licensed and unlicensed patents (Rubin, 2006). A natural extension to this model is to add, α_i matched-pair fixed effects to account for the fixed characteristics of matched pairs:

$$CITES_{i,t} = f\left(\psi_L\, EVER\_LICENSED_i + \psi_{LP}\, LICENSED_{i,t} + \psi_P\, POST_{i,t} + \alpha_i + \delta_{i,t} + \gamma_t + \varepsilon_{i,t}\right). \tag{4}$$

A more rigorous approach to estimating a difference-in-difference regression relaxes the assumption of strict matched-pair design, instead relying on matching for covariate balance. In this approach, patent-level fixed effects account for all fixed unobserved heterogeneity of individual patents. This regression is estimated with the equation

$$CITES_{i,t} = f\left(\psi_{LP}\, LICENSED_{i,t} + \sigma_i + \delta_{i,t} + \gamma_t + \varepsilon_{i,t}\right) \tag{5}$$

where σ_i represent patent-level fixed effects. As before, ψ_{LP} is the coefficient of interest. However, in this equation, the coefficient represents the difference in citations from licensing relative to the change in citations for unlicensed patents at similar relative ages.

Finally, the difference-in-difference regression can be disaggregated to estimate yearly effects of licensing, meaning the effect of licensing on citations in specific years prior to and following licensing. This regression is estimated with the equation

$$CITES_{i,t} = f\left(\sum_{j=1\ldots10} \psi_{PRE_j}\, PRE\_LICENSE(j)_{i,t} + \sum_{k=1\ldots10} \psi_{POST_k}\, POST\_LICENSE(k)_{i,t} + \sigma_i + \delta_{i,t} + \gamma_t + \varepsilon_{i,t}\right) \tag{6}$$

which estimates ten coefficients on yearly difference-in-difference effects prior to licensing, ψ_{PRE_j}, and ten yearly coefficients after licensing, ψ_{POST_k}. Matching to improve balance in pre-licensing covariates should effectively make each estimated ψ_{PRE_j} close to zero, as these coefficients represent differences in citation rates prior to licensing.

Each of these regression models uses annual forward citations as the dependent variable. Most simply and easiest to interpret is a linear model which is well suited to applications with many fixed effects (the incidental parameter problem with non-linear models is not a concern with OLS). However, because citations are a right-skewed (Scherer and Harhoff, 2000) count variable, there are several options available to specify alternative functional forms. Non-linear models for count data, such as the negative binomial model, explicitly account for features of count data and can be more appropriate in some contexts. The negative binomial model is a more-flexible extension of the Poisson regression but cannot account for patents that never receive any citations (see Angrist and Pischke [2009] for a discussion of tradeoffs between linear models and non-linear models suited for different types of dependent variables). Finally, for comparability to other studies in the literature, a log-linear specification can also be estimated. To avoid the problem of the negative binomial model in accounting for never-cited patents, other studies in the patent literature transform the dependent variable by adding 1 before taking the log (Murray and Stern, 2007). For comparability to the previous literature, I present this functional form, but it is known to be problematic.

5. Results

This section presents the results from the application of the text analysis, matching, and regression frameworks presented in Section 4. First, I present the main regression results from applying Equations (1)-(6) to the preferred matched dataset of patents. I then examine the extent to which diffusion may be localized to one firm repeatedly innovating. Turning to mechanisms, one concern is that licensing could drive strategic patenting in technology areas that competitor firms now see as more desirable to enter. I account for this by looking at the citation rates of the citing patents themselves, noting that strategic patents are not cited often.

5.1. Diffusion after Licensing

The results of applying Equations (1)-(6) to the preferred matched dataset are presented in Table 6. Models (1)-(6) in Table 6 correspond to Equations (1)-(6). Model (1) shows that in the post-licensing period, licensed patents receive 0.344 (SE = 0.100) more citations than unlicensed patents. In this model and in models (2)-(4), the post-period for unlicensed patents is defined by the matched licensed patent. A cross-section regression of this form is typically biased because pre-treatment citation rates between licensed and unlicensed patents could differ. However, this is mitigated by the matching algorithm which used pre-treatment citations as a balancing covariate. Nevertheless, this regression may still suffer from other forms of omitted variable bias due to differences in licensed and unlicensed patents.

Model (2) in Table 6 shows that licensed patents receive 0.520 (SE = 0.155) additional citations per year after being licensed relative to their pre-license citation rate. Again this regression is biased if there are other factors that occur simultaneously at the time of licensing. Age and citing-year fixed effects do help reduce some of these concerns, but the lack of a suitable control group in this regression may bias this estimate.

Models (3) and (4) in Table 6 present difference-in-difference regressions that now control for heterogeneity in time-invariant differences between licensed and unlicensed patents as well as differences between pre-license and post-license periods. Model (4) adds fixed effects for each matched pair, which controls for a more specific layer of time-invariant heterogeneity. Because the matching algorithm already balanced licensed and unlicensed patents on pre-treatment citations, it is not surprising that the coefficient estimates in these regressions are similar to the results of Model (1) and that the coefficient on EVER_LICENSED is close to zero. In model (3), I estimate an effect of licensing of 0.328 (S.E. = 0.091) citations per year without matched-pair fixed effects and in model (4), I estimate an effect of licensing, which includes matched-pair fixed effects, of 0.335 (S.E. = 0.091).

Model (5) extends Model (4) by replaced matched-pair fixed effects with patent-level fixed effects. This relaxes the assumption of matching by allowing for separate estimates at the patent-level instead of the pair-level. As a result, EVER_LICENSED and POST drop and the coefficient on LICENSED remains the difference-in-difference coefficient of interest. I estimate that licensed patents receive 0.223 (SE = 0.066) additional citations after being licensed relative to unlicensed patents over similar time periods. This estimate represents a 31% increase in the citation rate for licensed patents before they are licensed (from 0.71 cites/year for eventually licensed patents before they are licensed).

Finally, Model (6) in Table 6 extends Model (5) by estimating yearly difference-in-difference effects of licensing. The coefficients presented in this model are each relative to the citation rate in the year the patent (or its matched pair) is licensed. The coefficient on the pre-licensing dummy variables are not distinguishable from zero (except for the year eight years prior to licensing, although this may be an anomaly), indicating that matching on pre-treatment citations worked. In the post-licensing period, the coefficient on the second year after licensing dummy is also not distinguishable from zero. However, beginning in the third year after licensing, the difference-in-difference rate is statistically significant. This effect holds for years three through eight after licensing (although the dummy on seven years after licensing is not statistically significant). During this period, citations are between 0.253 – 0.465 cites/year higher to licensed patents relative to unlicensed patents of comparable age. This represents a 36 – 65% increase in the citation rate for licensed patents due to licensing. Nine to ten years after licensing, the estimated difference-in-difference coefficient drops, but this is likely due to insufficient data on patents that have been licensed for this length of time. The surprisingly negative coefficient on the ten year post-licensing dummy is also the least precisely estimated of the post-period dummies. The results of Model (6) are presented graphically in Figure 7.

I assess sensitivity to functional form in Table 7. Models (1) and (2) in Table 7 replicate Models (5) and (6) in Table 6 for comparison. Models (3) and (4) in Table 7 apply a negative binomial functional form to Equations (5) and (6). Because of restrictions in the negative binomial model, patents with no citations over the panel are dropped, reducing the number of patent-year observations from 9,357 to 7,640. Nevertheless, the estimated coefficients in the negative binomial regressions closely match the OLS results in sign and statistical significance. Models (5) and (6) present results from an augmented log-linear regression where the dependent variable is transformed by the function $\log(CITES + 1)$ to account for its skewness. Again, results match closely to the OLS results in both sign and statistical significance.

Next, I assess sensitivity to the matching algorithm. I rely heavily on matching for causal identification, so assessing sensitivity to different decisions in the matching algorithm is important. Table 8 shows results from five matching algorithms applied to estimating Equation (5). Figure 8 displays these estimates graphically and Figure 9 displays the results of the thirteen matching approaches applied to estimating Equation 6. Model (2) in the table is the "preferred specification" and replicates Model (5) in Table 6. The other models implement matching algorithms as described in Section 4.1. Overall, the choice of number of topics does not appear to affect the results as much as choice of what to do with the topics (matching on estimated hazard, Mahalanobis distance, or CEM on topic peaks). Other than the estimate in Model (6) for the 50-topic model with CEM on topic peaks with ties resolved by the estimated hazard, the estimates in models (9) – (13) which rely on exact matching on primary USPTO classification are systematically larger than the other estimates. These estimates also have larger standard errors, most likely due to having dropped a large number of observations to create exact matches on the USPTO classes. These findings are generally consistent with the Rosenbaum bounds approach shown in Table 5. Overall, the range of estimates across matching approaches represent a 31 – 48% increase in the citation rate to licensed patents after licensing. The thirteen approaches to matching produce similar estimates, and each approach confirms the overall qualitative thesis of this paper. In fact, the preferred specification that I focus on is at the lower end of this range of estimates.

5.2. Exclusive vs. Non-Exclusive Licenses

I collected information from the Labs on which patents were exclusively licensed or non-exclusively licensed. In general, Lab technology transfer offices prefer to offer less exclusive licenses so as to increase the number of users who can access their technologies, whereas licensees prefer greater exclusivity so as to protect their right. In my sample, 49% (430 of 877 patents in the full sample) of licensed patents are exclusively licensed. Table 9 adds in a separate variable that interacts the LICENSED dummy with a dummy indicating whether the license was on an exclusive basis. The coefficient on the interaction term is positive but not statistically significant, suggesting that greater exclusivity of licensing only marginally increases the rate of innovation spillovers. Because non-exclusive licenses still typically are exclusive within their field of use or within a geographic region, this finding is not completely surprising.

In terms of mechanisms, as described in Section X, a license agreement can cause greater follow-on innovation either through a signaling effect or a learning-by-doing effect. A non-exclusive license and exclusive license would seem to suggest an equal signal but provide greater incentives for learning-by-doing, as a firm is likely to have a greater market share under exclusive licensing. This result suggests (weakly) that the learning-by-doing effect is greater than the signaling effect as exclusive licensing adds approximately 50% additional citations.

5.3. Breadth vs. Concentration of Spillovers

When follow-on innovation occurs only within a licensing firm, there is no positive externality associated with licensing. Therefore, the broad diffusion of follow-on innovation is important to understand to evaluate the impact of licensing.

Licensing grants a single firm the right to commercialize a patented invention in a field of use. One concern with the empirical findings presented in Section 5.1 is that the increased citation rate to licensed patents could be driven by the licensing firm repeatedly developing follow-on inventions from its licensed patent. This would suggest that while licensing leads to follow-on innovation, benefits from induced innovations would be captured by the licensing firm exclusively. To investigate this concern, I create a new dependent variable that counts citations only from patents with assignees who have not already filed a patent citing this same patent. For example, if a patent is cited five times by a single firm, I only count this as one unique citation. Table 10 presents results from regressions with this modified dependent variable. Model (1) replicates Model (5) from Table 6 for comparison. Model (2) uses the dependent variable of only citations from first-time citers. This model estimate positive and significant difference-in-difference coefficients, showing that knowledge diffusion from licensing does occur beyond just the licensing firm.

Comparing the magnitudes of the coefficients in Model (1) to Model (2) in Table10 reveals that citations are still concentrated in assignees to a degree. Although I am not able to observe which citing assignees are the licensees, the most conservative interpretation would be to assume that all citations from repeated assignees could be from the licensee. Therefore, comparing these coefficients suggests that at least 76% (0.169 / 0.223) of the estimated effect in Model (1) is driven by citations from assignees other than the licensing firm.

