Knowledge spillovers and corporate investment in scientific research*

Ashish Arora[†] Sharon Belenzon[‡] Lia Sheer [§]

March 9, 2020

Abstract

Using data on 800,000 publications by American public firms between 1980 and 2015, and patent citations to these publications, we study how corporate investment in research is linked to its use in the firm's inventions, and to spillovers to rivals. Firms produce more research when it is used internally, but less research when it is used by rivals. This tradeoff may be related to the decline in upstream research relative to downstream development in corporate R&D. Over our sample period, the propensity to cite corporate science has increased for internal and rival patents, but citations by rivals have increased faster.

*We thank Nick Bloom, Tim Bresnahan, Wesley Cohen, Dietmar Harhoff, David Hounshell, Richard Freeman, Alfonso Gambardella, Shane Greenstein, Brian Lucking, Matt Marx, Petra Moser, David Mowery, Andrea Patacconi, Mark Schankerman, Scott Stern, Manuel Trajtenberg, John Van Reenen, and seminar participants at NBER summer institute, Harvard, Stanford, Berkeley, Maryland, Munich, and Bocconi for helpful comments. We thank Bernardo Dionisi, Honggi Lee, Dror Shvadron, and JK Suh for excellent research assistance. All remaining errors are ours.

[†]Duke University, Fuqua School of Business and NBER

[‡]Duke University, Fuqua School of Business and NBER

[§]Duke University, Fuqua School of Business

1 Introduction

Although scientific research is typically thought of as being performed by universities and funded by the government, for-profit firms fund and perform a surprisingly large fraction of research in the United States. In 2017, the business sector funded about $85 billion in basic and applied research, accounting for about 22% of R&D funded by business, and 43% of all research in the United States.[1] Many significant scientific breakthroughs have come from scientists working not in universities but in corporate labs owned by companies such as Du Pont, ICI, Merck, Xerox, IBM, and, of course, AT&T. Companies invest in scientific research expecting that it will lead, directly or indirectly, to the creation of new products and processes. The upshot is that though economists often speak of R&D as a single construct, it is useful to distinguish between research ("R") and development ("D"). Upstream research is an input into downstream development, which more directly related to the innovations. In Vannevar Bush's words (Bush (1945): 241), "New products and new processes do not appear full-grown. They are founded on new principles and new conceptions, which in turn are painstakingly developed by research in the purest realms of science."

The two components of R&D differ in another respect as well. Whereas new products and processes can be protected from potential imitators by patents, copyrights, and trade secrecy, research, even by corporate scientists, is typically disclosed in the form of scientific publications. Research, therefore, is more likely to result in knowledge spillovers (e.g., Dasgupta and David (1994), Arrow (1962), Nelson (1959)) than downstream development. When these spillovers accrue to rivals, they are not just externalities; they may actually *reduce* private returns from research, so much so that Rosenberg (1990) posed the question of why firms invest in scientific research in the first instance. Rosenberg suggested that research often produced commercially valuable findings, but even absent that, research investments enabled firms to benefit from academic science. The subsequent literature has offered several possible mechanisms for private

[1]Henceforth, we shall use scientific research interchangeably with research. The business sector also performed about $86 billion in research in 2017. The data on business research come from the National Science Foundation, National Center for Science and Engineering Statistics 2019. National Patterns of R&D Resources: 2017–18 Data Update. NSF 20-307. Available at https://ncses.nsf.gov/pubs/nsf20307.

returns from research. The common feature of these explanations is that they do not require that the firm invests in research for use in its own inventive activity, and spillovers to other firms do not reduce returns from research.[2]

In this paper, we argue that private returns to corporate research depend on the balance between two opposing forces: the benefits from the use of science in own inventions, and costs of spillovers to rivals.[3] Changes in the balance between internal use and spillovers may be related to the changes in the declining share of research in corporate R&D. Specifically, as we discuss in Section 6 below, increases in spillovers relative to internal use may be a proximate cause of the decline in corporate research that the literature has documented (Mowery, 2009; Arora et al., 2019).

Our analysis includes all publicly traded American firms with at least one year of positive R&D expenditures and at least one patent over the period 1980-2015. The final sample consists of an unbalanced panel of 3,807 firms and 53,110 firm-year observations. We measure the use of internal research in invention by citations made by the firm's patents to its own scientific publications. Spillouts are measured by citations from the patents of rivals to the focal firm's publications.

With these newly constructed data, we present two main findings. First, we show that there is a positive relationship between the market value of a firm and its stock of scientific output. This relationship is stronger when the firm's patents use the science that the firm's scientists produce. Conversely, a firm's stock of scientific output is less valuable to the firm when its rivals use its science. Second, and consistent with this, we find that a firm produces more scientific publications if it is more likely to use the science in its patents, but produces

[2]These explanations include absorptive capacity (Cohen and Levinthal, 1989; Cockburn and Henderson, 1998; Griffith et al., 2004, 2006; Aghion and Jaravel, 2015), incentives for high-skilled scientist-inventors (Stern, 2004; Henderson and Cockburn, 1994; Audretsch and Stephan, 1996; Cockburn and Henderson, 1998), and enhancing reputation to attract investors, prospective customers or regulators (Azoulay, 2002; Hicks, 1995).

[3]Xerox's Palo Alto Research Center (PARC) is a case in point. PARC was one of the most innovative corporate research lab in the 1970s. Failures by Xerox to commercialize PARC discoveries, which frequently spilled over to companies such as Apple, Microsoft, and 3Com, are frequently cited as reasons for its ultimate demise. Yet, the benefits Xerox obtained from PARC's research in areas that were closer to Xerox's core business, such as the laser printer, were substantial. These inventions allowed the firm to recoup its investment in PARC despite the spillovers, at least for a time.

fewer publications if the science is more likely to be used by rivals' inventions.

Although we do not claim our estimates of these relationships as causal, the patterns of association are consistent with the notion that firms obtain value from scientific research if they have been able to use it for their inventions, but the value is reduced when knowledge spills out to rivals. The relationships endure even after controlling for firm fixed effects, as well as a variety of time-varying firm characteristics. We also present estimates where we instrument for citations by rivals using tax credits as instruments for patenting by rivals, following Bloom et al. (2013) and Lucking et al. (2018).

Our work connects to two streams of research in the economics of innovation literature. One stream of prior work has relied on confidential data to distinguish between research and development, such as Mansfield (1980), and Griliches (1986). Using a sample of approximately 1000 large manufacturing firms from 1957 through 1977, Zvi Griliches found that firms that spent a larger share of R&D on upstream research were substantially more productive. In a more recent paper, Akcigit et al. (2017) use confidential data on French firms to distinguish between basic and applied research. They argue that spillovers from basic research are broader than from applied research.

Instead of using confidential data on R&D inputs, we use publicly available data on the outputs, namely publications and patents. Our data enables us to trace knowledge flows from research to innovation, using patent citations to corporate publications. This enables us to measure the extent to which the firm's research is being used internally and the extent to which its rivals are using the research. We can, therefore, empirically explore the tradeoff between internal use and spillovers to rivals in a firm's decision to invest in research. Despite these differences, our empirical results are consistent with the findings in this literature. Consistent with Griliches (1986), we find that research is privately valuable, in part because research enhances the firm's inventive activities. Further, consistent with Akcigit et al. (2017), we find that knowledge spillovers are associated with publications, but not with patents.

Another literature has focused on knowledge spillovers. Building on Jaffe (1986), who measures spillovers using external R&D, Bloom et al. (2013) (hereafter, BSV) distinguish between the R&D expenditures of product-market rivals and technology rivals. While the latter captures a pure spillover, which should improve the focal firm's innovation outcomes, the former captures a rent-stealing effect: An increase in the knowledge base of close competitors would hurt the focal firm in the product market.

We build on BSV with three main differences. First, we consider upstream research as an input into downstream invention. Research is proxied by scientific papers, and is a more potent source of spillovers than inventions, which are protected by patents. Second, we introduce a direct measure of spillovers. While previous work typically measures potential spillovers by the (weighted) sum of R&D performed by other firms, we measure knowledge flows directly as patent citations to science produced by a focal firm.[4] Third, we focus on knowledge spill-outs, as opposed to knowledge spill-ins. That is, we examine how the use of own knowledge by outsiders affects the focal firm, rather than how a focal firm is affected by knowledge produced by other firms.[5] Unlike knowledge spillovers in general, which benefit other firms but do not directly affect the performing firm, spill-outs to rivals directly reduce the rents from innovation. Spillouts are, therefore, a direct cost of research, not simply a beneficial externality.

Our paper contributes to the ongoing policy discussions on the apparent decline in inventiveness (Bloom et al., 2017) and the associated slowdown in productivity growth. If inventions build on science, particularly corporate science, then a decline in corporate science may be implicated in the declining novelty of inventions. Although university research has increased considerably, it may not be a perfect substitute for corporate research as an input into invention (Arora et al., 2019). We suggest that corporate science may have declined partly because spillovers have increased faster than internal use.

Our paper also contributes by developing new data and measures. We develop and validate a new measure of use of science in invention. Patent citations to science have been used to

[4] Although patent citations are imperfect measures of knowledge flow, Roach and Cohen (2013) judge patent to publication citations to be better sources of tracking flow of scientific knowledge than patent to patent citations, which have been more extensively used in the literature e.g., Jaffe et al. (1993)

[5] Our paper is closer to Belenzon (2012), which examines the relationship between private returns to R&D and the use of own research by the focal firm and by its rivals. In related work, Ceccagnoli (2005) investigates a model where some firms invest in R&D that can spill out to rivals, who may not invest in R&D. However, the paper does not trace spillouts, nor does it distinguish between research and development.

track the flow of academic science to commercial invention.[6] Ours is the first large scale study measuring the flow of corporate science to corporate invention, within and across firm boundaries, for a period of over a third of a century. We match publication records from Web of Science to front-page non-patent literature (NPL) references and link both to Compustat firms. In so doing, we also improve and extend the NBER patent database, adjusting for changes in corporate names and ownership. The outcome is a more accurate and comprehensive match between firms and their stock of patents and publications, which accounts for changes in names, and for mergers, acquisitions, and divestitures. This is further described in section 3 below and in the Data Appendix.

The paper proceeds as follows. Section 2 presents the analytical framework that guides our empirical investigation. Section 3 discusses the data, Section 4 outlines the econometric specifications, and Section 5 summarizes the results. Section 6 concludes with a discussion of how trends in spillouts and internal use may account for some of the documented shift in corporate R&D, away from research and towards more downstream activities.

2 Analytical framework

We outline a framework that follows BSV, to motivate our empirical investigation, relegating details to the Appendix. Consider two firms, indexed by 0 and 1. Both compete in the product market and both invest in innovation, d_0 and d_1, respectively. Research by firm 0 reduces the cost of innovation of firm 0, but may also spill-out to the rival firm, reducing its cost of innovation. For simplicity, we assume that firm 1 does not invest in research, and thus also ignore how research spillovers lower the cost of research itself.

There are three stages. In stage 3, the firms compete in the product market. Their product market profits depend upon their own innovation output and that of the rival. The reduced form profit functions are $\Pi_0(d_0, d_1)$ and $\Pi_1(d_0, d_1)$. In stage 2, firms choose their innovation output. The cost of innovation for firm 0 is $\phi(r_0; \lambda)d_0$, where r_0 is the investment in research

[6] For instance Azoulay et al. (2019) use patent citations to track the flow of NIH funded research to biomedical inventions and Bryan et al. (2020) compare front-page citations to in-text citations to a fixed set of journals.

by firm 0, and λ represents internal absorptive capacity or the ability to learn from internal research. We assume that ϕ is decreasing in its arguments and that $\dfrac{\partial^2 \phi}{\partial \lambda \partial r_0} \leq 0$, so that the ability to learn from internal research is more effective when there is more to learn from. The innovation cost for firm 1 is $s(r_0;\theta)d_1$, where θ represents the ability of the firm to learn from firm 0's research. We assume that $s(r_0;\theta)$ is decreasing in its arguments, and that $\dfrac{\partial^2 s}{\partial \theta \partial r_0} \leq 0$. Finally, we assume that $\Pi_0(d_0, d_1)$ is concave in its arguments, as is $\Pi_1(d_0, d_1)$.[7]

In stage 1, firm 0 chooses its research investment, r_0, to maximize $v_0 = \Pi_0(d_0, d_1) - \phi(r_0;\lambda)d_0 - \gamma r_0$, where γ is the unit cost of research, and d_0 and d_1 are determined by the equilibrium in stage 2.[8] The first-order condition for an interior optimum is:

$$\frac{\partial v_0}{\partial r_0} = \frac{\partial \Pi_0}{\partial d_1}\frac{\partial d_1}{\partial r_0} - d_0 \frac{\partial \phi}{\partial r_0} - \gamma = 0 \tag{1}$$

Equation **1** shows the key tradeoff we focus on. The first term on the RHS of equation **1** represents the cost of spillout to its rival. It has two components: $\dfrac{\partial \Pi(d_0, d_1)}{\partial d_1}$ is the fall in the focal firm's profits from an increase in the rival's innovation; the second component, $\dfrac{\partial d_1}{\partial r_0}$ represents how the rival's innovation output responds to knowledge spillovers and is non-negative unless strategic substitution outweighs the direct cost reducing effect. Thus, we expect that the first term is negative – knowledge spillover to rivals is a cost of doing research. The middle term represents how research lowers the cost of innovation, and thus represents the benefit from research. The last term is the marginal cost of research, γ.