5.4. Accounting for Strategic Patenting and Signaling

Sections 5.1 – 5.3 have established that licensing increases the rate of citation to licensed patents relative to unlicensed patents of comparable age. The increased rate of citations may not necessarily suggest that knowledge is diffusing if citations are accruing to licensed patents for strategic reasons. For example, when a National Lab patent is licensed to a firm, competitor firms may take this as a signal that firms are moving in a certain technological direction. In response these firms could file defensive or "strategic" patents in the area to protect their competitive positions rather than take the information signal as a useful indicator of technological promise, as suggested by Drivas et al. (2014). Strategic patents are for defensive purposes and do not represent actual knowledge spillovers, yet they still may drive increased citations to the licensed patent – these citations represent demarcations of prior art rather than inspiration for follow-on innovation. Previous studies have found that strategic patents are cited less often than patents that represent novel inventions (Harhoff et al., 2003). Therefore, to investigate the mechanisms underlying the results presented previously, I define two new independent variables. First, I create a new variable of citations that only counts citations from patents that themselves have been cited at least once. Second, I create a variable of citations that only counts citations from patents that have received at least the median annual rate of citations. This effectively drops all citation counts in the dependent variable from patents that received fewer than 0.27 citations per year.

Table 11 presents the results of regressions applying Equation (5) to three dependent variables: all citations, citations only from patents cited at least once, and citations only from patents with at least median citations per year. Model (1) in Table 11 again repeats Model (5) in Table 6. In my sample, 35% of citing patents are themselves never cited. If never-cited patents were proportionally represented as citers to all patents, then the estimate in Model (2) should be approximately 35% less than the estimate in Model (1). The estimate from Model (2) is 33% less than the estimate from

Model (1), suggesting that never-cited patents cite licensed patents approximately just as often as average. A similar comparison for Models (3) and (1) can be made. By definition, 50% of citing patents are cited less than the median of 0.27 cites per year. Therefore, if above-the-median citing patents were evenly distributed, the expected coefficient in Model (3) would be half of the coefficient in Model (1). Table 7 shows that the coefficient in Model (6) is indeed 50% smaller than the coefficient in Model (1). Taken together, these results provide evidence that the additional citations to licensed patents due to having been licensed are from patents that are representative of average citing patents and do not tend to be less often cited themselves. Given that strategic patents are cited less often, this provides evidence against a strong strategic patenting effect.

6. Discussion and Conclusion

Public innovation policy is oriented toward enhancing basic scientific understanding and developing fundamental inventions without immediate commercial application. Yet public R&D has led to new technologies, such as the Internet, GPS, and radar, which have dramatically altered the economy and improved well-being. While initial investment in these inventions relied on public support, development of the majority of these inventions into the products and processes that made these inventions revolutionary required large private investment.

Beginning with the Stevenson-Wydler and Bayh-Dole Acts of 1980, national policy reforms over the past 30 years have greatly increased the intellectual property protection surrounding publicly sponsored inventions, decreasing public access to utilizing these technologies. Yet despite this decreased access, the diffusion of technological knowledge created by public R&D funding has not similarly decreased.

In this paper, I have shown that in the context of five U.S. National Labs, technology transfer agreements that license patents to private firms have increased the rate of spillovers from publicly sponsored inventions. This empirical finding, corroborated through a variety of statistical methods to account for unobserved heterogeneity and selection bias, provides new evidence for the role of intellectual property protection in increasing the benefits to publicly funded R&D. By making technological knowledge appropriable, patenting allows public institutions to transfer the right to utilize a public invention for the purpose of commercializing the technology in the "market for ideas" (Gans and Stern, 2010). In this paper, I find evidence to support that the incentives to commercialize a licensed technology leads to a net positive change in follow-on invention and that this effect is driven not just by the licensing firm, but by invention in firms that did not have access to the licensed technology but may have still gained experience with downstream products and processes. This implies that at least for the case of publicly sponsored inventions, the spillover effects of licensing a patent cannot be fully appropriated by the licensing firm.

This research is also an important improvement on existing approaches to evaluating technology transfer, which have often relied on short-term and easy-to-measure metrics. Recent policy documents have called for improved measurement of technology transfer efforts, and the focus of this paper on measuring social impacts of technology transfer through its effect on follow-on innovation would be a useful contribution to this policy discussion. (U.S. Government Accountability Office, 2009; The White House, 2011, 2014; Stepp et al., 2013)

Methodologically, the matching algorithms presented in this paper could be usefully extended to study other problems in the innovation literature. In this work, I have focused on patent abstracts to find comparable patents, but future work could apply text classification to the claims within patents to understand which patents draw on more diverse prior art or establish new technologies that are cross-disciplinary. This would be a particularly useful extension of this mixed-membership model.

Clearly, not all innovations (and not even all important innovations) are captured by patents. While the results presented in this paper apply to patented inventions, the recurrent problems associated with drawing inferences about innovation more broadly from patent data persist. However, the context for this work is over transactions in the market for innovations, and these markets rely on patents heavily to develop contractible assets. Therefore, the conclusions of this work do have important implications for understanding the general dynamics of innovation and knowledge diffusion.

Licensing a technology to a private firm requires IP protection. In this paper, I show that licensing an already patented invention increases the rate of knowledge diffusion. The magnitude of the effect I estimate can be usefully compared to other studies that have studied the effect of patenting on knowledge diffusion to provide insight on whether patenting and licensing considered together positively or negatively effects knowledge diffusion relative to putting a publicly discovered invention into the public domain. Galasso and Schankerman (2013) estimate that removing patent protection on highly valuable patents increases the citation rate to these patents by 50%. Murray and Stern (2007) find that citations to academic papers decline 10-20% when a patent that covers the same invention disclosed in the paper is granted[17]. The estimates I present in this paper would suggest that licensing cuts the effect of patenting on knowledge diffusion impediment by more than half. This suggests that for the sole objective of increasing knowledge diffusion, patented inventions should be licensed but unpatented inventions should not be patented because diffusion effects are even higher for inventions in the public domain. However, from the perspective of maximizing the net social benefit of public R&D over time, policy must assess the tradeoff between developing discovered inventions for use in the short-run through greater IP protection with greater knowledge diffusion to create new inventions in the long-run. At the margin, technology transfer is a win-win policy as it both drives private investment into

[17] While citations to papers and patents are very different, the magnitude of the effect is still a useful benchmark.

developing technologies in the short-run and also increases the rate of follow-on innovation for the long-run.

References

113th Congress (2014). *H.R. 5120 Department of Energy Laboratory Modernization and Technology Transfer Act of 2014.* Available at: http://thomas.loc.gov/cgi-bin/bdquery/z?d113:hr5120:

Abadie A., and G. Imbens (2006). Large Sample Properties of Matching Estimators for Average Treatment Effects *Econometrica.*

Adams J.D., E.P. Chiang, and J.L. Jensen (2003). The Influence of Federal Laboratory R&D on Industrial Research *Review of Economics and Statistics*.

Alexopoulos M. (2011). Read All about It!! What Happens Following a Technology Shock? *American Economic Review*.

Angrist J.D., and J.-S. Pischke (2009). *Mostly Harmless Econometrics an Empiricist's Companion*. Princeton University Press, Princeton, (ISBN: 9781400829828 1400829828 1282608096 9781282608092).

Arora A. (1995). Licensing Tacit Knowledge: Intellectual Property Rights And The Market For Know-How *Economics of Innovation and New Technology*.

Arora A., A. Fosfuri, and A. Gambardella (2004). *Markets for Technology: The Economics of Innovation and Corporate Strategy*. MIT, Cambridge, Mass.; London, (ISBN: 0262511819 9780262511810 0262011905 9780262011907).

Arrow K. (1962). Economic Welfare and the Allocation of Resources for Invention *The Rate and Direction of Inventive Activity: Economic and Social Factors*.

Arthur W.B. (2009). *The Nature of Technology: What It Is and How It Evolves*. Free Press, (ISBN: 978-1416544050).

Black D.A., and J.A. Smith (2004). How Robust Is the Evidence on the Effects of College Quality? Evidence from Matching *Journal of Econometrics*.

Blei D. (2010). Introduction to Probabilistic Topic Models. Available at: cs.princeton.edu/~blei/papers/Blei2011.pdf.

Blei D., and J.D. Lafferty (2007). A Correlated Topic Model of Science *The Annals of Applied Statistics*.

Blei D., A. Ng, and M. Jordan (2003). Latent Dirichlet Allocation *Journal of Machine Learning Research*.

Bozeman B. (2000). Technology Transfer and Public Policy: A Review of Research and Theory *Research Policy*.

Cannady C. (2013). *Technology Licensing and Development Agreements*. Oxford University Press, Oxford [UK] ; New York, (ISBN: 9780195385137).

Cohen L., and R. Noll (1996). The Future of the National Laboratories *Proceedings of the National Academy of Sciences*.

Cox D.R. (1984). *Analysis of Survival Data*. Chapman and Hall, London ; New York, 201 pp., (ISBN: 041224490X).

DOE Office of Energy Efficiency and Renewable Energy (2012). *FY 2009 - 2012 DOE Technology Transfer Data*.

DOE Office of Energy Efficiency and Renewable Energy (2014). *Energy Innovation Portal*. Available at: http://techportal.eere.energy.gov/.

DOE Office of Science (2013). The Office of Science Laboratories. Available at: http://science.energy.gov/laboratories/.

DOE Office of Scientific and Technical Information (2014a). *DOepatents*. Available at: http://www.osti.gov/doepatents/.

DOE Office of Scientific and Technical Information (2014b). *Department of Energy National Laboratories*. Available at: http://www.osti.gov/accomplishments/nuggets/natlabsA-L.html.

DOE Technology Transfer Working Group (2013). *Licensing Guide and Sample License*. Available at: http://technologytransfer.energy.gov/LicensingGuideFINAL.pdf.

Drivas K., Z. Lei, and B. Wright (2014). Academic Patent Licenses: Roadblocks or Signposts for Nonlicensee Cumulative Innovation? *SSRN Working Paper*.

Federal Laboratory Consortium for Technology Transfer (2011). *Technology Transfer Desk Reference: A Comprehensive Guide to Technology Transfer*. Available at: https://secure.federallabs.org/pdf/T2_Desk_Reference.pdf.

Feinerer I., K. Hornik, and D. Meyer (2008). Text Mining Infrastructure in R *Journal of Statistical Software*.

Fleming L. (2001). Recombinant Uncertainty in Technological Search *Management Science*.

Furman J.L., and S. Stern (2011). Climbing atop the Shoulders of Giants: The Impact of Institutions on Cumulative Research *American Economic Review*.

Galasso A., and M. Schankerman (2013). Patents and Cumulative Innovation: Causal Evidence from the Courts *Rotman School of Management Working Paper*.

Gans J.S., and S. Stern (2010). Is There a Market for Ideas? *Industrial and Corporate Change*.

Google (2014). *Google Patents*. Available at: www.google.com/patents.

Green J., and S. Scotchmer (1995). On the Division of Profit in Sequential Innovation *RAND Journal of Economics*.

Grimmer J., and B.M. Stewart (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts *Political Analysis*.

Gruen B., and K. Hornik (2011). *Topicmodels: Topic Models. R Package Version 0.0-8.* Available at: http://CRAN.R-project.org/package=topicmodels.