2.1 Value

The tradeoff presented in Equation **1** motivates our empirical analyses. In Table 4, we test whether research is more valuable to the extent that it is used by the firm, but less valuable if it is used by rivals. These considerations also apply to investment in research, but with one important difference: the production of research depends on how the marginal returns to

[7]For technical reasons, we also assume that $\frac{\partial^2 \Pi_0}{\partial {d_0}^2} = c_0 < 0$, $\frac{\partial^2 \Pi_1}{\partial {d_1}^2} = c_1 < 0$ and $\frac{\partial^2 \Pi_0}{\partial {d_1}^2} \leq 0$.

[8]Firm 1's payoff is $v_1 = \Pi_1(d_0, d_1) - s(r_0;\theta)d_1$.

research are affected by spillouts and internal use. The marginal returns also depend on the innovation response of rivals, and in particular, on whether innovation decisions are strategic substitutes or complements.[9] Under plausible assumptions, internal use in innovation would increase, and spillouts to rivals would decrease, investment in research. We explore these relationships empirically in Table 5. In Table 7, we explore empirically whether the use of scientific knowledge makes inventive activity more productive.

Result 1: Value increases with internal use $\frac{\partial v_0}{\partial \lambda} > 0$.[10]

Applying the envelope theorem, we get $\frac{\partial v_0}{\partial \lambda} = -d_0 \frac{\partial \phi}{\partial \lambda} + \frac{\partial \Pi_0}{\partial d_1}\frac{\partial d_1}{\partial \lambda}$. Strategic interactions affect the sign of $\frac{\partial d_1}{\partial \lambda}$, which is negative under strategic substitutability, zero if there are no strategic interactions in the product market, but positive if there are strategic complementarities (see Appendix equation 6).[11] Strategic substitutability is sufficient but not necessary to guarantee that internal use increases value.

Result 2: Value falls with rival spillouts $\frac{\partial v_0}{\partial \theta} < 0$.

Applying the envelope theorem, we get $\frac{\partial v_0}{\partial \theta} = \frac{\partial \Pi_0}{\partial d_1}\frac{\partial d_1}{\partial \theta} < 0$. The first term on the RHS is negative by assumption, and the second term is positive by Appendix equation 4. Thus, spillovers of its research to rivals decrease the value of the focal firm.

2.2 Research

The impact on research will depend upon how the marginal return to research $\frac{\partial v_0}{\partial r_0}$ is affected by spillouts and internal use, and in turn, on the nature and significance of strategic interactions. Formally, the sign of $\frac{\partial r_0}{\partial \lambda}$ is the same as the sign of $\frac{\partial^2 v_0}{\partial r_0 \partial \lambda}$, and similarly for spillouts, θ.

[9] If firm 1 also invests in research, then the results below should be interpreted as holding r_1 constant. Therefore, results 3 and 4, for instance, characterize firm $0's$ reaction function from stage 1.

[10] Holds if strategic substitutes or no strategic interactions

[11] Strategic interactions may be absent, if, for instance, the focal firm operates in two markets, old and new, but innovates only for the new market. It is plausible that knowledge spillouts to the rival enable the latter to innovate and enter the old market, thereby reducing the focal firm's profits, but without inducing any response in innovation from the focal firm.

Unlike the case for value, with strategic interactions, the impact on research requires additional assumptions. Without strategic interactions, we get the expected results, namely that the firm will invest less in research if spillouts increase and more in research if internal use increases, which we show more formally in Appendix equations 11 and 13, respectively.

Result 3: Research increases with internal use: $\frac{\partial r_0}{\partial \lambda} \geq 0$.

Result 4: Research falls with rival spillouts $\frac{\partial r_0}{\partial \theta} \leq 0$.

We summarize these results as follows:

VARIABLE	VALUE	RESEARCH
Spillout to rivals θ	Decrease	Decrease
Internal use λ	Increase	Increase

(a) no strategic interactions

VARIABLE	VALUE	RESEARCH
Spillout to rivals θ	Decrease	Ambiguous
Internal use λ	Increase*	Ambiguous

(b) Strategic interactions; *no strategic complementarity

3 Data

We combine data from six sources: (i) company and accounting information from S&P North American Compustat, (ii) scientific publications from Web of Science (WoS), (iii) patent and non-patent literature (NPL) citations from PatStat; (iv) subsidiary data from ORBIS, (v) acquisition data from SDC platinum, and (vi) company name changes from WRDS's "CRSP Monthly Stock". Appendix Table A1 summarizes the definition and data source for the main variables used in our empirical analysis. We revise and extend the NBER 2006 patent dataset (Hall et al. (2001) and Bessen (2009)). We re-construct the NBER data from 1980 and extend it to 2015 while introducing several improvements to accommodate changes in corporate names and ownership structures. We use scientific publications as our measure of the production of scientific knowledge and patents as our measure of inventive activity. We treat a citation by a patent to a corporate publication as an indicator that the patented invention used the knowledge in the publication. For this purpose, we also develop new data on corporate publications matched to NPL citations. The Data Appendix provides details on our data construction efforts. We discuss them briefly below.

Our data cover only publicly listed firms and their subsidiaries. Therefore, we examine how our coverage for patent assignees and publication authors compares to aggregate data from NSF.[12] For patents, the 2018 S&E Indicators' Appendix Table 8-1 states that private sector firms are granted 92,481 (83%) out of 110,759 patents assigned to U.S. owners in 2010. For the same year, Compustat firms receive 58,833 patents, or 64% of all private sector patents. Using data from 2002 to 2015 available from the NSF S&E indicators, our sample firms account for about 63% of granted private sector patents. For publications, the 2018 S&E Indicators' Appendix Table 5-41 states that the "industry sector" authored 32,074 (8%) out of 409,853 total U.S. S&E articles in the Elsevier Scopus database in 2010. For the same year, Compustat firms author approximately 24,000 papers, or about 80% of all industry papers identified in Scopus.[13]

Accounting panel data. We start with all North American Compustat records and select companies with positive R&D expenses for at least one year during our sample period of 1980-2015. We exclude firms that are not headquartered in the United States and firms without patents. As in Bloom et al. (2013), we further restrict the sample to manufacturing firms. Our final sample consists of an unbalanced panel of 3,807 firms and 53,110 firm-year observations.

Approximately 30% of the Compustat firms in our sample changed their name at least once, making it challenging to match publication and patent data to firms. Accounting for name changes is challenging because there is no single source that tracks different names of the same firm, and to the best of our knowledge, this has not been done previously on a large scale. We identify name changes in two ways: (i) we link Compustat records to WRDS's "CRSP Monthly Stock" file, which records historical names for each month a security is traded, and (ii) perform extensive manual checks using SEC filings to verify all related names for our sample

[12]Available at https://nsf.gov/statistics/2018/nsb20181/data/appendix

[13]The NSF S&E Indicators are based on Scopus whose coverage of scientific journals is slightly different from the Web of Science, which we use.

period.

The second major challenge comes from ownership changes. A parent company and a majority-owned subsidiary may have different identification numbers and records in Compustat. Moreover, a single company may correspond to multiple firm identifiers due to changes in ownership (such as mergers, acquisitions, and spinoffs). We identify ownership structures and ownership changes in three ways. First, we match our sample firms to ORBIS ownership files for the years 2002-2015 for annual subsidiary information (using each publication year as a separate "snapshot" of ownership structure).[14] Second, for firms that exit Compustat before 2002, we manually collect subsidiary names based on SEC filings and rely on the NBER patent database for pre-2002 ownership data. Third, we match our firms to M&A data from SDC Platinum to supplement information on ownership changes.

Corporate publications. We match our sample firms to the Web of Science database. We include articles from journals covered in the "Science Citation Index" and "Conference Proceedings Citation Index - Science", excluding social sciences, arts and humanities articles. Using the affiliation field and all historical company names, we identify approximately 800 thousand articles with at least one author employed by our sample of Compustat firms or their majority-owned subsidiaries at the time, published between 1980 and 2015.

Corporate Patents. We match patents to our sample of Compustat firms and their subsidiaries. We account for firm name changes as well as M&A reassignment of patents based on SDC and ORBIS data. As with publications, when ownership of the patenting entity changes, the stock of patents associated with the entity are reallocated to the new owner. We match approximately 1.3 million patents to our sample firms and their subsidiaries.

Patent citation to corporate publications. We match non-patent literature (NPL) citations to publications as our measure of the use of corporate science in invention. Using all patents granted in the period 1980-2015, we perform a many-to-many match between NPL citations and WoS publications (approximately 10 million citations matched to 800 thousand corporate publications), allowing for more than one publication to be matched to each citation.

[14] The year 2002 is the first year with reliable coverage of ownership information in ORBIS.

For each possible match, we construct a score that captures the degree of textual overlap between the free-text NPL format and the structured WoS record, which includes the following fields: article title, journal, and authors. To exclude mismatches, we use a more detailed secondary matching algorithm that is based on different WoS fields: standardized authors' names, number of authors, article title, journal name, and year of publication. The matching algorithm accounts for misspelling, unstructured text, incomplete references, and other issues that may cause mismatches. We manually verify the accuracy of the matches. We then focus on citations made by our sample of corporate patents. This process resulted in 70 thousand unique corporate cited publications, by 140 thousand unique corporate citing patents.[15]

3.1 Descriptive statistics

Our main sample and variables are at the firm-year level. Table 1 presents descriptive statistics for our main variables over the sample period. Our sample includes a wide distribution of firm sizes: market value ranging from 6 million dollars (10th percentile) to 4 billion dollars (90th percentile) and sales ranging from 3 million dollars (10th percentile) to 3.7 billion dollars (90th percentile). About 70% of firms have at least one publication during the sample period (by construction, all firms have at least one patent). These firms produce, on average, 19 publications per year. The distribution of publications is highly skewed, with the median firm producing one publication per year. We observe a similar pattern for patents, with an average of 24 patents per firm-year and a median of 2 patents.

[15]Papers and patents are matched "dynamically", such that, for instance, if a sample firm merges with another firm, the patents of the merged firm are included in the stock of patents linked to the Compustat record from that point onward, but not before. Most importantly, we can identify more accurately an internal or external citation based on the owner of the citing patent and that of the cited paper at the time the paper was published.

Table 1: Summary Statistics for Main Variables

					Distribution		
VARIABLE	# Obs.	# Firms	Mean	Std. Dev.	10th	50th	90th
Scientific publications count	41,664	2,781	19	101	0	1	22
Scientific publications stock	41,664	2,781	110	636	0	5	113
Patents count	53,110	3,807	24	138	0	2	36
Patents stock	53,110	3,807	128	677	1	8	192
R&D expenditures ($mm)	53,110	3,807	96	501	0.41	8	124
R&D stock ($mm)	53,110	3,807	432	2,410	1.3	32	524
Market value ($mm)	53,110	3,807	3,381	21,351	6	130	3,962
Tobin's Q	53,110	3,807	4	6	0	2	17
Sales ($mm)	53,110	3,807	2,253	11,470	3	119	3,640
Assets ($mm)	53,110	3,807	1,684	10,065	2	58	2,315

Notes: This table provides summary statistics for the main variables used in the econometric analysis for the sample period of 1980-2015.

Table 2 provides an additional descriptive analysis of citation patterns. Among the 2,781 publishing firms, 734 cite at least once their own publications in their patents, and 984 firms produce publications that are cited by other firms. Cited publications receive substantially more external citations than internal citations (8.6 vs. 3.9). Yet, the number of external citations drops sharply when accounting for the product market proximity between the citing and cited firms (from 8.6 to 3.1), indicating that a substantial portion of spillouts are unlikely to be harmful to the focal firm.

Further, in an additional analysis (not reported in the table) we find that publications that are cited internally are almost ten times more likely to receive an external citations. Publications that are cited by the firm's own patents receive 1.1 external citations, compared to only 0.1 external citations for publications that are not internally cited. Furthermore, we find that firms with above-mean share of internal citations of total citations received have more productive R&D programs (measured by number of publications and patents per dollar of R&D) and are more R&D intensive (measured by R&D expenditures over sales). These patterns are consistent with our main premise that although research may spill out to rivals, as long as the benefit of internal use offsets the private cost of spillouts, firms might have sufficient incentives to invest in scientific research.