Harhoff D., F.M. Scherer, and K. Vopel (2003). Citations, Family Size, Opposition and the Value of Patent Rights *Research Policy*.

Hausman N. (2010). Effects of University Innovation on Local Economic Growth and Entrepreneurship.

Heckman J., and S. Navarro-Lozano (2004). Using Matching, Instrumental Variables, and Control Functions to Estimate Economic Choice Models *Review of Economics and Statistics*.

Hellmann T. (2007). The Role of Patents for Bridging the Science to Market Gap *Journal of Economic Behavior & Organization*.

Henderson R., A.B. Jaffe, and M. Trajtenberg (1998). Universities as a Source of Commercial Technology: A Detailed Analysis of University Patenting, 1965–1988 *Review of Economics and Statistics*.

Hirabayashi J. (2003).. Revisiting the USPTO Concordance Between the US Patent Classification and the Standard Industrial Classification Systems *WIPO-OECD Workshop on Statistics in the Patent Field*.

Hoberg G., and G. Phillips (2010). Dynamic Text-Based Industries and Endogenous Product Differentiation *NBER Working Paper Series*.

Hopkins D., and G. King (2010). A Method of Automated Nonparametric Content Analysis for Social Science *American Journal of Political Science*.

Hornik K. (2007). *Snowball: Snowball Stemmers. R Package Version 0.0-1.*

Iacus S.M., G. King, and G. Porro (2012). Causal Inference without Balance Checking: Coarsened Exact Matching *Political Analysis*.

Imbens G.W. (2004). Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review *Review of Economics and Statistics*.

Jaffe A., and J. Lerner (2001). Reinventing Public R&D: Patent Policy and the Commercialization of National Laboratory Technologies *The RAND Journal of Economics*.

Jaffe A.B., and M. Trajtenberg (1996). Flows of Knowledge from Universities and Federal Laboratories: Modeling the Flow of Patent Citations over Time and across Institutional and Geographic Boundaries *Proceedings of the National Academy of Sciences*.

Jaffe A.B., and M. Trajtenberg (1999). International Knowledge Flows: Evidence From Patent Citations *Economics of Innovation and New Technology*.

Jaffe A., and M. Trajtenberg (Eds.) **(2002).** *Patents, Citations, and Innovations*. MIT Press.

Jaffe A.B., M. Trajtenberg, and M.S. Fogarty (2000). Knowledge Spillovers and Patent Citations: Evidence from a Survey of Inventors *American Economic Review*.

Jaffe A., M. Trajtenberg, and R. Henderson (1993). Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations *The Quarterly Journal of Economics*.

Kaplan S., and K. Vakili (2014). The Double-Edged Sword of Recombination in Breakthrough Innovation *Strategic Management Journal*.

King G., J. Pan, and M.E. Roberts (2013). How Censorship in China Allows Government Criticism but Silences Collective Expression *American Political Science Review*.

LBNL, Innovation and Partnerships Office (2014). *Licensing Overview*. Available at: http://ipo.lbl.gov/for-industry/licensing-berkeley-lab-technologies/.

Logar N., V. Narayanamurti, and L.D. Anadón (2014). Reforming U.S. Energy Innovation Institutions: Maximizing the Return on Investment *Transforming U.S. Energy Innovation*.

Margolis R.M., and D.M. Kammen (1999). Evidence of under-Investment in Energy R&D in the United States and the Impact of Federal Policy *Energy Policy*.

Merton R.K. (1973). *The Sociology of Science: Theoretical and Empirical Investigations*. University of Chicago Press, Chicago, (ISBN: 0226520919 9780226520919 0226520927 9780226520926).

Mostellar F., and D. Wallace (1964). *Inference and Disputed Authorship*. University of Chicago Press.

Mowery D., N. Richard, S. Bhaven, and A. Ziedonis (2001). The Growth of Patenting and Licensing by U.S. Universities: An Assessment of the Effects of the Bayh-Dole Act of 1980 *Research Policy*.

Mowery D.C., B.N. Sampat, and A.A. Ziedonis (2002). Learning to Patent: Institutional Experience, Learning, and the Characteristics of U.S. University Patents After the Bayh-Dole Act, 1981-1992 *Management Science*.

Murray F., and S. Stern (2007). Do Formal Intellectual Property Rights Hinder the Free Flow of Scientific Knowledge? *Journal of Economic Behavior & Organization*.

National Institute of Standards and Technology, and G. Anderson (2011). *Federal Technology Transfer Data*. Available at: http://www.nist.gov/tpo/publications/federal-laboratory-techtransfer-reports.cfm.

National Science Board (2014). *Science and Engineering Indicators 2014*. National Science Foundation, Arlington, VA.

Nelson R. (1959). The Simple Economics of Basic Scientific Research *The Journal of Political Economy*.

Partha D., and P.A. David (1994). Toward a New Economics of Science *Research Policy*.

PNNL, Technology Transfer (2014). *Our Licensing Guidelines*. Pacific Northwest National Lab. Available at: http://www.pnl.gov/business/techtransfer/guidelines.asp.

Quinn K.M., B.L. Monroe, M. Colaresi, M.H. Crespin, and D.R. Radev (2010). How to Analyze Political Attention with Minimal Assumptions and Costs *American Journal of Political Science*.

Rooksby J. (2013). Innovation and Litigation: Tensions Between Universities and Patents and How to Fix Them *Yale Journal of Law & Technology*.

Rosenbaum P. (2002). *Observational Studies*. Springer, New York, NY.

Rosenbaum P. (2005). Observational Study *Encyclopedia of Statistics in Behavioral Science*.

Rosenberg N. (1982). *Inside the Black Box: Technology and Economics*. Cambridge University Press, Cambridge ; New York, 304 pp., (ISBN: 0521248086).

Rubin D. (2006). *Matched Sampling for Causal Effects*. Cambridge University Press, Cambridge, UK.

Scherer F.M. (1982). Inter-Industry Technology Flows and Productivity Growth *The Review of Economics and Statistics*.

Scherer F.., and D. Harhoff (2000). Technology Policy for a World of Skew-Distributed Outcomes *Research Policy*.

Scotchmer S. (1991). Standing on the Shoulders of Giants: Cumulative Research and the Patent Law *Journal of Economic Perspectives*.

Singh J., and A. Agrawal (2011). Recruiting for Ideas: How Firms Exploit the Prior Inventions of New Hires *Management Science*.

Stange K. (2011). A Longitudinal Analysis of the Relationship Between Fertility Timing and Schooling *Demography*.

Stepp M., S. Pool, N. Loris, and J. Spencer (2013). *Turning the Page: Reimagining the National Labs in the 21st Century Innovation Economy*. The Information Technology and Innovation Foundation, The Center for American Progress, and The Heritage Foundation. Available at: http://cdn.americanprogress.org/wp-content/uploads/2013/06/2013-turning-the-page-3.pdf.

Stevenson-Wydler Technology Innovation Act of 1980 (1980).

The White House (2011). *Presidential Memorandum - Accelerating Technology Transfer and Commercialization of Federal Research in Support of High-Growth Businesses*. Office of the Press Secretary. Available at: http://www.whitehouse.gov/the-press-office/2011/10/28/presidential-memorandum-accelerating-technology-transfer-and-commerciali.

The White House (2014). *The President's Management Agenda for Fiscal Year 2015: Opportunity for All: Creating a 21st Century Government*. Available at: http://www.whitehouse.gov/sites/default/files/omb/budget/fy2015/assets/fact_sheets/creating-a-21st-century-government.pdf.

Thompson P., and M. Fox-Kean (2005). Patent Citations and the Geography of Knowledge Spillovers: A Reassessment *The American Economic Review*.

Turner A. (2004). *PNNL's Body Scanner Garners Federal Commercialization Award*. Pacific Northwest National Lab. Available at: http://www.pnl.gov/news/archives/2004/04-70.htm.

U.S. Government Accountability Office (2009). *Technology Transfer: Clearer Priorities and Greater Use of Innovative Approaches Could Increase the Effectiveness of Technology Transfer at Department of Energy Laboratories*. Available at: http://www.gao.gov/assets/300/290963.pdf.

U.S. Patent and Trademark Office (2014). *Patent Full-Text Databases*. Available at: http://patft.uspto.gov/netahtml/PTO/.

Venugopalan S., and V. Rai (2014). Topic Based Classification and Pattern Identification in Patents *Technological Forecasting and Social Change*.

Weitzman M. (1998). Recombinant Growth *The Quarterly Journal of Economics*.

Westwick P. (2003). *The National Labs: Science in an American System, 1947-1974*. Harvard University Press, (ISBN: 978-0674009486).

Williams H. (2013). Intellectual Property Rights and Innovation: Evidence from the Human Genome *Journal of Political Economy*.

Wright B.D., K. Drivas, Z. Lei, and S.A. Merrill (2014). Technology Transfer: Industry-Funded Academic Inventions Boost Innovation *Nature*.

Figures and Tables

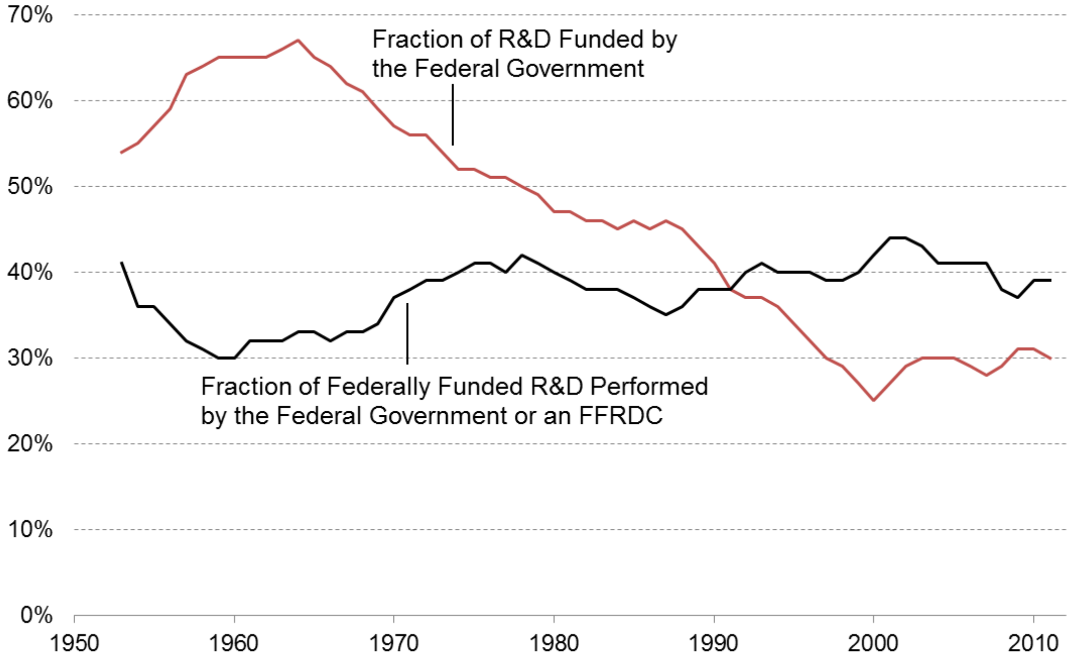


Figure 1. Time series of federal funding for R&D. The red line shows the fraction of R&D funded by the federal government from 1953 – 2011. The black line shows the fraction of federally funded R&D that is performed by the federal government directly or by a federally funded R&D center (FFRDC). Data and definitions adopted from the National Science Board (2014).