Table 2: Summary Statistics for Patent Citations to Corporate Publications

	(1)	(2)	(3)	(4)
VARIABLE	No. of firms w. positive values	Citations per firm-year	No. of citing patents per firm-year	No. of cited publications per firm-year
Internal patent citations	734	3.88	1.95	2.33
External patent citations, corporate	984	8.63	6.25	4.40
External patent citations, rival	975	3.09	2.15	1.74

Notes: This table provides summary statistics for patent citations to corporate scientific publications by our sample firms. The sample is at the firm-year level and includes only firms with at least one cited publication.

3.2 Validating patent citations to scientific articles as a measure of use of science in invention

We use patent citations to scientific publications to measure the use of knowledge. Although patent citations are widely used, they are also widely criticized as imperfect measures of knowledge flows (Jaffe and Trajtenberg, 2002; Duguet and MacGarvie, 2005; Roach and Cohen, 2013). Roach and Cohen (2013) point out, however, that patent to publication citations, though imperfect, are much better than patent to patent citations at tracing knowledge flows, especially from public research to firms. Our interest is in knowledge flows from corporate research to other firms. To validate our measure of use of science, NPL citation to scientific articles, we use the Carnegie Mellon Survey (CMS) data on industrial R&D (Cohen et al., 2000). As part of the survey, lab directors in R&D performing firms were asked about the extent to which their R&D projects used scientific knowledge from various sources. Of the firms in our sample, 772 are also covered in the CMS, with patents granted between 1991 and 1999 (a total of 29,318 patents).

Table 3 confirms that firms whose patents cite scientific publications also reported that science contributed to their R&D projects, even after controlling for firm size, number of backward patent citations to other patents, and complete sets of four-digit industry SIC codes and year dummies. Furthermore, the fields of science that contributed the most to a firm are also those whose publications the firm's patents cite, and firms that draw on public science also tend to cite public science in their patents. Column 4 is especially important. It documents a strong relationship between our measure of patent citations to corporate science and the reported the value of other firms' research as a source into own invention.[16]

4 Econometric framework

The analytical framework in section 2 provides two sets of results. First, that research would increase the value of the firm to the extent it is used internally, but that it would be less valuable to the extent that it spills out to rivals. Second, and consistent with this, the firm would produce less research if it is more likely to spill out to rivals, and more research if the firm is more likely to use it internally.

4.1 Market value equation

We follow Bloom et al. (2013) and their predecessors (Griliches (1986) and Hall et al. (2005)) and estimate the following Tobin's Q specification (bold indicates vector representation):

$$\begin{aligned} \ln \frac{Value_{it}}{Assets_{it}} =& \alpha_0 \frac{G_{it-2}}{Assets_{it-2}} + \alpha_1 \ln\left(Internal\ citations_{it-2}\right) + \alpha_2 \ln\left(Rival\ citations_{it-2}\right) \\ & + \boldsymbol{Z}'_{it-2}\gamma + \boldsymbol{\eta_i} + \boldsymbol{\tau_t} + \epsilon_{it} \end{aligned} \tag{2}$$

Tobin's Q is market value over assets. G is knowledge assets, measured as the perpetual stocks of R&D, publications, and patents. The variable λ in section 2 reflects the extent to which internal science will reduce innovation costs. We measure it using *Internal cites*, the cumulative number of citations made by the focal firm's patents to its own publications. More precisely, $Internal\ citations_{it-2}$ counts all citations made by the focal firm's patents to its own research published up to and including year t-2.

Our measure of spillouts, *Rival citations*, is the cumulative number of citations made by

[16] We thank Michael Roach and Wesley Cohen for providing the Carnegie Mellon survey data to us.

Table 3: Supporting Evidence from Carnegie Mellon Survey

	Dependent variable: CMS questions				
	(1)	(2)	(3)	(4)	(5)
Response to CMS questions:	Public Research Findings (Q.18)		Main Research Field Findings (Q.22)	Other Firms' Research Findings (Q.16)	Share of Basic Research (Q.45)
Citations to top 200 universities articles	0.337				
	(0.146)				
Citations to public science articles		0.246			1.821
		(0.120)			(0.697)
Citations to articles in main research field			0.148		
			(0.065)		
Citations to corporate articles				0.453	
				(0.161)	
Citations to patents	0.001	0.001	-0.002	-0.003	-0.043
	(0.006)	(0.006)	(0.005)	(0.007)	(0.037)
ln(Sales)	0.078	0.074	0.040	-0.016	0.023
	(0.032)	(0.034)	(0.020)	(0.027)	(0.174)
Industry dummies	Yes	Yes	Yes	Yes	Yes
Observations	555	555	495	555	557
R-squared	0.39	0.39	0.46	0.41	0.39

Notes: This table presents OLS estimation results for the relationship between patent citation to publications per patent and the 1994 Carnegie Mellon survey (CMS) questions response (Cohen et al., 2000) related to the importance of research findings as an input to the firm's R&D projects. The analysis is at the patent level. Columns 1 & 2 are based on CMS Q18: "During the last three years, what percentage of your R&D unit's projects made use of the following research outputs produced by universities or government research institutes and labs?". Column 3 is based on CMS data Q.22: "Referring to the fields listed above, indicate the field whose research findings in general (not just university and government research) contributed the most to your R&D activities during the last three years". Column 4 is based on CMS data Q.16: "Below are some sources of activities or information on the R&D activities or innovations of other firms in your industry. Please score each of following in terms of the importance of that information's contribution to a recently completed major project". Here we use the average score to the source "Other firms' research findings." Column 5 is based on CMS data Q.45: "Approximately what percentage of your R&D effort is: a. Basic Research; b. Applied Research; c. Design and/or Development; d. Technical Service. The sample includes only patenting firms. In Column 3, the sample is restricted to firms that indicated their main research field in Q22 (excluding 'Others' category). For citations to articles in main research field, publications were classified into research fields based on Web of Science journal subject category. Citations to corporate articles include patent citations to publications by our sample of Compustat firms. Citations to patents include patent citations to patents. Standard errors in brackets are robust to arbitrary heteroscedasticity.

rivals to focal firm's publications.[17] We follow BSV and measure product-market proximity as the Mahalanobis similarity of vectors representing the shares of industry segment sales for each pair of firms (labeled as SEG). Industry segments are from Compustat's operating segments database. Citations received by firm i from firm j are weighted by SEG_{ij}, the "distance" of citing firm j from the cited focal firm i in the product market. For example, if Firm i and Firm j have similar sales shares across operating segments, the proximity score of the firms would be high. The Mahalonobis distance allows industry relatedness to be firm-specific by accounting for how dominant each firm is in an industry, with higher weights assigned for more dominant competitors. For instance, if Firm i and Firm j's sales both account for a large share of total industry sales in industry segment A, then segment A would be given a high weight in determining the proximity score between Firm i and Firm j. If internal use is valuable (see Result 1 in Section 2), we expect $\hat{\alpha_1} > 0$, and if spillouts reduce the value of research (Result 2), we expect $\hat{\alpha_2} < 0$.[18]

We focus on *spill-outs*, in contrast to the earlier literature (e.g., Jaffe (1986); Bloom et al. (2013)), which has stressed *spill-ins* or incoming knowledge flows. To facilitate comparison with that literature, we also present specifications that control for potential incoming knowledge flows. Accordingly, $\boldsymbol{Z}$ is a vector of controls, including sum of stocks of R&D, patents, and publications by other firms weighted according to the proximity of these firms to the focal firm in the product and technology spaces. As in Bloom et al. (2013), $SPILLSIC_{it}$ is the sum of weighted R&D by product market rivals and is computed as $\sum_j SEG_{ij} \times GRD_{jt}$. GRD_{jt} is

[17] $Rival\ citations_{it-2} = \sum_{j=1}^{N} n_{ij} \times SEG_{ij}$, where n_{ij} is the number of citations from patents of firm j to publications by firm i published up to year t-2 (inclusive), and SEG_{ij} is the product market proximity between the two firms. To get SEG_{ij} we follow BSV's procedure and weight the share of firm $i's$ sales in industry segment s (defined by 4-digit SIC codes) by the market share of firm i in industry segment s. Define W_i as the vector, whose individual component w_{is} is the share of segment s in firm $i's$ total sales, multiplied by the share of firm i in the total sales in segment s. The proximity between Firm i and Firm j is the cosine similarity of the vectors: $SEG_{ij} = \dfrac{W_i' \cdot W_j'}{|W_i'||W_j'|}$.

[18] If a firm invests in research to signal quality to regulators and customers or to attract talented researchers, citations of its publications by others would validate its claims to quality and reinforce the signal. That is, external citations, rather than representing profit-reducing spillovers, would increase profits. Similarly, higher-quality research, which is more likely to garner citations, would be positively related to profits. Thus, if citations by rivals is negatively related to value, this strongly suggests that spillouts of knowledge to rivals reduce profits from research.

the perpetual R&D stock of a potential rival firm j. Similarly, $SPILLTECH_{it}$ is the sum of outsiders R&D stock weighted by the technology distance as $\sum_j TEC_{ij} \times GRD_{jt}$. Technology proximity, TEC_{ij}, is measured analogously as the Mahalanobis similarity of vectors representing the shares of patents across 4-digit international patent classes (IPC) for each pair of firms, i, j. That is, we use the same formulation as for SEG_{ij}, but instead of share of sales by industry segments we use share of patents by IPC.

4.2 Publication equation

The relationship between scientific research with internal use and spillouts is specified as follows:

$$\begin{aligned} \ln(Publications_{it}) =& \beta_0 + \beta_1 \ln(Internal\ citations_{it-2}) + \beta_2 \ln(Rival\ citations_{it-2}) \\ &+ \boldsymbol{Z}'_{it-2}\boldsymbol{\gamma} + \boldsymbol{\eta_i} + \boldsymbol{\tau_t} + \epsilon_{it} \end{aligned} \tag{3}$$

Our coefficients of interest are β_1 and β_2. From Results 3 and 4 from Section 2 we expect $\hat{\beta}_1 > 0$ and $\hat{\beta}_2 < 0$, respectively. One possible concern is that firms with a higher number of publications or patents are more likely to show internal citations from patents to publications, which would bias $\hat{\beta}_1$ upward. To mitigate this concern, all specifications include firm fixed effects as well as firm controls for scale such as patent and R&D stocks. Furthermore, we only include citations received up to two years before the relevant year to mitigate concerns that the relationship between the number of publications and internal citation is merely due to common shocks (e.g., shocks to research opportunity that affects both the number of publications and the number of potentially citing patents by the focal firm).[19] We also present results when citations by rivals are instrumented using variation in state R&D taxes (see Section 4.4 below), and control for the sum of R&D by rivals and by firms overlapping in the technology space as in the Tobin's Q specification.

[19]The temporal structure of citations and publications is further clarified in Figure 8 in the Data Appendix.

4.3 Patent equation

We estimate the following patent production function to assess our premise that research and its use affect downstream invention and R&D productivity.

$$\begin{aligned} \ln(Patents)_{it} =&\omega_0 + \omega_1 Publications\ stock_{it-2} + \omega_2 Internal\ citations_{it-2} + \\ &\omega_3\ citation\ to\ rivals_{it-2} + \omega_4 \ln(R\&D\ stock)_{it-2} + \boldsymbol{\eta_i} + \boldsymbol{\tau_t} + \epsilon_{it} \end{aligned} \tag{4}$$

In Equation 4, $Patents_{it}$ is the flow of patents for firm i in year t. The framework presented in Section 2 is based on the assumption that the use of science, be it internal (λ) or rival (θ), would reduce the cost of innovation by reducing ϕ and s, respectively. We proxy internal use with $Internal\ citation$, measured as the cumulative citations to own publications, as before. To proxy for the the use of external science, we use $Citations\ to\ rivals$, the cumulative citations to rivals' publications.[20] If the use of science leads to more downstream invention, we expect $\hat{\omega}_2 > 0$ and $\hat{\omega}_3 > 0$. However, internal research can also enhance downstream invention indirectly by directing the firm's inventive activities to more promising avenues for invention (Nelson, 1982). The direct effect is captured by our measure of internal use, ω_2. The indirect effect should be captured by ω_1, the coefficient of the stock of publications. In this interpretation, including both publications and internal use offers a potentially informative "horse race" between the two mechanisms.