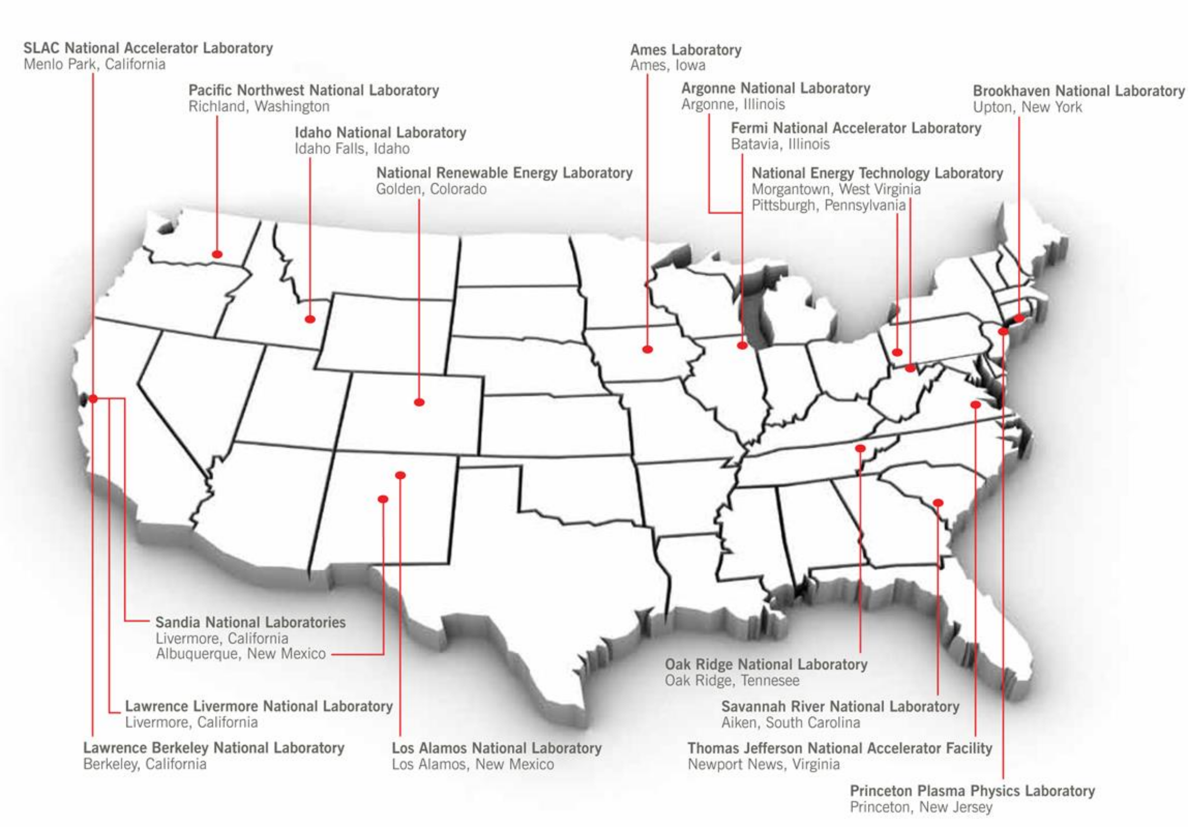

SLAC National Accelerator Laboratory
Menlo Park, California
Pacific Northwest National Laboratory
Richland, Washington
Idaho National Laboratory
Idaho Falls, Idaho
National Renewable Energy Laboratory
Golden, Colorado
Ames Laboratory
Ames, Iowa
Argonne National Laboratory
Argonne, Illinois
Fermi National Accelerator Laboratory
Batavia, Illinois
National Energy Technology Laboratory
Morgantown, West Virginia
Pittsburgh, Pennsylvania
Brookhaven National Laboratory
Upton, New York
Sandia National Laboratories
Livermore, California
Albuquerque, New Mexico
Lawrence Livermore National Laboratory
Livermore, California
Lawrence Berkeley National Laboratory
Berkeley, California
Los Alamos National Laboratory
Los Alamos, New Mexico
Oak Ridge National Laboratory
Oak Ridge, Tennesee
Savannah River National Laboratory
Aiken, South Carolina
Thomas Jefferson National Accelerator Facility
Newport News, Virginia
Princeton Plasma Physics Laboratory
Princeton, New Jersey

Figure 2. Location of the 17 U.S. National Labs under DOE. The U.S. Department of Energy is the steward for seventeen National Labs. Within DOE, ten of the labs are managed by the Office of Science, three are managed by the National Nuclear Security Administration, and the remaining four are managed by the Office of Energy Efficiency & Renewable Energy, the Office of Fossil Energy, the Office of Nuclear Energy, and the Office of Environmental Management. The five labs that I study in this paper are the Brookhaven, Lawrence Berkeley, Pacific Northwest, Sandia, and National Energy Technology Laboratories. Figure credit: DOE Office of Science (2013).

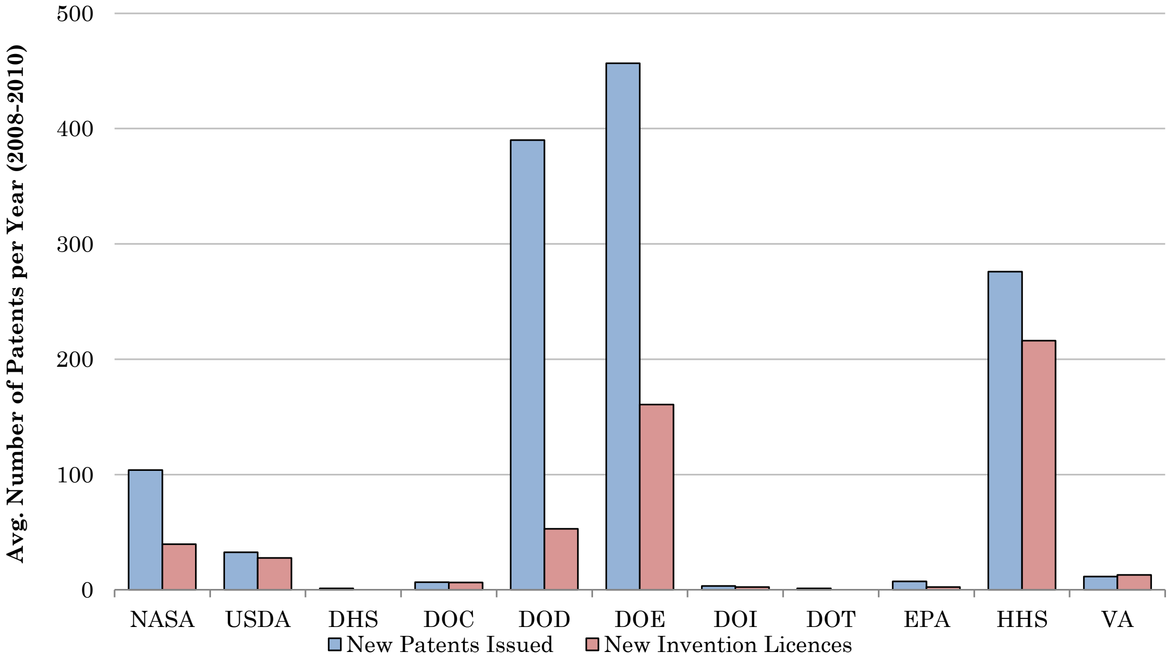


Figure 3. Patenting and Licensing of Federally Funded R&D by Sponsoring Agency (2008 – 2010). Data collected from federal agencies by the National Institute of Standards and Technology (2011).

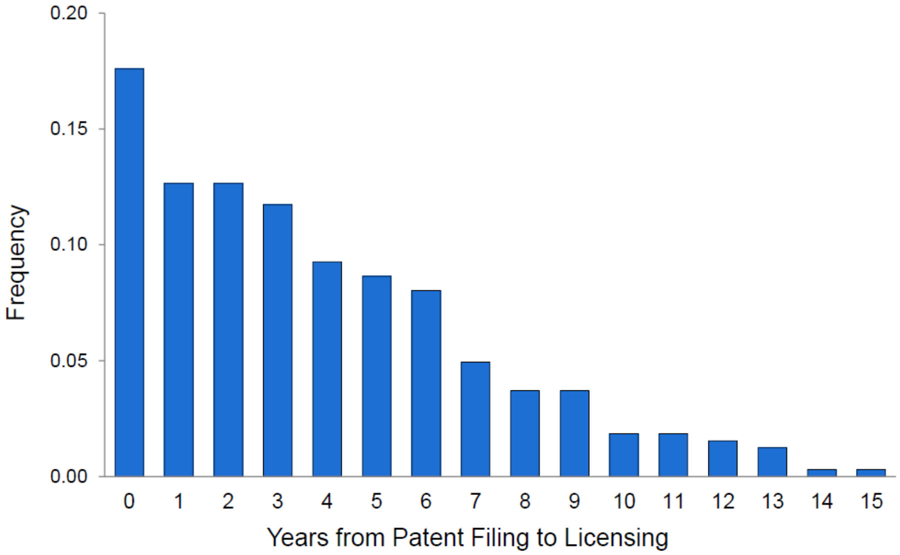
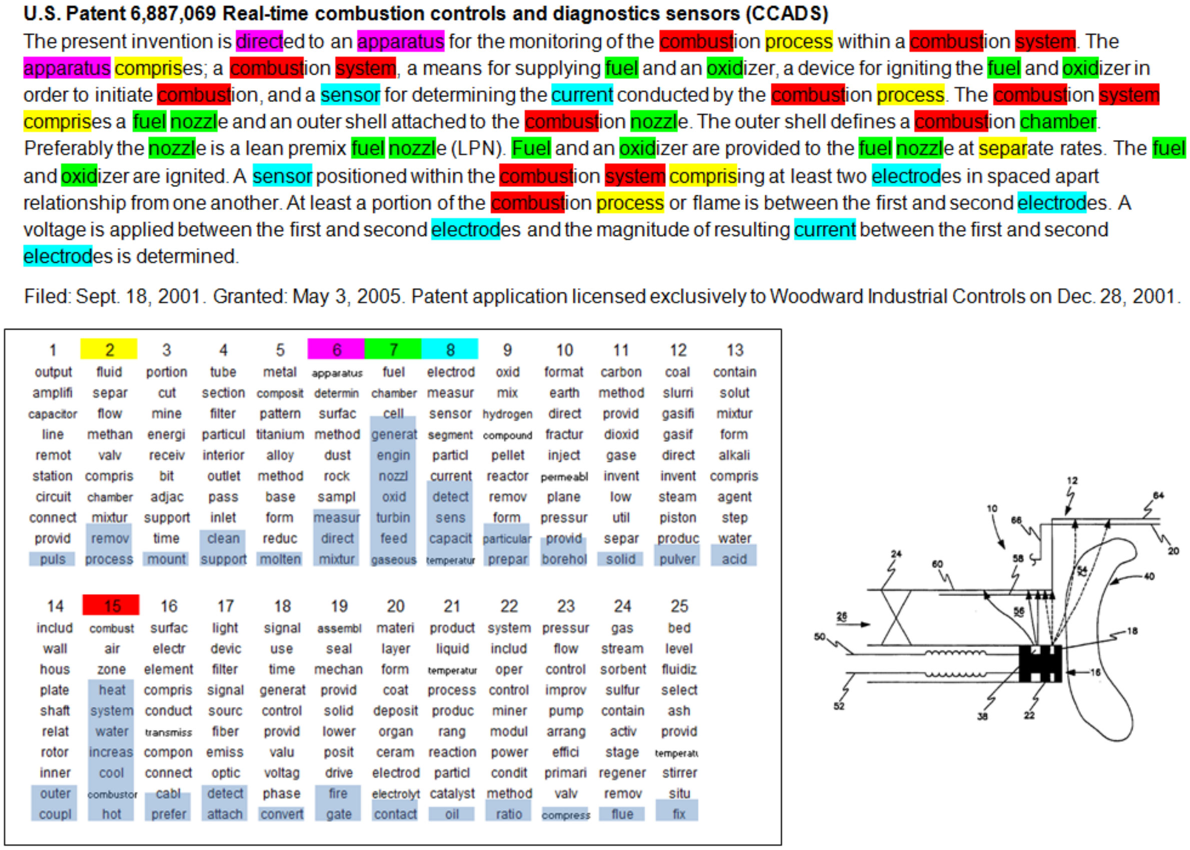