4.4 Instrumental variable strategy

Investment in research and citations made by external patents may be driven by common unobserved time-varying effects, leading to an upward bias in the OLS estimate of the spillout coefficients $\hat{\alpha_2}$ and $\hat{\beta}_2$ in the Tobin's Q and publication equations, respectively. For instance, if a particular line of scientific inquiry becomes economically promising, research in that field may receive more citations as other firms ramp up inventive activity in that field. Similarly,

[20]Section 2 had ignored research by rivals. Empirically, we explore how the use of rival's research conditions innovation by the focal firm by allowing its innovation to depend also on the use of external science.

an expansion in demand may increase research by the focal firm as well as its use by others. Formally, a firm and its rivals may have common shocks to the marginal benefits of R&D. These common shocks would result in a positive correlation between the research conducted by the focal firm and the patenting output of its rivals, and hence, between the research conducted by the focal firm and the citations the research receives from patents filed by rivals.

We follow Bloom et al. (2013) and Lucking et al. (2018) and exploit state variation in tax credits as an instrument for rivals' patenting.[21] In effect, our IV strategy is to use the variation in the cost of R&D as a source of exogenous variation in the extent of inventive activity (i.e., patents) to purge *Rival citations* of unobserved shocks to scientific opportunities or demand, which are common to the focal firm and its rivals (see Section 2 of the Appendix for more technical detail). R&D tax credits affect the marginal cost of R&D, but not the benefit. Therefore, they offer a source of variation in R&D that is independent of the confounding variation. For each sample firm, we calculate its cost of R&D and regress the number of patents against this cost. The predicted number of patents from this regression is used as our input into calculating a focal firm-specific aggregate number of predicted patents by its rivals, where the aggregation is based on the weighting procedures discussed in Section 4.1. The aggregate rival patents are used as our instrument for *Rival citations* in the publication and stock market value regressions.

We implement the IV approach by first projecting our logged patent count variable on both state and federal tax credit components of R&D user costs. Next, we calculate the predicted value of logged patent-count using the regression estimates, $\hat{\pi}_{it}$. For each firm i, we compute $Rival\hat{P}AT_{it} = \sum_i SEG_{ij}\hat{\pi}_{jt}$, where SEG_{ij} is the distance in product space between firm j and focal firm i, using Mahalnobis distance described earlier. Finally, we use $Rival\hat{P}AT_{it-2}$ as an instrument for *Rival citations*$_{it-2}$. This indirect procedure follows Bloom et al. (2013) and is used because *Rival citations*$_{it-2}$ is itself constructed by weighting citations from firms by the Mahalonobis distance. For all IV results, we use bootstrapped standard errors to correct for potential bias arising from using the predicted first-stage instrument.

[21] The coefficient estimate of internal use remains vulnerable to an upward bias from such common shocks.

5 Estimation results

5.1 Market value equation

Table 4 presents the estimation results for market value. Column 1 regresses Tobin's Q against R&D stock, with an R&D coefficient estimate of 0.11. Column 2 breaks up R&D stock into publication ("R") and patent ("D") stocks, indicating that two-thirds of the estimate of R&D is driven by development, where one third is attributed to research. Column 3 adds internal and rival citations. As expected, the coefficient estimate of *Internal citations* is positive ($\hat{\alpha_1} > 0$), and the estimate of *Rival citations* is negative $\hat{\alpha_2} < 0$. Both estimates are statistically different from zero.[22]

Column 4 presents the second stage IV estimates.[23] The estimates indicate that 0.63 internal citations offset the value destroyed by one rival citation.[24] The coefficient estimate of *Rival citations* is significantly larger in magnitude than the coefficient in the previous specifications. As discussed in section 4.4, a variety of factors may result in an upward bias in the OLS estimator. One possible explanation for the higher IV estimate is that rival citations are related to the unobserved quality of the firm's research. That is, rival citations may reflect the quality of the firm's science as well as spillouts. We expect that the former is positively associated with value, whereas the latter is negatively related. Another possible source of bias is common shocks. For example, shocks to demand or technical opportunities would result in a spurious positive correlation between patent citations received from rivals and the firm's

[22]In unreported analysis we follow BSV and estimate a specification that includes a 5-degrees polynomial expansion of R&D stock with internal and rival citations. The coefficient estimates of internal and rival citations are similar to those reported here (0.044 with a standard error of 0.011, and -0.042 with a standard error of 0.012, respectively).

[23]Column 1 in Table A2 presents the first stage results of regressing *Rival citations* against $Rival\hat{P}AT_{it-2}$. As expected, more patenting by rivals leads to more citations by these rivals to the focal firm's publications. Table A3 presents the estimation results of regressing rival patents, which generates our instrument $Rival\hat{P}AT_{it-2}$, on rival cost of R&D (Column 1). There is a strong negative relationship between rival R&D cost with rival patenting and rival R&D expenditures (Column 2).

[24]Average and median values for *Cumulative internal citations* are 58 and 422, and for *Cumulative rival citations* are 63 and 371, respectively. The marginal effect of an additional rival citation, evaluated at the sample mean, is $(-0.41 \times 4/64) \times 0.3 = -0.01$ (mean SEG value is 0.3, mean Tobin's Q is 4, and 64 is one plus average *Cumulative rival citations*). The same calculation for internal use is: $(0.23 \times 4/59) = 0.016$. Hence, to offset one rival citation, 0.010/0.016=0.63 internal citations are needed

own value. Measurement error is possible as well. We do not depreciate citations from rivals. However, citations received in the distant past may be a noisier proxy for the likelihood of future spillouts. The instrument variable estimation would purge some of this noise, reducing attenuation bias.

The literature on knowledge spillovers has used aggregate R&D by firms overlapping in technology (Jaffe (1986)), and by firms competing in the product market (BSV) to proxy for potential incoming knowledge flows or spill-ins. Accordingly, we include these measures in Columns 5-7. Following BSV, these are labeled $SPILLTECH$ and $SPILLSIC$, respectively. By so doing, we also adjust for the knowledge use patent citations to publications may miss: others may benefit from a firm's research without necessarily citing it in their patents. The results in Columns 5 of Table 4 show that the use of a firm's scientific knowledge by rivals as measured by citations continues to be negatively related to value, even after controlling for external R&D by rivals and other related firms. Although the coefficient of citations by rival patents drops in magnitude (citations by rival patents is positively related to R&D by rivals), it remains comparable to the coefficient on citations by internal patents, and statistically different from zero. Second, R&D by rivals and by firms operating in similar technology fields is negatively related to value, indicating that perhaps possible knowledge spill-ins are being offset by other rent reducing effects, such as more intense product market competition, or preemption in the technology space.

To study these opposing effects of external R&D, we replace aggregate external R&D with disaggregated measures of research and innovation by other firms, using the publications and patents by these firms in Columns 6 and 7, respectively. If publications are a source of external knowledge that firms can use freely in their own inventions, then we would expect external publications (be it of rivals or technically related firms) to be positively related to value. Patents also disclose knowledge, but this is knowledge over some or all of which the patentee has claimed property rights. External patents potentially also preempt the focal firm from inventing in the related technical space. Moreover, patents might presage forthcoming innovations that might reduce the focal firm's profits in the product market. Therefore, external patents may be negatively related to value.

Indeed, Columns 6 and 7 confirm the expected results. In Column 6, we see that the coefficient of publications by rivals, SPILLSIC-PUB, is positive, whereas the coefficient of rival patents, SPILLSIC-PAT is negative. Similarly, In column 7, we see that the coefficient of publications by firms operating in similar technology areas, SPILLTECH-PUB, is positive, whereas the coefficient of patents, SPILLTECH-PAT is negative. Overall, these results reinforce the basic premise of this paper that the different constituents of R&D, namely research and development, have very different economic properties. Upstream research is the more potent source of spillovers, whereas downstream development activities are more likely to be covered by means of appropriation, such as intellectual property rights or secrecy, and thus, more likely also to have an important market-stealing component. As a result, private returns to research are positively related to its internal use in invention, but negatively related to its use in invention by close product market rivals.

5.2 Publications equation

The tradeoff between internal use and spillouts will be reflected in the decisions to invest in research as well. However, there is an important nuance. The production of research depends not on the average return but rather the marginal returns to research. The latter, in turn, depends on the anticipated response by rivals. For instance, as shown in Section 2, an increase in internal use has a direct effect of increasing the marginal return, but an indirect effect that depends on how the rival responds. If there is strategic substitutability, the indirect effect is to reinforce the direct effect through an increase in the firm's investment in innovation. With strategic complementarity, the indirect effect is in the opposite direction. Similarly, an increase in spillouts has a direct effect on reducing the marginal return to research, but the indirect effect depends upon the innovation decisions of the focal firm and its rivals. Put differently, the effect of spillouts and internal use on the incentives for research, though governed by the same

Table 4: Stock Market Value and Patent Citations to Corporate Science

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Dependent variable:	ln(Tobin's Q)						
	R&D	"R" vs. "D"	Internal and rival citations	2nd Stage IV, Rival	Spill-ins, Jaffe and BSV	SPILLSIC, "R" vs. "D"	SPILLTECH, "R" vs. "D"
$Publication\ stock_{t-2}/Assets_{t-2}$		0.031	0.032	0.035	0.031	0.030	0.029
		(0.005)	(0.005)	(0.006)	(0.005)	(0.005)	(0.005)
$Patent\ stock_{t-2}/Assets_{t-2}$		0.074	0.074	0.074	0.075	0.076	0.076
		(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
$R\&D\ stock_{t-2}/Assets_{t-2}$	0.106						
	(0.004)						
$ln(Cumulative\ internal\ citations)_{t-2}$			0.025	0.228	0.028	0.027	0.025
			(0.011)	(0.112)	(0.011)	(0.011)	(0.011)
$ln(Cumulative\ rival\ citations)_{t-2}$			-0.036	-0.407	-0.026	-0.020	-0.024
			(0.012)	(0.201)	(0.012)	(0.012)	(0.012)
$ln(SPILLSIC,\ GRD)_{t-2}$					-0.127		-0.204
					(0.039)		(0.037)
$ln(SPILLSIC,\ PUB)_{t-2}$						0.179	
						(0.040)	
$ln(SPILLSIC,\ PAT)_{t-2}$						-0.291	
						(0.032)	
$ln(SPILLTECH,\ GRD)_{t-2}$					-0.244	-0.395	
					(0.062)	(0.055)	
$ln(SPILLTECH,\ PUB)_{t-2}$							0.085
							(0.056)
$ln(SPILLTECH,\ PAT)_{t-2}$							-0.277
							(0.043)
Firm fixed-effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Dependent variable sample average	4.127	4.127	4.127	4.127	4.127	4.127	4.127
Weak identification (Kleibergen-Paap)				F=39.462			
Number of firms	3,701	3,701	3,701		3,701	3,701	3,701
Observations	39,994	39,994	39,994	39,667	39,994	39,994	39,994
R-squared	0.69	0.68	0.68	0.06	0.68	0.68	0.68

Notes: This table presents OLS estimation results for the relationship between Tobin's Q with internal and rival citations. *Internal citations* counts citations made by the focal firm's patents to its own publications. *Rival citations* counts citations to the focal firm's publications from other firms, weighted by their product-market distance. SPILLSIC, SPILLSIC-PUB, and SPILLSIC-PAT are the product market distance weighted sum of all other firms' R&D, publication, and patent stocks, respectively. SPILLTECH, SPILLTECH-PUB, and SPILLTECH-PAT are the technology-distance weighted sum of all other firms' R&D, publication, and patent stocks, respectively. One is added to logged values. Standard errors in brackets are robust to arbitrary heteroscedasticity.

economic forces that affect value, can be empirically different.

Table 5 presents the estimation results for publications. Overall, our results are consistent with the view that a firm's investment in research depends, among other things, on how its research is used internally and externally.[25] Column 1 includes *Internal citations*. As expected, $\hat{\beta}_1$ is positive and statistically significant, indicating that firms produce more publications if their past publications were used internally.

Not all citations to science are equally relevant for investing in research. We expect internal citations to be more relevant to a firm's decision to invest in research when the cited publication (i) is more recent (and thus less likely to be merely a background reference), and (ii) is cited by the firm's valuable patents. These predictions are confirmed in Columns 2-3. Column 2 distinguishes between citations to old and new science. Internal citations to new science consist of citations to articles published no earlier than five years from the grant year of the citing patent, and citations to all earlier articles are treated as citations to old science. Only the coefficient estimate of citations to recent science is positive and statistically significant (the estimates on new and old science are statistically different from each other with a p-value < 0.01). Column 3 distinguishes between citations made by high and low quality patents.[26] The coefficient estimate for high-quality publications is positive and statistically significant, while the estimate for low-quality patents is statistically zero (the estimates are statistically different from each other with a p-value < 0.01).

In summary, Columns 1-3 are consistent with the view that internal patent citations to science that matter for the production of future science are citations that come from high-quality patents of the sponsoring firm to recent, high-quality, publications. These results further bolster the view that scientific output is an input into downstream inventive activity, and that to justify investment in research, managers need to demonstrate that the knowledge produced is useful for the downstream inventive activity of the sponsoring firm.