Figure 4. Distribution in the Lag between Patent Filing and Licensing



U.S. Patent 6,887,069 Real-time combustion controls and diagnostics sensors (CCADS)

The present invention is directed to an apparatus for the monitoring of the combustion process within a combustion system. The apparatus comprises; a combustion system, a means for supplying fuel and an oxidizer, a device for igniting the fuel and oxidizer in order to initiate combustion, and a sensor for determining the current conducted by the combustion process. The combustion system comprises a fuel nozzle and an outer shell attached to the combustion nozzle. The outer shell defines a combustion chamber. Preferably the nozzle is a lean premix fuel nozzle (LPN). Fuel and an oxidizer are provided to the fuel nozzle at separate rates. The fuel and oxidizer are ignited. A sensor positioned within the combustion system comprising at least two electrodes in spaced apart relationship from one another. At least a portion of the combustion process or flame is between the first and second electrodes. A voltage is applied between the first and second electrodes and the magnitude of resulting current between the first and second electrodes is determined.

Filed: Sept. 18, 2001. Granted: May 3, 2005. Patent application licensed exclusively to Woodward Industrial Controls on Dec. 28, 2001.

1	2	3	4	5	6	7	8	9	10	11	12	13
output	fluid	portion	tube	metal	apparatus	fuel	electrod	oxid	format	carbon	coal	contain
amplifi	separ	cut	section	composit	determin	chamber	measur	mix	earth	method	slurri	solut
capacitor	flow	mine	filter	pattern	surfac	cell	sensor	hydrogen	direct	provid	gasifi	mixtur
line	methan	energi	particul	titanium	method	generat	segment	compound	fractur	dioxid	gasif	form
remot	valv	receiv	interior	alloy	dust	engin	particl	pellet	inject	gase	direct	alkali
station	compris	bit	outlet	method	rock	nozzl	current	reactor	permeabl	invent	invent	compris
circuit	chamber	adjac	pass	base	sampl	oxid	detect	remov	plane	low	steam	agent
connect	mixtur	support	inlet	form	measur	turbin	sens	form	pressur	util	piston	step
provid	remov	time	clean	reduc	direct	feed	capacit	particular	provid	separ	produc	water
puls	process	mount	support	molten	mixtur	gaseous	temperatur	prepar	borehol	solid	pulver	acid

14	15	16	17	18	19	20	21	22	23	24	25
includ	combust	surfac	light	signal	assembl	materi	product	system	pressur	gas	bed
wall	air	electr	devic	use	seal	layer	liquid	includ	flow	stream	level
hous	zone	element	filter	time	mechan	form	temperatur	oper	control	sorbent	fluidiz
plate	heat	compris	signal	generat	provid	coat	process	control	improv	sulfur	select
shaft	system	conduct	sourc	control	solid	deposit	produc	miner	pump	contain	ash
relat	water	transmiss	fiber	provid	lower	organ	rang	modul	arrang	activ	provid
rotor	increas	compon	emiss	valu	posit	ceram	reaction	power	effici	stage	temperatu
inner	cool	connect	optic	voltag	drive	electrod	particl	condit	primari	regener	stirrer
outer	combustor	cabl	detect	phase	fire	electrolyt	catalyst	method	valv	remov	situ
coupl	hot	prefer	attach	convert	gate	contact	oil	ratio	compress	flue	fix

Figure 5. Example classification of a patent using the LDA model. In this example, I show how U.S. patent 6,887,069 was classified under and LDA model with 25 topics applied to the subgroup of patents in my dataset from the National Energy Technology Lab. The patent abstract is shown at the top, with individual words highlighted in colors corresponding to the five most frequent topics for this document. In the box below the abstract, the topic distributions for the corpus are shown with words within topics arranged in descending order of likelihood to occur within the topic. Overlaid on top of the word-within-topic distributions is the distribution of topics within this particular patent, shown in blue bars. This distribution is the key output of the LDA model for my analysis as it gives a continuous measure of the patent abstract’s substantive content.

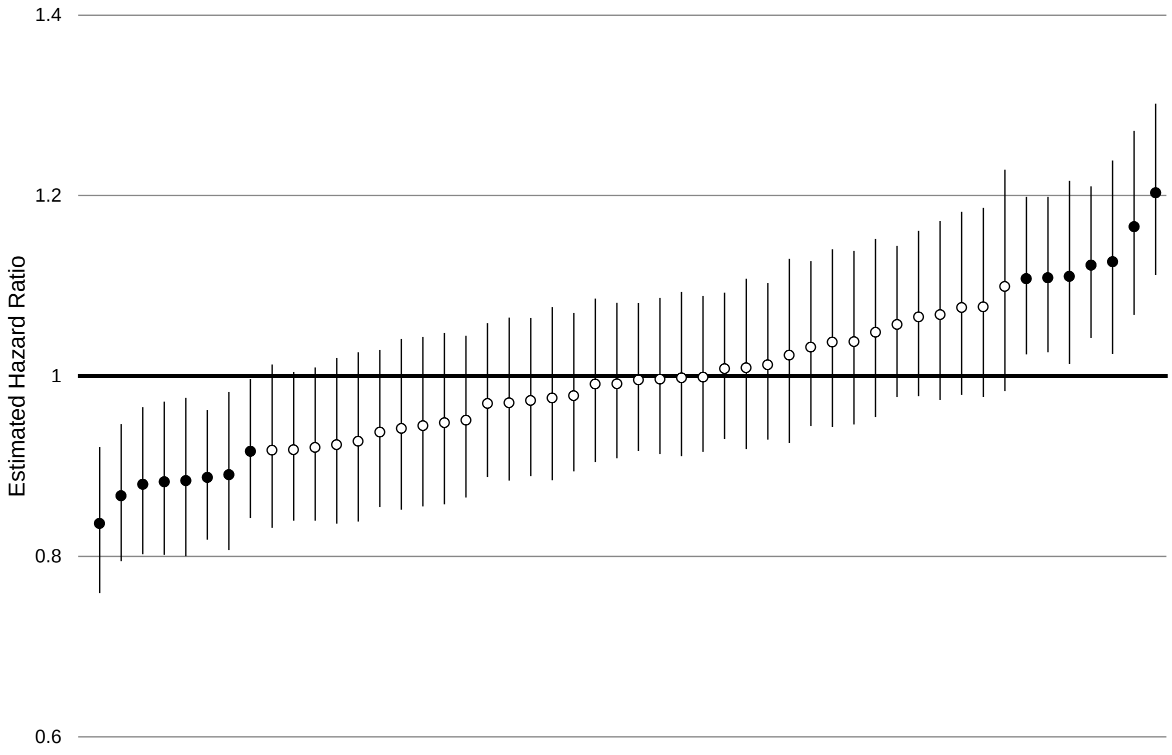


Figure 6. Estimated Hazard Ratios for Each Topic. Estimated hazard ratios can be used to predict how likely a given patent is to be licensed. Topics are arranged from lowest to highest hazard, where HR = 1 means no effect on probability of license. These estimates inform the predicted HR's for each patent, the key input for matching

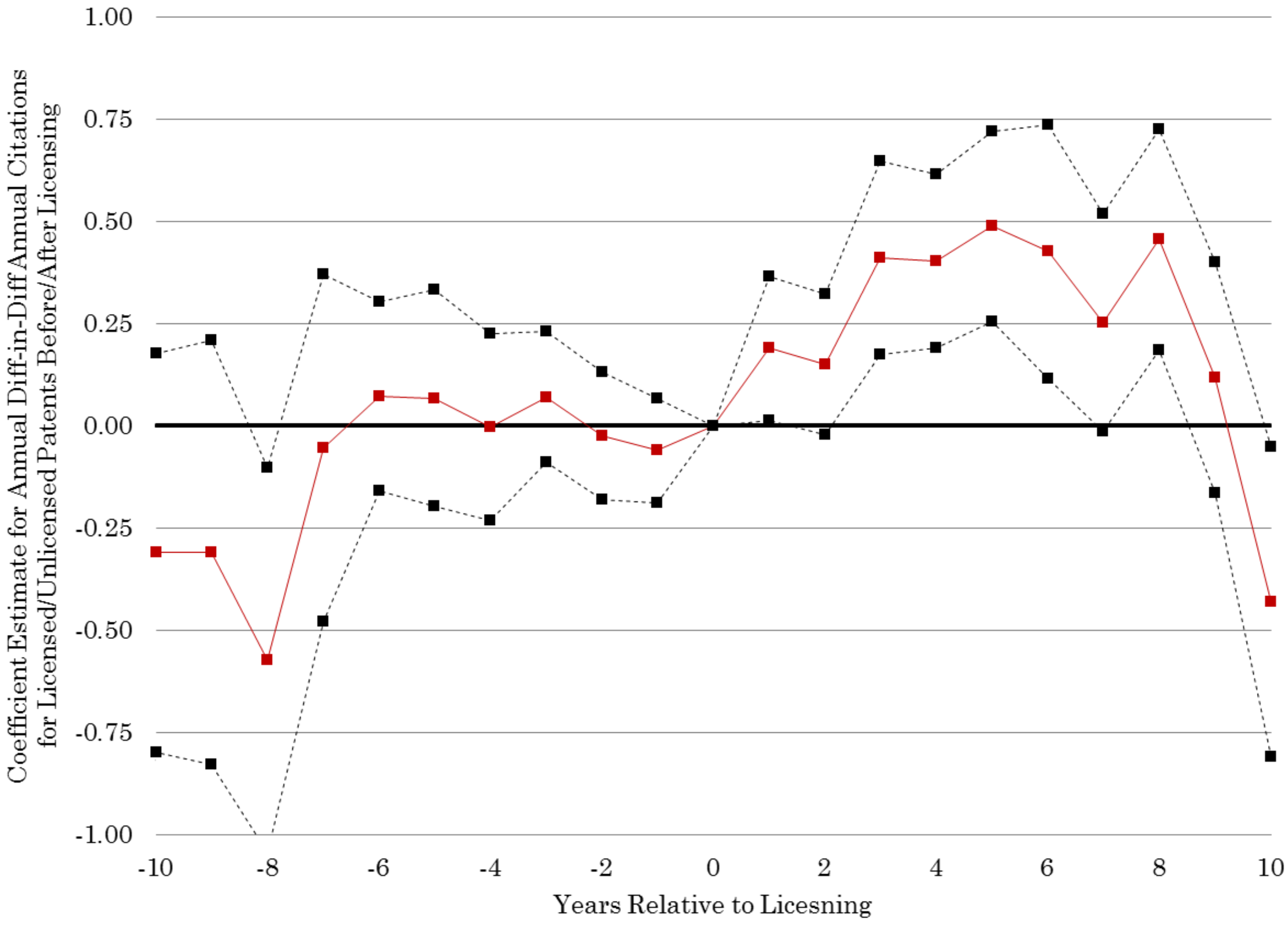


Figure 7. Annual Difference-in-Difference Estimates. Estimated diff-in-diff coefficients associated with individual years before and after licensing. Beginning 2 years after licensing, citations increase by about 0.3-0.4 cites/year through 7 years after licensing. Quantitative estimates shown in Model (6) in Table 2