[25] Average and median values for *Internal citations* are 3 and 34, and for *Rival citations* are 3 and 23, respectively.

[26] Patent quality is based on the number of citations a patent receives divided by the average number of citations received by all patents granted in the same year as the focal patent. Patents are classified into high and low quality using median value from the corporate patents sample.

Column 4 adds *Rival citations*. The coefficient estimate $\hat{\beta}_2$ is negative and statistically significant.[27]. A firm whose research is used in its own inventive activity is likely to continue investing in research. However, a firm whose research spills over to rivals is likely to reduce its investment. Notice that this result nets out potentially offsetting effects. The direct effect of spillouts is to reduce the payoff from research by increasing innovation by rivals. However, if innovation strategies are strategic complements, an increase in innovation by rivals would induce the focal firm to increase innovation as well, which increases the marginal payoff to research, potentially offsetting some or all of the direct effect.

Column 5 presents the estimates from instrumenting *Rival citations* with $Rival\hat{P}AT_{it-2}$.[28] The estimate of *Rival citations* is larger in magnitude, indicating a larger negative effect of rival citations on focal publications. Based on the estimates from Column 5, the negative effect of an additional rival citation is offset by 0.64 internal citation (similar to the offsetting effect from Table 4).[29]

Columns 6-8 add controls for potential spill-ins. Similar to our findings from Table 4, the coefficient of publications by rivals, SPILLSIC-PUB, is positive, whereas the coefficient of rival patents, SPILLSIC-PAT is negative (but statistically not different from zero). In column 8, the coefficient of publications by firms operating in similar technology areas, SPILLTECH-PUB, and the coefficient of patents, SPILLTECH-PAT are both positive. This is a point of difference from the market value estimates, which show a negative estimate for the coefficient on SPILLTECH-PAT. A possible explanation is that while more patenting by firms operating in similar technology fields hurt the profits of the focal firm, strategic complementarities in technology markets lead to more patents and, consequently, more research. That is, the average effect of SPILLTECH-PAT is negative, while its effect on the marginal value of research is

[27] The coefficient estimates on internal and rival citations remain statistically significant also when we cluster standard errors by firms. The standard error on internal citations rises to 0.019 and on rival citations to 0.022. The coefficient estimates remain unchanged

[28] Column 2 of Table A2 presents the first stage results of regressing *Rival citations* against $Rival\hat{P}AT_{it-2}$, with the expected sign.

[29] An additional *Rival citation* lowers publications by $(-0.146 \times 20/4) \times 0.3 = 0.21$ (0.3 is average SEG value– the contribution of an additional citation by a rival, one plus average *Rival citations* is 4, and one plus average annual publications is 20). An additional *Internal citation* increases publications by $(0.083 \times 20/5) = 0.33$ (one plus average *Internal citation* is 5).

positive.

Table 6 presents a number of checks to confirm that the results are robust to a variety of changes in specifications and measurements. Column 1 controls for lagged dependent variable to mitigate a concern that citations capture a serial correlation in publications over time. Column 2 controls for internal patent citations to own patents ("self-citations") to mitigate a concern that $\hat{\beta}_1$ captures a patent "self-citation" effect (Hall et al., 2005; Belenzon, 2012). Column 3 presents an alternative measure of internal citations – the share of internal citations in all patent citations received by the focal firm's publications. Column 4 presents an alternative specification for rival citations – the share of rival citations in the total number of citations received. Column 5 presents estimates from a Negative Binomial publications count specification with pre-sample fixed effects (5-year pre-sample average number of publications) following Blundell et al. (1999). Column 6 presents results for Inverse hyperbolic sine transformation, and Column 7 presents the results of estimating a flexible lag structure model. Our key results remain robust.[30]

5.3 Patent equation

The framework developed in Section 2 assumes that scientific knowledge lowers the cost of innovation (or equivalently, increases the efficiency of investments in innovation). In Table 7, we directly explore whether the use of science enhances innovation. As is customary in the literature, we use patents to measure the flow of innovations produced by a firm, controlling for R&D investment. Our interest is at how the number of patents produced holding fixed R&D stock is related to the focal firm's publication stock and to the use by the focal firm of outside publications.

[30] A potential concern is related to who adds citations to science and whether this reflects the use of the scientific discovery in the patented invention. For instance, firms that invent on a large scale are more likely to use in-house patent attorneys. Such attorneys are likely to know the focal firm's research better than outside research, possibly leading to more internal citations. In unreported specifications we examine this issue. We construct data on patent attorneys for the corporate patents in our sample. We harmonize their names and assign them unique ids. For each patent attorney, we calculate the share of her patents filed for a focal firm of the total patents she files. If this share exceeds 80%, we classify the attorney as "in-house" (results are not sensitive to that threshold). There are 63,010 unique patent attorneys for our sample patents. Of them, 3,525 are classified as in-house attorneys. Controlling for the share of patents filed by internal attorneys does not affect coefficient estimates of the variables of interest. The coefficient estimate of in-house attorneys is positive, as expected, but it is not statistically different from zero. The results are available on request. We thank an anonymous reviewer for alerting us to this issue.

Table 5: Use and Publication Output

Dependent variable:	ln(1+number of publications)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Internal citations	New vs. Old science	High vs. Low Quality Patents	Rival citations	2nd Stage IV, Rival	Spill-ins, Jaffe and BSV	SPILLSIC, "R" vs. "D"	SPILLTECH, "R" vs. "D"
$ln(Internal\ citation\ to\ publications)_{t-2}$	0.037			0.069	0.093	0.067	0.066	0.066
	(0.009)			(0.010)	(0.014)	(0.010)	(0.010)	(0.010)
$NEW\ publications$		0.127						
		(0.012)						
$OLD\ publications$		-0.048						
		(0.010)						
$High\ quality\ citing\ patents$			0.051					
			(0.011)					
$Low\ quality\ citing\ patents$			-0.020					
			(0.012)					
$ln(Rival\ citation\ to\ publications)_{t-2}$				-0.076	-0.146	-0.085	-0.088	-0.086
				(0.013)	(0.026)	(0.013)	(0.013)	(0.013)
$ln(SPILLSIC,\ GRD)_{t-2}$						0.015		-0.035
						(0.022)		(0.022)
$ln(SPILLSIC,\ PUB)_{t-2}$							0.132	
							(0.022)	
$ln(SPILLSIC,\ PAT)_{t-2}$							-0.051	
							(0.019)	
$ln(SPILLTECH,\ GRD)_{t-2}$						0.152	0.089	
						(0.035)	(0.031)	
$ln(SPILLTECH,\ PUB)_{t-2}$								0.281
								(0.032)
$ln(SPILLTECH,\ PAT)_{t-2}$								0.136
								(0.026)
$ln(R\&D\ stock)_{t-2}$	0.157	0.156	0.157	0.158	0.159	0.155	0.154	0.154
	(0.006)	(0.006)	(0.006)	(0.006)	(0.005)	(0.006)	(0.006)	(0.006)
$ln(Patent\ stock)_{t-2}$	0.058	0.058	0.058	0.060	0.062	0.058	0.060	0.056
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Firm fixed-effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Dependent variable average:	16	16	16	16	16	16	16	16
Weak identification(Kleibergen-Paap)					F=116.358			
Number of firms	3,807	3,807	3,807	3,807	3,807	3,807	3,807	3,807
Observations	45,474	45,474	45,474	45,474	45,474	45,474	45,474	45,474
R-squared	0.89	0.89	0.89	0.89	0.08	0.89	0.89	0.89

Notes: This table presents OLS estimation results for the relationship of publications to internal and rival citations. All the citation variables include patent citations at year t-2 to publications published up to year t-2. SPILLSIC is the product market distance weighted sum of all other firms' R&D stocks, and SPILLTECH is the technology-distance weighted sum of all other firms' R&D stocks. All specifications include a dummy variable that receives the value of one for firms that never published up to the focal year; a dummy variable that receives the value of one for firms without granted patents in that year; and a dummy variable that receives the value of one for firms without annual patent citations to own publications. Standard errors (in brackets) are robust to arbitrary heteroscedasticity.

Table 6: Robustness checks

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Dependent variable:		ln(1+Number of publications)			Number of publications	ASIN(Number of publications)	ln(1+Number of publications)
	Lagged publications	Patent self-citations	Share of internal citations	Share of rival citations	Neg. Binomial	Inverse hyperbolic sine	Flexible lag structure
$ln(Internal\ citation\ stock)_{t-2}$	0.028	0.051			0.147	0.071	
	(0.008)	(0.010)			(0.028)	(0.020)	
$ln(External\ citation\ stock,\ RIVAL)_{t-2}$	-0.044	-0.079			-0.092	-0.095	
	(0.010)	(0.013)			(0.030)	(0.030)	
$ln(R\&Dstock)_{t-2}$	0.081	0.155	0.160	0.159	0.335	0.186	0.146
	(0.004)	(0.006)	(0.006)	(0.006)	(0.028)	(0.020)	(0.019)
$ln(Patent\ stock)_{t-2}$	0.027	0.045	0.063	0.061	0.147	0.070	0.070
	(0.003)	(0.004)	(0.004)	(0.004)	(0.021)	(0.011)	(0.011)
$ln(Patent\ self-citations)_{t-2}$		0.034					
		(0.004)					
$Share\ of\ internal\ citations\ of\ citations\ received_{t-2}$			0.160				
			(0.029)				
$ln(Total\ citations\ received)_{t-2}$			-0.030				
			(0.007)				
$Share\ of\ rival\ citations\ of\ external\ citations\ received_{t-2}$				-0.135			
				(0.028)			
$ln(External\ citations\ received)_{t-2}$				-0.010			
				(0.008)			
$ln(Internal\ citation\ to\ publications)_{t-1}$							0.071
							(0.014)
$ln(Internal\ citation\ to\ publications)_{t-2}$							0.029
							(0.012)
$ln(Internal\ citation\ to\ publications)_{t-3}$							0.021
							(0.011)
$ln(Internal\ citation\ to\ publications)_{t-4}$							-0.005
							(0.012)
$ln(Internal\ citation\ to\ publications)_{t-5}$							-0.005
							(0.016)
$ln(External\ citation\ to\ publications,\ RIVAL)_{t-1}$							-0.007
							(0.019)
$ln(External\ citation\ to\ publications,\ RIVAL)_{t-2}$							-0.021
							(0.016)
$ln(External\ citation\ to\ publications,\ RIVAL)_{t-3}$							-0.030
							(0.016)
$ln(External\ citation\ to\ publications,\ RIVAL)_{t-4}$							-0.060
							(0.016)
$ln(External\ citation\ to\ publications,\ RIVAL)_{t-5}$							-0.055
							(0.019)
Pre-sample FE					0.463		
					(0.025)		
$ln(1 + Number\ of\ publications)_{t-1}$	0.468						
	(0.007)						
Firm fixed-effects	Yes	Yes	Yes	Yes	No	Yes	Yes
Industry dummies	No	No	No	No	Yes	No	No
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Dependent variable sample average	15.9	15.9	15.9	15.9	13.821	11.3	18.227
Number of firms	3,807	3,807	3,807	3,807	3029	3,807	3029
Observations	45,474	45,474	45,474	45,474	34,869	45,474	34,869
R-squared	0.91	0.89	0.89	0.89		0.88	0.90

Notes: This table presents robustness checks for the relationship of publications to internal and rival citations. All specifications include a dummy variable that receives the value of one for firms that never published up to the focal year; a dummy variable that receives the value of one for firms without yearly granted patents; and a dummy variable that receives the value of one for firms without annual patent citations to own publications. Standard errors in brackets are robust to arbitrary heteroscedasticity.

Column 1 includes publications stock without controlling for citations. As expected, publications are positively related to patenting. Column 2 adds internal use. Interestingly, the coefficient estimate of publications stock barely moves, indicating that scientific research may also enhance the productivity of a firm's innovation efforts by guiding it towards more fruitful avenues of research, in addition to directly producing commercially valuable discoveries. Put differently, scientific discoveries may directly lead to new products and processes, which would be captured by patent citations to internal publications. However, scientific knowledge may also guide innovation activities into more productive avenues and away from less productive ones. This feature of "science as a map" is reflected by the association between publication stock and patenting flow (Fleming and Sorenson, 2004; Nagaraj and Stern, 2020). Columns 3 and 4 add citations by the focal firm to publications by rivals, and to all external publications. As expected, these variables are positively related to patenting output of the focal firm. Knowledge produced by other firms can also enhance innovation by the focal firm. These results are also robust to alternative specifications, such as normalizing number of patents by R&D expenditures and weighting patent by the number of citations they receive from other patents.