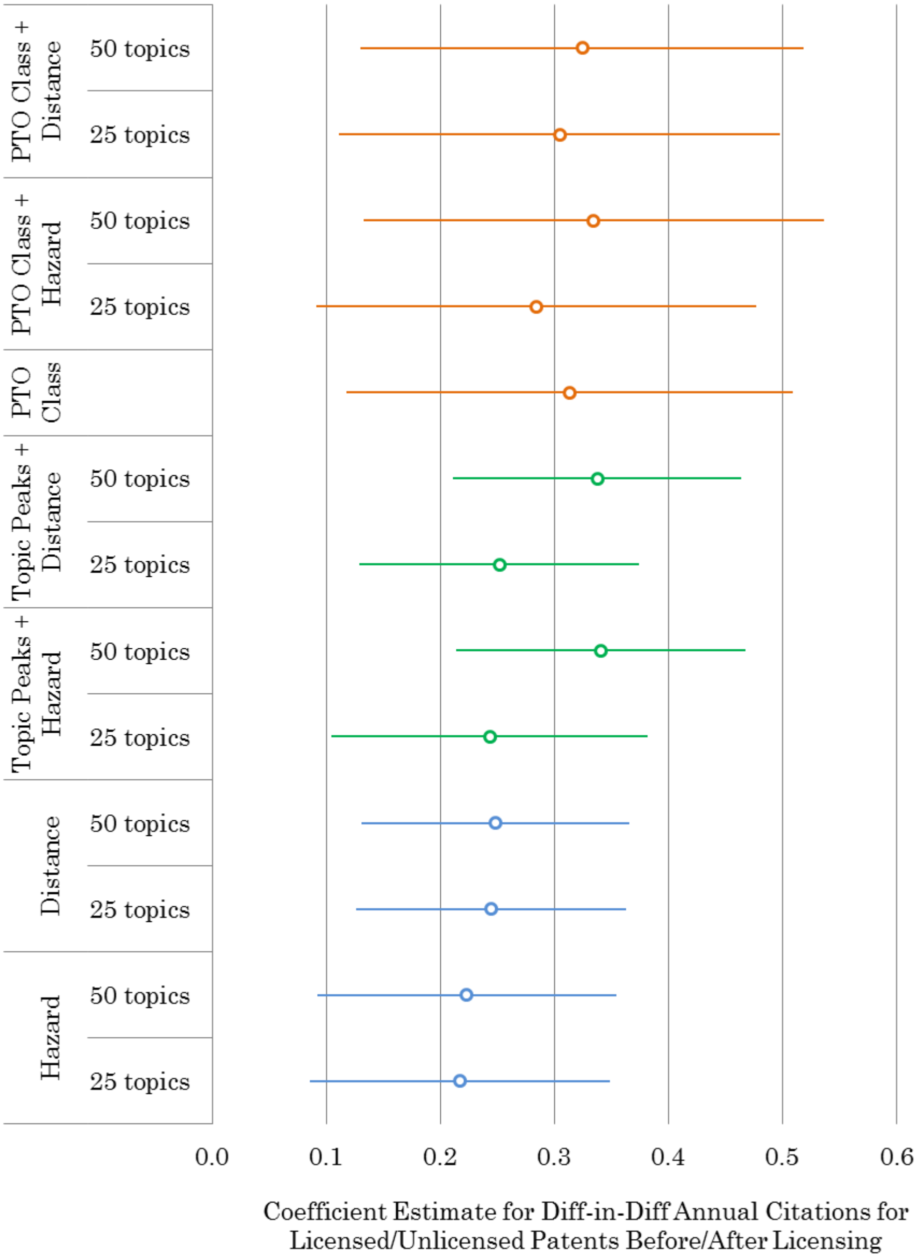


Figure 8. Sensitivity of coefficient estimates under different matching procedures. Estimates and 95% confidence intervals are shown for the matching procedures described in Section 4.1 for the regression model described by Equation 5.

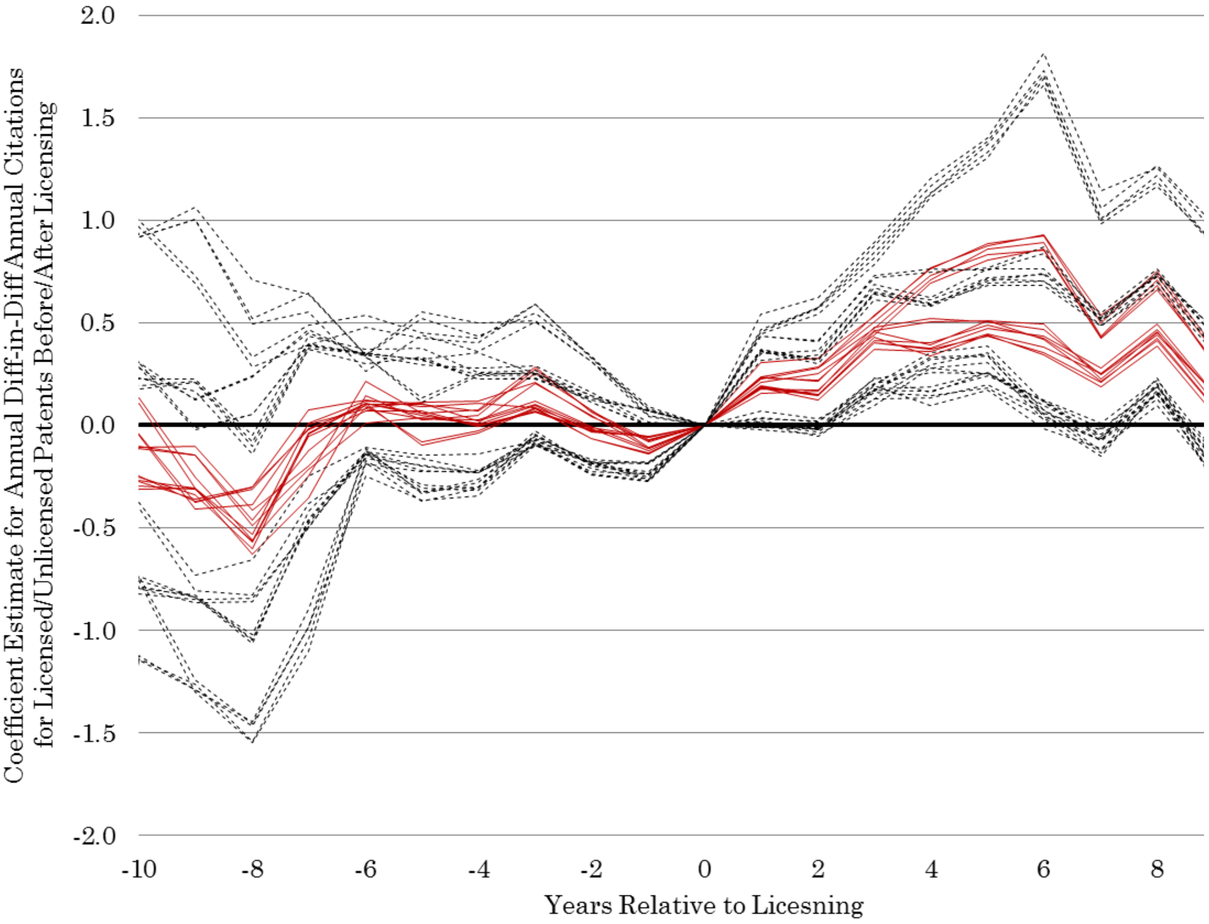


Figure 9. Sensitivity of coefficient estimates under different matching procedures. Estimates (red lines) and 95% confidence intervals (black dashed lines) are shown for the matching procedures described in Section 4.1 for the regression model described by Equation 6.

Table 1. Summary Statistics of R&D, Patenting, and Licensing by Lab Operator. Summary of R&D expenditure, patenting, and licensing for the National Labs disaggregated by Lab operator type and also showing the sample of five labs examined in this paper. Summaries of ratios of patenting pre R&D expenditure and licensing income per license also shown. All variables are three-year averages for 2009-2011 except for R&D expenditure, which is taken just at 2011 levels. For comparison, in 2011, all U.S. universities patented at a rate of 0.19 patent applications and 0.07 granted patents per million dollars of R&D (DOE Office of Energy Efficiency and Renewable Energy, 2012; National Science Board, 2014).

Operator Type	R&D Expenditure ($ mil, FY 2011)	Patents Filed per R&D Expenditure (patents / $ mil)	Patents Granted per R&D Expenditure (patents / $ mil)	Patents Licensed per Patent Filed	Patent Licensing Income ($ mil)	Patent Licensing Income per Licensed Patent ($ / patent)	Patent Licensing Income as Fraction of R&D Expenditure
University	2,453	0.07	0.03	11%	13.48	62,235	0.74%
Non-Profit	3,568	0.07	0.03	35%	12.80	18,525	0.28%
Government	753	0.01	0.01	13%	0.05	4,674	0.00%
Industry	6,569	0.07	0.03	13%	12.81	25,305	0.19%
Aggregate	13,343	0.06	0.03	18%	39.15	27,484	0.30%
Sample	5,441	0.06	0.03	23%	15.68	20,057	0.23%

Table 2. Descriptive Statistics at the Patent-Level and the Patent-Year-Level. Summaries of key parameters for all patents and the subset of patents that are ever licensed aggregated by observations at the patent-level and the patent-year level.