Column 5 controls for SPILLSIC-PAT, the stock of patents by rivals. Whereas publications represent knowledge that can be freely used in downstream inventions, a firm may respond to patents by rivals by increasing its own innovation (strategic complementarity) or reducing it (strategic substitutes). The estimate of SPILLSIC-PAT is positive and statistically significant, pointing to a possible strategic complementarity (see also Appendix equation 4). Interestingly, controlling for SPILLSIC-PAT does not materially change any of the other coefficient estimates. Column 6 replaces SPILLSIC-PAT with SPILLTECH-PAT, the stock of patents by the broader set of firms that overlap in technology space with the focal firm, without any material change in the coefficients of use of internal and external science.[31]

[31] The coefficient of SPILLTECH-PAT is positive and also larger in magnitude than the coefficient of SPILLSIC-PAT. This may be due to shocks to technical opportunities that may be common to firms overlapping in technology space, or strategic complementarities, consistent with the positive coefficient of SPILLTECH-PAT in the Tobin's Q equation.

Table 7: Use of science and R&D Productivity

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable:	ln(1+Number of patents)					
	Publications stock	Internal use	Spill-ins from rivals	All spill-ins	Rival patents, BSV	Rival patents, Jaffe
$ln(Publications\ stock)_{t-2}$	0.282	0.222	0.217	0.220	0.203	0.199
	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
$ln(citation\ to\ internal\ publications\ per\ patent)_{t-2}$		0.128	0.085	0.092	0.084	0.106
		(0.026)	(0.027)	(0.027)	(0.027)	(0.026)
$ln(citation\ to\ rival\ publications\ per\ patent)_{t-2}$			0.189		0.172	0.186
			(0.023)		(0.022)	(0.022)
$ln(citation\ to\ all\ external\ publications\ per\ patent)_{t-2}$				0.057		
				(0.008)		
$ln(SPILLSIC,\ PAT)_{t-2}$					0.245	
					(0.007)	
$ln(SPILLTECH,\ PAT)_{t-2}$						0.924
						(0.032)
$ln(R\&D\ stock)_{t-2}$	0.240	0.258	0.255	0.257	0.245	0.238
	(0.008)	(0.008)	(0.008)	(0.008)	(0.007)	(0.007)
Firm fixed-effects	Yes	Yes	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes
Dependent variable sample average	26.79	26.79	26.79	26.79	26.79	26.79
Number of firms	3,807	3,807	3,807	3,807	3,807	3,807
Observations	45,496	45,496	45,496	45,496	45,496	45,496
R-squared	0.86	0.87	0.87	0.87	0.87	0.87

Notes: This table presents OLS estimation results of a patent equation, which examines the relationship between R&D productivity and citations to science. One is added to all citation variables. All specifications include a dummy variable that receives the value of one for firms without citations to publications at the focal year. Standard errors (in brackets) are robust to arbitrary heteroscedasticity.

6 Conclusion and discussion

Our findings support the view that firms invest in research to feed downstream technology development. On the other hand, when the knowledge spills over to rivals, profits fall. Over time, firms will invest less in research if the output of their research becomes relatively less important for the technology they develop or if it is more likely to spillout to rivals. Figure 1, which is based on National Science Foundation data, shows that the share of basic and applied research in the total domestic R&D funded and performed by corporations, the "R" of R&D, in the United States has declined from over 31 percent in 1985 to about 20 percent in 2008 and has remained at that level thereafter.[32] The share of research in total R&D performed by business shows a similar, albeit less dramatic, decline, from a peak of around 30 percent in 1991 to about 20 percent in 2008, and rising modestly thereafter. These trends suggest that the composition of corporate R&D is shifting away from "R" and towards "D".

Impressionistic accounts also indicate that many leading American corporations began to withdraw from upstream scientific research in the 1980s (Mowery, 2009). Many corporate labs were shut down or were oriented towards more applied activities. Bell Labs was separated from its parent company AT&T and placed under Lucent in 1996; Xerox PARC was spun off into a separate company in 2002, and Du Pont closed its Central Research and Development Laboratory in 2016.[33] These trends are also reflected in Figure 2, which presents trends in the annual number of publications ("R") and patent ("D") divided by sales, for our sample firms. The corporate publication rate fell by about 60% over the sample period, whereas patenting rates do not show any clear trend. The pattern suggests that the composition of corporate R&D is changing over time, with less "R" and more "D" (Arora et al., 2018, 2019).

Our results suggest that changes over time in the importance and magnitude of spillovers and internal use may be an important proximate cause of the decline in corporate production

[32]Based on Tables 2-4, National Science Foundation, National Center for Science and Engineering Statistics 2019. National Patterns of R&D Resources: 2017–18 Data Update. NSF 20-307. Alexandria, VA. Available at https://ncses.nsf.gov/pubs/nsf20307.

[33]DuPont was a major producer of chemistry research. In the 1960s, DuPont researchers published more articles in the *Journal of the American Chemical Society* than MIT and Caltech combined.

of scientific research. Figure 3 shows that the propensity of patents to cite corporate science (measured as the ratio between citations to corporate science per patent and total number of available corporate publications in a given year) has been rising over time for both internal and rival citations, but whereas citations by rivals have increased by about 350%, internal citations have increased by only about 150%. Simply put, Figure 3 shows that more knowledge is spilling over to close competitors.[34] The estimated elasticity of publications with respect to rival citations from Column 5 in Table 5 is -0.146, implying that the increase in rival citations can account for about 50% of the decline in corporate publications over the sample period. A similar "back of the envelope" calculation, with 0.083 as the elasticity of publications with respect to internal citations, implies that the rise in internal citations would account for about 12% additional publications produced over the same period. Hence, the net change in the use of corporate science can potentially account for up to 40% of the fall in corporate publications between 1980 and 2015.

The decline in corporate participation in science, even as inventions themselves become more dependent on science, is piquant. Even as firms make greater use of the scientific knowledge produced by others, they themselves are less willing to produce such knowledge, preferring to focus attention and resources from upstream research to downstream development: from "R" to "D". This shift, though likely privately profitable, is not without social costs. The results in Table 7 point to one such possible cost. The declining corporate engagement in research may be contributing to the reported decline in R&D productivity and the associated decline in productivity growth (e.g., Bloom et al. (2017)). Our findings suggest that a more careful

[34] These trends are pithily illustrated in the following quote from a former Bell Labs researcher (Odlyzko (1995), p.4):

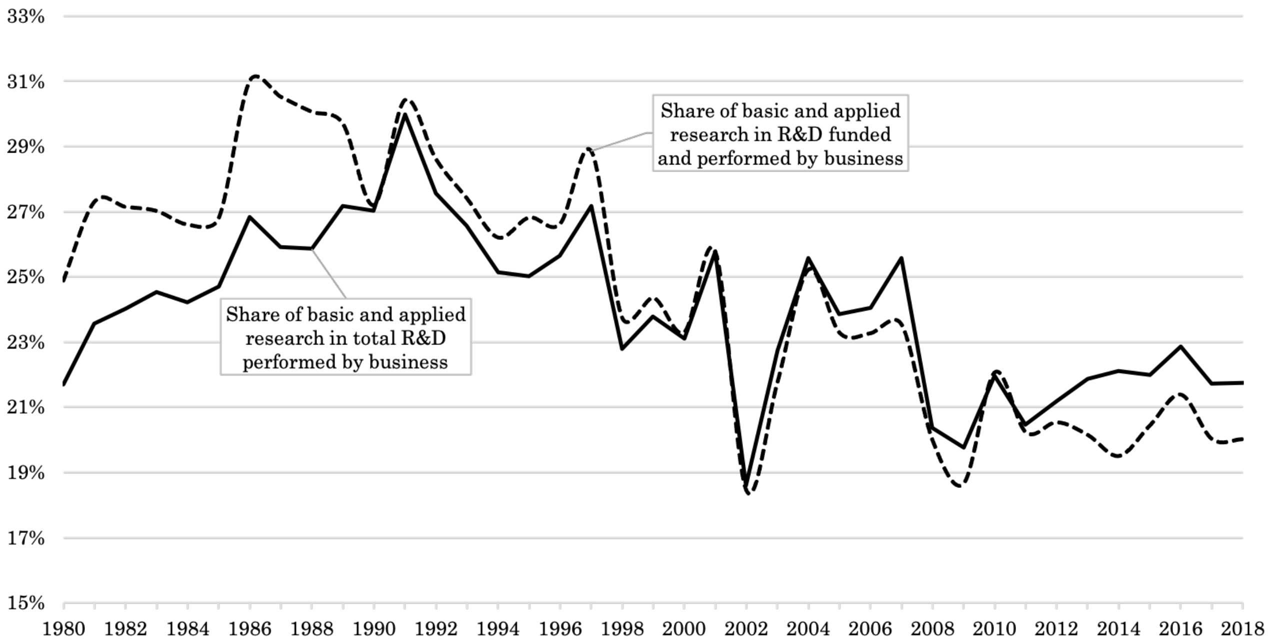
> Xerography was invented by Carlson in 1937, but it was only commercialized by Xerox in 1950. Furthermore, there was so little interest in this technology that during the few years surrounding commercialization, Xerox was able to invent and patent a whole range of related techniques, while there was hardly any activity by other institutions. [... By contrast] when Bednorz and Mueller announced their discovery of high-temperature superconductivity at the IBM Zurich lab in 1987, it took only a few weeks for groups at University of Houston, University of Alabama, Bell Labs, and other places to make important further discoveries. Thus even if high-temperature superconductivity had developed into a commercially significant field, IBM would have had to share the financial benefits with others who held patents that would have been crucial to developments of products.

investigation of the link between a possible decline of R&D productivity and the decline in scientific research represents a useful line of further inquiry.

A more obvious cost is that firms contribute less to the pool of knowledge available to advance innovation. That firms benefit from a public good, but are unwilling to contribute to it may be ironic, but certainly not surprising to economists. Our findings raise a different question: Why have spillovers become more significant over time? The obvious candidate answer, namely product market competition (Aghion et al. (2005)), has to contend with recent research pointing to growing market concentration and the rise of superstar firms (Autor et al.). It may well be that though firms enjoy market power, their stay at the top is more short-lived due to greater competition for the market (Segal and Whinston, 2007). Spillovers may also have increased due to advances in information technology and the speed and efficiency with which knowledge diffuses.

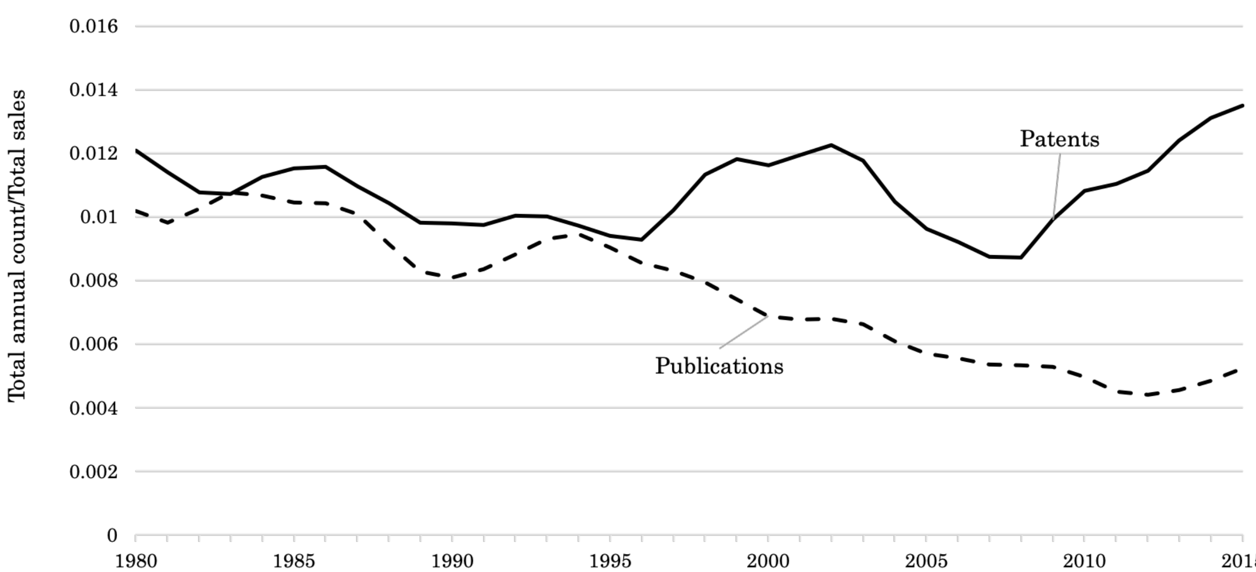
Changes in intellectual property protection may also affect spillovers (Galasso and Schankerman, 2014; Moser, 2005). Here, once again, there are contradictory impulses. On the one hand, it is widely acknowledged that intellectual property protection in the United States was strengthened in the 1980s. On the other hand, the last decade has seen a push back, with several court cases weakening patent protection. Moreover, there are significant differences across industries as well. Life sciences, where intellectual property protection is the strongest (Williams (2013)), have seen the smallest decline, perhaps because profit-reducing spillovers are the least widespread there, compared to sectors such as materials, chemicals, and information technology. Moreover, if scientific findings have become more broadly applicable, even without changes in the patent regime, the patents filed by a firm may cover a smaller fraction of the applications of its scientific discoveries. These speculations indicate that an important direction for future research is to understand what lies behind the increases in the volume and importance of knowledge spillovers across firms.

Figure 1: Research in Business R&D, 1980-2018



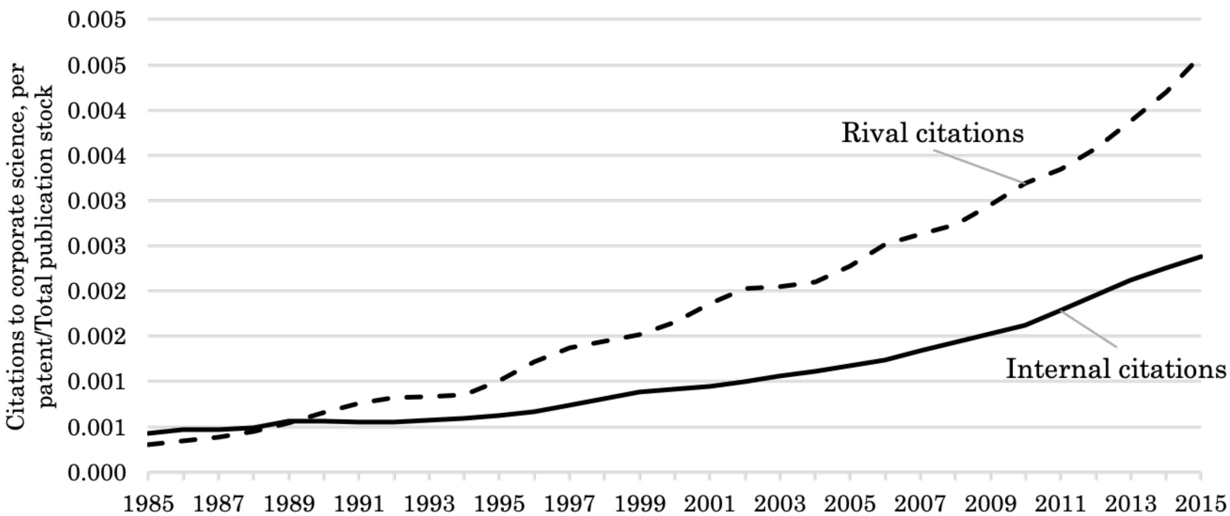
Notes: Data for this plot are generated from tables 2-4, National Science Foundation, National Center for Science and Engineering Statistics 2019. National Patterns of R&D Resources: 2017–18 Data Update. NSF 20-307. Alexandria, VA. Available at https://ncses.nsf.gov/pubs/nsf20307.

Figure 2: Trends in Corporate Scientific Publications and Patents, 1985-2015



Notes: The figure presents the sum of annual publications and patents divided by the sum of annual sales across our complete sample firms from 1980 through 2015.

Figure 3: Trends in Use of Science by Corporate Patents, 1985-2015



Notes: This figure presents trends in the propensity of corporate patents to cite corporate science, measured as the ratio between citations to science per corporate patent divided by total number of corporate publications in each year (y-axis values are multiplied by 1000). The sample is conditional on firm-year observations with at least one granted patent. Rival citations weight external citations by the product market proximity between the citing and cited firms.

References

Aghion, P., Bloom, N., Blundell, R., Griffith, R., and Howitt, P. (2005). Competition and innovation: An inverted-U relationship. The Quarterly Journal of Economics, 120(2):701–728.

Aghion, P. and Jaravel, X. (2015). Knowledge spillovers, innovation and growth. Economic Journal, 125(583):533–573.

Akcigit, U., Grigsby, J., and Nicholas, T. (2017). The rise of American ingenuity: Innovation and inventors of the golden age. NBER Working Papers: No. 23047.

Arora, A., Belenzon, S., and Patacconi, A. (2018). The decline of science in corporate R&D. Strategic Management Journal, 39(1):3–32.

Arora, A., Belenzon, S., Patacconi, A., and Suh, J. (2019). The Changing Structure of American Innovation: Some Cautionary Remarks for Economic Growth. NBER Working Papers: No. 25893.

Arrow, K. (1962). Economic welfare and the allocation of resources for invention. In The rate and direction of inventive activity: Economic and social factors, pages 609–626.

Audretsch, D. B. and Stephan, P. E. (1996). Company-scientist locational links: The case of biotechnology. American Economic Review, 86(3):641–652.

Autor, D., Dorn, D., Katz, L. F., Patterson, C., and Van Reenen, J. Cep discussion paper no 1482 revised may 2019 (replaces may 2017 version) the fall of the labor share and the rise of superstar firms.

Azoulay, P. (2002). Do pharmaceutical sales respond to scientific evidence? Journal of Economics Management Strategy, 11(4):551–594.

Azoulay, P., Li, D., Zivin, J., and Sampat, B. (2019). Public r&d investments and private-sector patenting: Evidence from nih funding rules. The Review of economic studies, 86(1):117–152.

Belenzon, S. (2012). Cumulative innovation and market value: evidence from patent citations. Economic Journal, 122(559):265–285.

Bessen, J. (2009). Nber pdp project user documentation. Unpublished documentation.

Bloom, N., Jones, C. I., Van Reenen, J., and Webb, M. (2017). Are ideas getting harder to find? NBER Working Papers: No. 23782.

Bloom, N., Schankerman, M., and Van Reenen, J. (2013). Identifying technology spillovers and product market rivalry. Econometrica, 81(4):1347–1393.

Blundell, R., Griffith, R., and van Reenen, J. (1999). Market Share, Market Value and Innovation in a Panel of British Manufacturing Firms. The Review of Economic Studies, 66(3):529–554.

Bryan, K. A., Ozcan, Y., and Sampat, B. (2020). In-text patent citations: A user's guide. Research Policy, 49(4):103946.

Bush, V. (1945). Science: The endless frontier. Transactions of the Kansas Academy of Science (1903-1945), 48(3):231–264.

Ceccagnoli, M. (2005). Firm heterogeneity, imitation, and the incentives for cost reducing r&d effort. The Journal of Industrial Economics, 53(1):83–100.

Cockburn, I. M. and Henderson, R. M. (1998). Absorptive capacity, coauthoring behavior, and the organization of research in drug discovery. The Journal of Industrial Economics, 46(2):157–182.

Cohen, W. M. and Levinthal, D. A. (1989). Innovation and learning: the two faces of R & D. Economic Journal, 99(397):569–596.

Cohen, W. M., Nelson, R. R., and Walsh, J. P. (2000). Protecting their intellectual assets: Appropriability conditions and why us manufacturing firms patent (or not). NBER Working Papers: No. 7552.

Dasgupta, P. and David, P. A. (1994). Toward a new economics of science. Research policy, 23(5):487–521.

Duguet, E. and MacGarvie, M. (2005). How well do patent citations measure flows of technology? evidence from french innovation surveys. Economics of Innovation and New Technology, 14(5):375–393.

Fleming, L. and Sorenson, O. (2004). Science as a map in technological search. Strategic management journal, 25(8-9):909–928.

Galasso, A. and Schankerman, M. (2014). Patents and cumulative innovation: Causal evidence from the courts. The Quarterly Journal of Economics, 130(1):317–369.

Griffith, R., Harrison, R., and Van Reenen, J. (2006). How special is the special relationship? using the impact of us R&D spillovers on UK firms as a test of technology sourcing. American Economic Review, 96(5):1859–1875.

Griffith, R., Redding, S., and Reenen, J. V. (2004). Mapping the two faces of R&D: Productivity growth in a panel of oecd industries. Review of Economics and Statistics, 86(4):883–895.

Griliches, Z. (1986). Productivity, R and D, and basic research at the firm level in the 1970's. American Economic Review, 76(1):141–154.

Hall, B. H. (1993). R&D tax policy during the 1980s: success or failure? Tax Policy and the Economy, 7:1–35.

Hall, B. H., Jaffe, A., and Trajtenberg, M. (2005). Market value and patent citations. RAND Journal of Economics, pages 16–38.

Hall, B. H., Jaffe, A. B., and Trajtenberg, M. (2001). The NBER patent citation data file: Lessons, insights and methodological tools. NBER Working Papers: No. 8498.

Hall, R. E. and Jorgenson, D. W. (1967). Tax policy and investment behavior. American Economic Review, 57(3):391–414.

Henderson, R. and Cockburn, I. (1994). Measuring competence? exploring firm effects in pharmaceutical research. Strategic Management Journal, 15:63–84.

Hicks, D. (1995). Published papers, tacit competencies and corporate management of the public/private character of knowledge. Industrial and Corporate Change, 4(2):401–424.

Jaffe, A. B. (1986). Technological opportunity and spillovers of R&D: evidence from firms' patents, profits and market value. NBER Working Papers: No. 1815.

Jaffe, A. B. and Trajtenberg, M. (2002). Patents, citations, and innovations: A window on the knowledge economy. MIT press.

Jaffe, A. B., Trajtenberg, M., and Henderson, R. (1993). Geographic localization of knowledge spillovers as evidenced by patent citations. the Quarterly journal of Economics, 108(3):577–598.

Lucking, B., Bloom, N., and Van Reenen, J. (2018). Have R&D spillovers changed? NBER Working Papers: No. 24622.

Mansfield, E. (1980). Basic research and productivity increase in manufacturing. The American Economic Review, 70(5):863–873.

Moser, P. (2005). How do patent laws influence innovation? evidence from nineteenth-century world's fairs. American Economic Review, 95(4):1214–1236.

Mowery, D. C. (2009). Plus ca change: Industrial R&D in the "third industrial revolution". Industrial and Corporate Change, 18(1):1–50.

Nagaraj, A. and Stern, S. (2020). The economics of maps. Journal of Economic Perspectives, 34(1):196–221.

Nelson, R. R. (1959). The simple economics of basic scientific research. Journal of Political Economy, 67(3):297–306.

Nelson, R. R. (1982). The role of knowledge in r&d efficiency. The Quarterly Journal of Economics, 97(3):453–470.

Odlyzko, A. (1995). The decline of unfettered research. Unpublished, University of Minnesota, http://www. dtc. umn. edu/˜ odlyzko/doc/decline. txt.

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

Rosenberg, N. (1990). Why do firms do basic research (with their own money)? Research policy, 19(2):165–174.

Segal, I. and Whinston, M. D. (2007). Antitrust in innovative industries. American Economic Review, 97(5):1703–1730.

Stern, S. (2004). Do scientists pay to be scientists? Management Science, 50(6):835–853.

Williams, H. L. (2013). Intellectual property rights and innovation: Evidence from the human genome. Journal of Political Economy, 121(1):1–27.

Wilson, D. J. (2009). Beggar thy neighbor? the in-state, out-of-state, and aggregate effects of R&D tax credits. Review of Economics and Statistics, 91(2):431–436.

Appendix

1 Analytical framework

The appendix provides more details on the analytical framework in the text. For completeness, we repeat some details here. There are two firms, indexed by 0 and 1. Both compete in the product market, and both invest in innovation, d_0 and d_1 respectively. For now, only firm 0 is assumed to have the capability to invest in research, r_0. Research reduces the cost of innovation of firm 0. However, we also allow research to spill-out to the rival firm, and reduce the cost of innovation of the rival firm.

1.1 Setup

There are three stages. In stage 3, the firms compete in the product market. Their product market performance depends the quality of their products and the cost of producing them. We assume that cost and quality depend upon the innovation output, d_i.

Assuming some Nash Equilibrium in the product market, we can write the reduced form profit function of firm 0 as $\Pi_0(d_0, d_1)$. The profit for firm 1 is similarly $\Pi_1(d_0, d_1)$. We assume that $\Pi_0(d_0, d_1)$ is increasing in the first argument and decreasing in the second. We further assume that $\Pi_0(d_0, d_1)$ is concave it its arguments. These assumptions mean that the firm's profit increases in its innovation output, albeit at a diminishing rate, and that innovation by rivals reduces the profits of the focal firm, also at a diminishing rate.[35] Further we assume that $\frac{\partial^2 \Pi_0{}^2}{\partial d_0} = c_0 < 0$, $\frac{\partial^2 \Pi_1}{\partial {d_1}^2} = c_1 < 0$ and $\frac{\partial^2 \Pi_0}{\partial {d_1}^2} \leq 0$.

In stage 2, firms choose their innovation output. Firm 0 chooses d_0 and firm 1 chooses d_1. Innovation output d_0 is produced at a cost given by $\phi(r_0; \lambda)d_0$. Recall that we assume $\phi(r_0; \lambda)$ decreases at a diminishing rate in r_0, and that $\frac{\partial^2 \phi}{\partial \lambda \partial r_0 < 0}$.