(a) Descriptive statistics at the patent-level

	All Patents					Ever Licensed Patents				
	Obs.	**Mean**	**St. Dev.**	**Min**	**Max**	**Obs.**	**Mean**	**St. Dev.**	**Min**	**Max**
Lab = LBNL	2,796	0.16	0.36	0	1	877	0.14	0.34	0	1
Lab = NETL	2,796	0.10	0.30	0	1	877	0.02	0.14	0	1
Lab = PNNL	2,796	0.23	0.42	0	1	877	0.54	0.50	0	1
Lab = SNL	2,796	0.44	0.50	0	1	877	0.27	0.44	0	1
Lab = BNL	2,796	0.07	0.25	0	1	877	0.04	0.19	0	1
Filing Date (Year)	2,796	2005.78	3.60	2000.09	2013.99	877	2004.34	3.06	2000.09	2012.80
Grant Date (Year)	2,796	2008.87	3.83	2001.32	2014.79	877	2007.61	3.46	2001.42	2014.79
License Date (Year)						877	2005.50	3.61	2000.12	2013.42
Grant Delay (Years)	2,796	3.09	1.49	0.25	11.17	877	3.26	1.61	0.46	11.17
Total Cites	2,796	7.92	18.51	0	252	877	12.14	22.75	0	236
Ever Licensed	2,796	0.31	0.46	0	1	877	1.00	0.00	0	1
Ever Licensed Exclusive	2,796	0.15	0.36	0	1	877	0.49	0.50	0	1

(b) Descriptive statistics at the patent-year-level

	All Patents					Ever Licensed Patents				
	Obs.	**Mean**	**St. Dev.**	**Min**	**Max**	**Obs.**	**Mean**	**St. Dev.**	**Min**	**Max**
Year	27,402	2008.94	3.69	2000	2014	9,852	2008.47	3.78	2000	2014
Age from Filing	27,402	5.06	3.69	0	14	9,852	5.53	3.78	0	14
Cites	27,402	0.80	2.32	0	51	9,852	1.08	2.58	0	42
Cites from Unique Assignees	27,402	0.38	1.06	0	33	9,852	0.63	1.50	0	33
Cites from Non-Zero Cited	27,402	0.43	1.62	0	35	9,852	0.59	1.85	0	34
Cites from Above Median Cited	27,402	0.33	1.37	0	30	9,852	0.45	1.50	0	24
Already Granted	27,402	0.69	0.46	0	1	9,852	0.71	0.45	0	1
Already Licensed	27,402	0.30	0.46	0	1	9,852	0.83	0.37	0	1
Already Licensed Exclusively	27,402	0.18	0.38	0	1	9,852	0.49	0.50	0	1

Table 3. Conditional means of annual citations for the full sample and matched sample approximating naïve single-difference and difference-in-difference regressions.

(a) Citations per year for patents that are never licensed and patents that are licensed during the time period of the panel

	Pre-Licensing	Post-Licensing	Post - Pre
Never Licensed	0.85	n/a	n/a
Ever Licensed	1.40	1.59	0.20
Licensed - Unlicensed	0.54	n/a	n/a

(b) Citations per year for patents in the preferred matched sample before and after licensing (or licensing of matched patent). The difference in difference estimate in the bottom right approximate the regression results shown below but without additional controls.

	Pre-Licensing	Post-Licensing	Post - Pre
Never Licensed	0.70	0.83	0.14
Ever Licensed	0.71	1.53	0.82
Licensed - Unlicensed	0.01	0.70	0.69

Table 4. Balance Check. Comparison of conditional means of key patent-level variables for licensed and unlicensed patents in the full sample and the preferred matched sample.

	Licensed Patents	All Unlicensed Patents	Difference (Full Sample)	p-value
Filing Year	2003.766	2005.855	2.088 ***	<0.0001
	(0.103)	(0.083)	(0.141)	
Grant Year	2007.022	2008.865	1.843 ***	<0.0001
	(0.117)	(0.088)	(0.152)	
Priority Year	2002.114	2004.684	2.570 ***	<0.0001
	(0.106)	(0.097)	(0.161)	
Grant Delay (Days)	1190.032	1099.118	-90.914 ***	<0.0001
	(19.864)	(11.843)	(22.073)	
Observations	877	1,919		

	Matched Licensed Patents	Matched Unlicensed Patents	Difference (Matched Sample)	p-value
Filing Year	2003.365	2003.474	0.109	0.5850
	(0.139)	(0.142)	(0.199)	
Grant Year	2006.756	2006.739	-0.016	0.9473
	(0.173)	(0.172)	(0.244)	
Priority Year	2002.422	2002.519	0.096	0.6607
	(0.156)	(0.155)	(0.220)	
Grant Delay (Days)	1231.788	1188.940	-42.847	0.2940
	(9.026)	(28.679)	(40.806)	
Observations	404	404		

* p < 0.5, ** p < 0.01, ***, p < 0.001

Table 5. Sensitivity Analysis with Different Matching Approaches. Average treatment effects, standard errors, t-statistics, and Rosenbaum Bounds sensitivity parameters for seven matching approaches. Estimates are for a subsample of the data of treated patents that are licensed in the same year that they are filed in (making all citations in the post-licensing period for ease of comparison).

Matching Model	ATT Difference in Citations (Licensed - Unlicensed)	Std. Error	t-Statistic	Γ
No Matching	0.71	0.11	6.66	n/a
Baseline	0.24	0.36	0.67	n/a
Baseline + PTO Class	0.62	0.16	3.97	1.34
Baseline + 25 topic model	0.58	0.18	3.21	1.64
Baseline + 50 topic model	0.63	0.17	3.82	1.45
Baseline + PTO Class + 25 topic model	0.55	0.18	3.11	1.20
Baseline + PTO Class + 50 topic model	0.69	0.20	3.43	1.59

*Note: Baseline propensity score model includes filing year and lab fixed effects. Standard errors and t-statistics are heteroskedasticity-consistent as proposed by Abadie and Imbens (2006) using 20 nearest neighbors.

Table 6. Baseline Regression Estimates.

	(1)	(2)	(3)	(4)	(5)	(6)
	Post-Licensing	Time-Series	Diff-in-Diff	Diff-in-Diff	Diff-in-Diff	Diff-in-Diff
	Cross-Section			matched pair FE	patent FE	patent FE
						yearly effects
Dependent Variable	All Cites	All Cites	All Cites	All Cites	All Cites	All Cites
Functional Form	OLS	OLS	OLS	OLS	OLS	OLS
Matched Sample	Preferred matching post-license period	Preferred matching licensed patents only	Preferred matching	Preferred matching	Preferred matching	Preferred matching
EVER_LICENSED	0.344 ***		0.015	-0.015		
	(0.100)		(0.021)	(0.015)		
LICENSED		0.520 ***	0.328 ***	0.335 ***	0.223 ***	
		(0.155)	(0.091)	(0.091)	(0.066)	
POST			0.145	-0.046		
			(0.110)	(0.067)		
PRE_LICENSE(10)						-0.294
						(0.249)
PRE_LICENSE(9)						-0.310
						(0.269)
PRE_LICENSE(8)						-0.566 *
						(0.246)
PRE_LICENSE(7)						-0.040
						(0.218)
PRE_LICENSE(6)						0.106
						(0.126)
PRE_LICENSE(5)						0.049
						(0.137)
PRE_LICENSE(4)						0.007
						(0.121)
PRE_LICENSE(3)						0.095
						(0.084)
PRE_LICENSE(2)						-0.031
						(0.080)
PRE_LICENSE(1)						-0.055
						(0.066)
POST_LICENSE(1)						0.184 *
						(0.091)
POST_LICENSE(2)						0.123
						(0.090)
POST_LICENSE(3)						0.372 **
						(0.123)
POST_LICENSE(4)						0.358 **
						(0.114)
POST_LICENSE(5)						0.445 ***
						(0.126)
POST_LICENSE(6)						0.381 *
						(0.164)
POST_LICENSE(7)						0.253
						(0.140)
POST_LICENSE(8)						0.465 ***
						(0.140)
POST_LICENSE(9)						0.155
						(0.144)
POST_LICENSE(10)						-0.379 *
						(0.191)
Age Fixed Effects?	Yes	Yes	Yes	Yes	Yes	Yes
Citing Year Fixed Effects?	Yes	Yes	Yes	Yes	Yes	Yes
Matched Pair Fixed Effects?	No	No	No	Yes	No	No
Patent Fixed Effects?	No	No	No	No	Yes	Yes
Observations	6,621	4,712	9,357	9,357	9,357	9,357
Number of Patents	808	404	808	808	808	808
Adj. R-Squared	0.094	0.127	0.096	0.086	0.093	0.102

* p < 0.5, ** p < 0.01, ***, p < 0.001

Table 7. Robustness to Functional Form.

	(1)	(2)	(3)	(4)	(5)	(6)
	Diff-in-Diff	Diff-in-Diff	Diff-in-Diff	Diff-in-Diff	Diff-in-Diff	Diff-in-Diff
	patent FE	patent FE	patent FE	patent FE	patent FE	patent FE
		yearly effects		yearly effects		yearly effects
Dependent Variable	All Cites	All Cites	All Cites	All Cites	ln(All Cites + 1)	ln(All Cites + 1)
Functional Form	OLS	OLS	Negative Binomial	Negative Binomial	OLS Log-Linear	OLS Log-Linear
Matched Sample	Preferred matching	Preferred matching	Preferred matching	Preferred matching	Preferred matching	Preferred matching
LICENSED	0.223 ***		0.265 ***		0.067 **	
	(0.066)		(0.068)		(0.022)	
PRE_LICENSE(10)		-0.294		-0.659		-0.180
		(0.249)		(0.455)		(0.093)
PRE_LICENSE(9)		-0.310		-0.966 *		-0.220 *
		(0.269)		(0.411)		(0.095)
PRE_LICENSE(8)		-0.566 *		-1.018 **		-0.248 **
		(0.246)		(0.340)		(0.096)
PRE_LICENSE(7)		-0.040		-0.422		-0.051
		(0.218)		(0.238)		(0.079)
PRE_LICENSE(6)		0.106		-0.170		0.012
		(0.126)		(0.191)		(0.049)
PRE_LICENSE(5)		0.049		-0.246		-0.026
		(0.137)		(0.174)		(0.050)
PRE_LICENSE(4)		0.007		-0.152		-0.034
		(0.121)		(0.155)		(0.042)
PRE_LICENSE(3)		0.095		0.036		0.030
		(0.084)		(0.130)		(0.035)
PRE_LICENSE(2)		-0.031		-0.076		-0.036
		(0.080)		(0.124)		(0.030)
PRE_LICENSE(1)		-0.055		0.001		-0.025
		(0.066)		(0.109)		(0.029)
POST_LICENSE(1)		0.184 *		0.231 *		0.040
		(0.091)		(0.098)		(0.029)
POST_LICENSE(2)		0.123		0.176		0.013
		(0.090)		(0.100)		(0.029)
POST_LICENSE(3)		0.372 **		0.375 ***		0.075 *
		(0.123)		(0.099)		(0.033)
POST_LICENSE(4)		0.358 **		0.396 ***		0.096 **
		(0.114)		(0.104)		(0.033)
POST_LICENSE(5)		0.445 ***		0.500 ***		0.119 ***
		(0.126)		(0.109)		(0.036)
POST_LICENSE(6)		0.381 *		0.486 ***		0.104 **
		(0.164)		(0.119)		(0.037)
POST_LICENSE(7)		0.253		0.464 ***		0.064
		(0.140)		(0.132)		(0.044)
POST_LICENSE(8)		0.465 ***		0.783 ***		0.182 ***
		(0.140)		(0.134)		(0.047)
POST_LICENSE(9)		0.155		0.605 ***		0.097
		(0.144)		(0.157)		(0.050)
POST_LICENSE(10)		-0.379 *		0.338		-0.043
		(0.191)		(0.195)		(0.062)
Age Fixed Effects?	Yes	Yes	Yes	Yes	Yes	Yes
Citing Year Fixed Effects?	Yes	Yes	Yes	Yes	Yes	Yes
Matched Pair Fixed Effects?	No	No	No	No	No	No
Patent Fixed Effects?	Yes	Yes	Yes	Yes	Yes	Yes
Observations	9,357	9,357	7,640	7,640	9,357	9,357
Number of Patents	808	808	615	615	808	808
Log Likelihood			-5,993.1	-5,967.6		
Wald Chi-Squared			839.6	880.4		
Adj. R-Squared	0.093	0.102			0.152	0.160

* p < 0.5, ** p < 0.01, ***, p < 0.001

Table 8. Robustness to Matching Design.

	(1)	(2)	(3)	(4)	(5)
	Diff-in-Diff	Diff-in-Diff	Diff-in-Diff	Diff-in-Diff	Diff-in-Diff
	patent FE	patent FE	patent FE	patent FE	patent FE
Dependent Variable	All Cites	All Cites	All Cites	All Cites	All Cites
Functional Form	OLS	OLS	OLS	OLS	OLS
Matched Sample	Hazard (25 topics)	Hazard (50 topics)	Distance (25 topics)	Distance (50 topics)	Topic Peaks + Hazard (25 topics)
LICENSED	0.217 ***	0.223 ***	0.244 ***	0.248 ***	0.243 ***
	(0.067)	(0.066)	(0.060)	(0.059)	(0.070)
Age Fixed Effects?	Yes	Yes	Yes	Yes	Yes
Citing Year Fixed Effects?	Yes	Yes	Yes	Yes	Yes
Matched Pair Fixed Effects?	No	No	No	No	No
Patent Fixed Effects?	Yes	Yes	Yes	Yes	Yes
Observations	9,377	9,357	9,366	9,441	8,134
Number of Patents	809	808	810	815	709
Adj. R-Squared	0.090	0.093	0.108	0.099	0.090

	(6)	(7)	(8)	(9)	(10)
	Diff-in-Diff	Diff-in-Diff	Diff-in-Diff	Diff-in-Diff	Diff-in-Diff
	patent FE	patent FE	patent FE	patent FE	patent FE
Dependent Variable	All Cites	All Cites	All Cites	All Cites	All Cites
Functional Form	OLS	OLS	OLS	OLS	OLS
Matched Sample	Topic Peaks + Hazard (50 topics)	Topic Peaks + Distance (25 topics)	Topic Peaks + Distance (50 topics)	PTO Primary Class	PTO Primary Class + Hazard (25 topics)
LICENSED	0.340 ***	0.251 ***	0.337 ***	0.313 **	0.284 **
	(0.064)	(0.062)	(0.064)	(0.099)	(0.097)
Age Fixed Effects?	Yes	Yes	Yes	Yes	Yes
Citing Year Fixed Effects?	Yes	Yes	Yes	Yes	Yes
Matched Pair Fixed Effects?	No	No	No	No	No
Patent Fixed Effects?	Yes	Yes	Yes	Yes	Yes
Observations	7,349	8,176	7,431	2,904	2,968
Number of Patents	648	711	655	273	277
Adj. R-Squared	0.099	0.103	0.099	0.097	0.088

	(11)	(12)	(13)
	Diff-in-Diff	Diff-in-Diff	Diff-in-Diff
	patent FE	patent FE	patent FE
Dependent Variable	All Cites	All Cites	All Cites
Functional Form	OLS	OLS	OLS
Matched Sample	PTO Primary Class + Hazard (50 topics)	PTO Primary Class + Distance (25 topics)	PTO Primary Class + Distance (50 topics)
LICENSED	0.334 ***	0.304 **	0.324 ***
	(0.102)	(0.098)	(0.098)
Age Fixed Effects?	Yes	Yes	Yes
Citing Year Fixed Effects?	Yes	Yes	Yes
Matched Pair Fixed Effects?	No	No	No
Patent Fixed Effects?	Yes	Yes	Yes
Observations	2,869	2,962	2,979
Number of Patents	268	278	280
Adj. R-Squared	0.087	0.093	0.092

* $p < 0.5$, ** $p < 0.01$, ***, $p < 0.001$

Table 9. Exclusive versus Non-Exclusive Licenses.

	(1)	(2)
	Diff-in-Diff	Diff-in-Diff
Dependent Variable	All Cites	All Cites
Functional Form	OLS	OLS
Matched Sample	Preferred matching	Preferred matching
LICENSED	0.223 ***	0.179 *
	(0.066)	(0.080)
EXCLUSIVE_LICENSED		0.091
		(0.097)
Age Fixed Effects?	Yes	Yes
Citing Year Fixed Effects?	Yes	Yes
Matched Pair Fixed Effects?	No	No
Patent Fixed Effects?	Yes	Yes
Observations	9,357	9,357
Number of Patents	808	808
Adj. R-Squared	0.093	0.093

* p < 0.5, ** p < 0.01, ***, p < 0.001

Table 10. Concentration of Diffusion.

	(1)	(2)
	Diff-in-Diff	Diff-in-Diff
Dependent Variable	All Cites	Cites from Uniquely Assigned Patents
Functional Form	OLS	OLS
Matched Sample	Preferred matching	Preferred matching
LICENSED	0.223 ***	0.169 ***
	(0.066)	(0.042)
Age Fixed Effects?	Yes	Yes
Citing Year Fixed Effects?	Yes	Yes
Matched Pair Fixed Effects?	No	No
Patent Fixed Effects?	Yes	Yes
Observations	9,357	9,357
Number of Patents	808	808
Adj. R-Squared	0.093	0.092

* p < 0.5, ** p < 0.01, ***, p < 0.001

Table 11. Accounting for Strategic Patenting.

	(1)	(2)	(3)
	Diff-in-Diff	Diff-in-Diff	Diff-in-Diff
Dependent Variable	All Cites	Cites from Non-Zero Cited Patents	Cites from Median Cited Patents
Functional Form	OLS	OLS	OLS
Matched Sample	Preferred matching	Preferred matching	Preferred matching
LICENSED	0.223 ***	0.150 ***	0.111 **
	(0.066)	(0.044)	(0.036)
Age Fixed Effects?	Yes	Yes	Yes
Citing Year Fixed Effects?	Yes	Yes	Yes
Matched Pair Fixed Effects?	No	No	No
Patent Fixed Effects?	Yes	Yes	Yes
Observations	9,357	9,357	9,357
Number of Patents	808	808	808
Adj. R-Squared	0.093	0.101	0.081

* p < 0.5, ** p < 0.01, ***, p < 0.001

Appendix I – The Latent Dirichlet Allocation Method

The general process to model the data begins by preprocessing the text, constructing a document-term matrix (DTM), and fitting an unsupervised model to the data. I use the R package “tm” (Feinerer et al., 2008) to preprocess the text and create the DTM and the R package “topicmodels” (Gruen and Hornik, 2011) to handle the topic modeling.

The preprocessing step involves six iterations over the corpus. First, I remove the metadata, which was stored in the first lines of the scraped files. Next, I strip excess white space, remove capitalization, and remove punctuation. Next I delete stopwords, commonly used words which carry little substantive meaning, such as “the” or “and,” using the tm package's list of English stopwords. Finally, I stem all words in the corpus using the “Snowball” package (Hornik, 2007). Stemming removes suffixes, such that the same word used in a different part of speech is recognized as the same word. For example, with stemming, the words “position,” “positioned,” and “positions” would be reduced to just the single word “position.” While stemming could potentially introduce additional bias into the analysis, the additional power gained by reducing the complexity of the underlying data is typically considered a worthwhile tradeoff.

In the next step, I construct the document term matrix (DTM). The DTM is a matrix with N rows and D columns, where N is the number of unique stemmed words in the corpus and D is the number of documents in the corpus. The DTM is a highly sparse matrix; in one example dataset I collected of 2,176 patents, there were 9,115 unique word stems and only 0.37% of the cells in the DTM were non-zero. The DTM is the basis for all subsequent text processing. By simplifying the data to only the DTM, two important and related features of the data are removed: 1) the ordering of words (including syntax and grammar), and 2) the proximity of words. These restrictions together are known as the “bag of words” assumption. Clearly, representing the corpus of documents by the DTM greatly simplifies analysis, but as with any simplification that removes information, important features of the raw data may be lost.

The approach I take to classifying the patent corpus is a probabilistic topic modeling approach using the latent Dirichlet allocation (LDA) model. This methodology is inspired by Blei and Lafferty (2007) and described in detail in Blei (2010). LDA requires that the number of topics be specified ex-ante (in a similar manner to other approaches I have experimented with, such as k-means, but unlike other possible methods, such as hierarchical clustering). A potential avenue for subsequent theoretical work could attempt to develop guidelines for thinking about the "optimal" (in some as-yet undefined way) number of topics to model.

The basic logic of LDA is to uncover the underlying structure of documents that likely generated the DTM extracted from the corpus. A "topic" is defined as a probability distribution over a finite vocabulary of words. Documents within the corpus have a distribution over the topics, while each word within the document is a draw from the distribution of words conditional on a topic. Beginning to move to a full parametric model, the distribution of unique words is Dirichlet[18] (i.e. the probability of a word occurring in a topic has a Dirichlet distribution) and the distribution of proportions of the topics within a document is a second Dirichlet distribution. For each of the N documents, the topics within the document have a Multinomial distribution with parameter drawn from the Dirichlet distribution of topics within the document. The individual words are drawn from a second Multinomial distribution conditional on the topic (drawn from the Multinomial distribution of topics) and the distribution of unique words in a topic (Gruen and Hornik, 2011).

[18] The Dirichlet distribution is a multivariate generalization of the Beta distribution. In each dimension, Dirichlet variates are bound on {0,1}. Dirichlet distributions are typically used to model probabilities of (>2) rivalrous events. The canonical examples of the use of Dirichlet distributions are modeling the lengths resulting from cutting a string of length 1 into several pieces, and the ratios of colored balls in an (Polya) urn.

More formally, each of the K topics, $\beta_1, \dots, \beta_K$ is defined by a distribution over the entire vocabulary of N words. Each of the D documents are composed of some relative frequency of the K topics, where $\theta_{d,k}$ is the proportion of the k^{th} topic in the d^{th} document. Similarly, words within a document are assigned to topics with $z_{d,n}$ the topic assignment ($z = 1, \dots, K$) for the n^{th} word in the d^{th} document. The actual data from the DTM is $w_{d,n}$, the (count or binary existence) of the n^{th} word in document d. This setup frames the natural language processing statistical model as a missing data problem, hence the name *latent* Dirichlet allocation. The probability model that describes the data-generation process is:

$$p(\beta, \theta, z, w) = \prod_{i=1 \dots K} p(\beta_i) \prod_{d=1 \dots D} p(\theta_d) \left(\prod_{n=1 \dots N} p(z_{d,n} | \theta_d) p(w_{d,n} | \beta, z_{d,n}) \right) \tag{7}$$

In words, this states that the (unconditional) distribution of topics is independent of the proportion of topics within a document which is independent of individual topic assignment of words conditional on the proportion of topics within the document, which is independent of the observed words conditional on the distribution of topics in the corpus and the assignment of topics to words in documents. The data-generation process can then be thought of as topics being generated for the entire corpus and topic frequencies generated for each document. Then, potential words within documents are assigned to topics (the same word can have positive probability in different topics), and finally observed words are generated given the overall corpus-level distribution of topics and the distribution of words within topics (Blei, 2010). Inference for the unobserved variables is made by conditioning on the observed words and using the Gibbs sampler to estimate the joint posterior distribution of the unobserved variables.

I base my implementation of the LDA algorithm on the R implementation provided in the “topicmodels” package (Gruen and Hornik, 2011). I fit the data with a full Bayesian model with diffuse weakly informative priors, using a Gibbs sampler. The LDA model estimates two distributions of interest: the distribution of topics within a document (for each document), and the distribution of topics within the corpus. Potential document-level covariates are the modal topic within a document (e.g. the focus of Kaplan and Vakili, 2011), or the (multi-dimensional) relative frequency of topics within a document. One potential application of the latter set of covariates could use matching algorithms on the vector of topics to identify similar pairs or groups of documents. Figure 5 gives a sense for the classification procedure described in this section.