In stage 1, the firm 0 choose its research investments, and the cost of research is modelled simply as γr_0. The payoff functions of the firms can be written as

$$v_0 = \Pi_0(d_0, d_1) - (\phi(r_0; \lambda)d_0 - \gamma r_0 \tag{1}$$
$$v_1 = \Pi_1(d_0, d_1) - s(r_0; \theta)d_1 \tag{2}$$

[35] This rules out instances where innovation by rivals has beneficial externalities such as expanding the market. The assumption about diminishing rates is helpful in signing the results.

1.2 Stage 2: Innovation

We assume a stable Nash Equilibrium exists at the innovation stage. The first order conditions are

$$\begin{aligned}\frac{\partial \Pi_0(d_0,d_1)}{\partial d_0} - \phi(r_0;\lambda) &= 0\\ \frac{\partial \Pi_1(d_1,d_0)}{\partial d_1} - s(r_0;\theta) &= 0\end{aligned} \tag{3}$$

the second order condition is

$$\begin{aligned}\frac{\partial^2 \Pi_0(d_0,d_1)}{\partial {d_0}^2} &< 0\\ \frac{\partial^2 \Pi_1(d_0,d_1)}{\partial {d_1}^2} &< 0\end{aligned}$$

1.2.1 Spillovers to rivals and innovation

At the second stage, consider an increase in the ability of the rival to learn from external knowledge, represented by $\Delta\theta$. The changes in innovation output are Δd_0 & Δd_1 respectively.

$$\begin{aligned}\frac{\partial^2 \Pi_0(d_0,d_1)}{\partial {d_0}^2}\Delta d_0 + \frac{\partial^2 \Pi_0(d_0,d_1)}{\partial d_0 \partial d_1}\Delta d_1 &= 0\\ \frac{\partial^2 \Pi_1(d_0,d_1)}{\partial {d_1}^2}\Delta d_1 + \frac{\partial^2 \Pi_1(d_1,d_0)}{\partial d_1 \partial d_0}\Delta d_0 &= \frac{\partial s}{\partial \theta}\Delta\theta\end{aligned}$$

Solving for the changes in innovation output we get

$$\begin{aligned}\frac{\partial d_0}{\partial \theta} &= \frac{1}{D}\frac{\partial^2 \Pi_1(d_1,d_0)}{\partial d_1 \partial d_0}\frac{\partial s}{\partial \theta}\\ \frac{\partial d_1}{\partial \theta} &= -\frac{1}{D}\frac{\partial^2 \Pi_0(d_0,d_1)}{\partial {d_0}^2}\frac{\partial s}{\partial \theta} \geq 0\end{aligned} \tag{4}$$

Where

$$D = \frac{\partial^2 \Pi_0(d_0,d_1)}{\partial {d_0}^2}\frac{\partial^2 \Pi_1(d_0,d_1)}{\partial {d_1}^2} - \frac{\partial^2 \Pi_0(d_0,d_1)}{\partial d_0 \partial d_1}\frac{\partial^2 \Pi_1(d_1,d_0)}{\partial d_1 d_0} \geq 0 \text{ at a stable equilibrium}$$

In 4, the inequality holds because profits are increasing at a diminishing rate, i.e., $\dfrac{\partial^2 \Pi_0(d_0,d_1)}{\partial {d_0}^2}$. As expected, the innovation output of the rival increases as it learns more. However, the change in focal firm's innovation depends on whether innovation outputs are strategic complements or substitutes. The innovation output of the focal firm decreases if it is a strategic substitute and increases otherwise.

Special case: no strategic interactions If there are no strategic interactions in the innovation game, $\frac{\partial^2 \Pi_1(d_1,d_0)}{\partial d_1 d_0} = 0$. The assumption of no strategic interaction in the innovation game, and imposing $\frac{\partial^2 \Pi_1}{\partial {d_1}^2} = c_1 < 0$ yields the following

$$
\begin{aligned}
\frac{\partial d_1}{\partial \theta} &= \frac{\partial s}{\partial \theta}\left(\frac{\partial^2 {\Pi_1}^2}{\partial d_1}\right)^{-1} = \frac{\partial s}{\partial \theta}\frac{1}{c_1} \geq 0 \\
\frac{\partial^2 d_1}{\partial \theta \partial r_0} &= \frac{\partial^2 s}{\partial \theta \partial r_0}\frac{1}{c_1} \geq 0 \\
\frac{\partial d_0}{\partial \theta} &= 0
\end{aligned}
\tag{5}
$$

1.2.2 Internal use and innovation

Consider the case when the focal firm is better able to use internal research in invention, represented by $\Delta\lambda$. The changes in innovation output are Δd_0 & Δd_1 respectively.

$$
\begin{aligned}
\frac{\partial^2 \Pi_0(d_0, d_1)}{\partial {d_0}^2}\Delta d_0 + \frac{\partial^2 \Pi_0(d_0, d_1)}{\partial d_0 \partial d_1}\Delta d_1 &= \frac{\partial \phi}{\partial \lambda}\Delta\lambda \\
\frac{\partial^2 \Pi_1(d_0, d_1)}{\partial {d_1}^2}\Delta d_1 + \frac{\partial^2 \Pi_1(d_1, d_0)}{\partial d_1 \partial d_0}\Delta d_0 &= 0
\end{aligned}
$$

Solving for the changes in innovation output we get

$$
\begin{aligned}
\frac{\partial d_0}{\partial \lambda} &= \frac{1}{D}\frac{\partial^2 \Pi_1(d_1, d_0)}{\partial {d_1}^2}\frac{\partial \phi}{\partial \lambda} \geq 0 \\
\frac{\partial d_1}{\partial \lambda} &= \frac{1}{D}\frac{\partial^2 \Pi_0(d_0, d_1)}{\partial d_0 \partial d_1}\frac{\partial \phi}{\partial \lambda}
\end{aligned}
\tag{6}
$$

Special case: Internal use with no strategic interactions

$$
\begin{aligned}
\frac{\partial d_0}{\partial \lambda} &= -\frac{1}{c_0}\frac{\partial \phi}{\partial \lambda} \geq 0 \\
\frac{\partial d_1}{\partial \lambda} &= 0
\end{aligned}
\tag{7}
$$

1.2.3 focal firm research and innovation

The effect on innovation outputs is

$$
\begin{aligned}
\frac{\partial d_0}{\partial r_0} &= \frac{1}{D}\left(\frac{\partial^2 \Pi_1(d_1, d_0)}{\partial {d_1}^2}\frac{\partial \phi}{\partial r_0} - \frac{\partial^2 \Pi_0(d_1, d_0)}{\partial d_1 \partial d_0}\frac{\partial s}{\partial r_0}\right) \\
\frac{\partial d_1}{\partial r_0} &= \frac{1}{D}\left(\frac{\partial^2 \Pi_0(d_1, d_0)}{\partial {d_0}^2}\frac{\partial s}{\partial r_0} - \frac{\partial^2 \Pi_1(d_1, d_0)}{\partial d_0 \partial d_1}\frac{\partial \phi}{\partial r_0}\right)
\end{aligned}
\tag{8}
$$

If innovation outputs are strategic complements, both expressions in 8 are positive. However, with strategic substitutes, one or the other (but not both) may decrease. When there are no strategic interactions in the innovation market then both innovation outputs will go up.

Formally, with no strategic interactions

$$
\begin{aligned}
\frac{\partial d_0}{\partial r_0} &= \frac{1}{c_0}\frac{\partial \phi}{\partial r_0} \geq 0 \\
\frac{\partial d_1}{\partial r_0} &= \frac{1}{c_1}\frac{\partial s}{\partial r_0} \geq 0 \\
\frac{\partial^2 d_1}{\partial r_0 \partial \theta} &= \frac{1}{c_1}\frac{\partial^2 s}{\partial r_0 \partial \theta} \geq 0
\end{aligned} \tag{9}
$$

1.3 Stage 1: Research

For firm 0, the first order condition for optimal r_0, is

$$
\frac{\partial \Pi(d_0, d_1)}{\partial d_1}\frac{\partial d_1}{\partial r_0} - \frac{\partial \phi}{\partial r_0} d_0 = \gamma
$$

By assumption, firm 1 does not invest in research.

1.3.1 Internal use

At any interior maximum, the sign of $\frac{\partial r_0}{\partial \lambda}$ is the same as the sign of $\frac{\partial^2 v_0}{\partial \lambda \partial r_0}$.

$$
\frac{\partial^2 v_0}{\partial \lambda \partial r_0} = -d_0 \frac{\partial^2 \phi}{\partial \lambda \partial r_0} - \frac{\partial \phi}{\partial \lambda}\frac{\partial d_0}{\partial r_0} + \frac{\partial \Pi_0}{\partial d_1}\frac{\partial^2 d_1}{\partial \lambda \partial r_0} + \frac{\partial^2 \Pi_0}{\partial {d_1}^2}\frac{\partial d_1}{\partial \lambda}\frac{\partial d_1}{\partial r_0} + \frac{\partial^2 \Pi_0}{\partial d_1 \partial d_0}\frac{\partial d_1}{\partial \lambda}\frac{\partial d_0}{\partial r_0} \tag{10}
$$

No strategic interactions implies

$$
\frac{\partial^2 v_0}{\partial \lambda \partial r_0} = -d_0 \frac{\partial^2 \phi}{\partial \lambda \partial r_0} - \frac{\partial \phi}{\partial \lambda}\frac{\partial d_0}{\partial r_0} \geq 0 \tag{11}
$$

With strategic interactions, 10 cannot be signed without further assumptions about the slope of the reaction function.

1.3.2 Spillouts

Recall that $\frac{\partial v_0}{\partial \theta} = \frac{\partial \Pi_0}{\partial d_1}\frac{\partial d_1}{\partial \theta}$, which implies

$$
\frac{\partial^2 v_0}{\partial r_0 \partial \theta} = \frac{\partial^2 \Pi_0}{\partial {d_1}^2}\frac{\partial d_1}{\partial r_0}\frac{\partial d_1}{\partial \theta} + \frac{\partial^2 d_1}{\partial r_0 \partial \theta}\frac{\partial \Pi_0}{\partial d_1} + \frac{\partial^2 \Pi_0}{\partial d_1 \partial d_0}\frac{\partial d_1}{\partial \theta}\frac{\partial d_0}{\partial r_0} \tag{12}
$$

No strategic interactions implies

$$
\begin{aligned}
\frac{\partial^2 v_0}{\partial \theta \partial r_0} &= \frac{\partial^2 \Pi_0}{\partial {d_1}^2}\frac{\partial d_1}{\partial r_0}\frac{\partial d_1}{\partial \theta} + \frac{\partial^2 d_1}{\partial r_0 \partial \theta}\frac{\partial \Pi_0}{\partial d_1} \\
&= \frac{\partial s}{\partial r_0}\frac{\partial s}{\partial \theta}\frac{1}{c_1} + \frac{1}{c_1}\frac{\partial^2 s}{\partial r_0 \partial \theta}\frac{\partial \Pi_0}{\partial d_1} \leq 0
\end{aligned} \tag{13}
$$

With strategic interactions, 12 requires additional assumptions about the slope of the rival's reaction function.

2 Additional details on the IV and first stage results

As in Bloom et al. (2013), we adopt the Hall-Jorgenson's user cost of capital for firm i in state s at time t (Hall and Jorgenson, 1967): $\frac{(1-D_{it})}{(1-\tau_{st})}\left[I_t+\delta-\frac{\Delta p_t}{p_{t-1}}\right]$, where D_{it} is the discounted tax credits and depreciation allowance and τ_{st} is firm's tax rate. $\left[I_t+\delta-\frac{\Delta p_t}{p_{t-1}}\right]$ is common to all firms and is therefore ignored, and only the component, $\rho^P_{it}=\frac{(1-D_{it})}{(1-\tau_{st})}$ is considered. ρ^P_{it} is further decomposed into federal and state components, where the state-level instrument is constructed using the estimates of state-specific R&D tax rates from Wilson (2009) and each firm's distribution of patent inventors across states. Formally, the R&D tax rate for firm i in year t based on state tax credit is: $\rho^S_{it}=\sum_s \theta_{ist}\rho^S_{st}$, where θ_{ist} is 10-year moving average of the fraction of firm i's patent inventors in state s in year t and ρ^S_{st} is the tax rate for state s in year t. The notation is borrowed directly from Bloom et al. (2013).

As in Bloom et al. (2013) and Hall (1993), the federal tax credit component is calculated by multiplying the difference between firm-specific base R&D expenditure and actual R&D expenditure with the appropriate credit rate. The definition of base R&D expenditure has changed in 1990 from a maximum of prior 3-year rolling average of R&D expenditures (or 50% of current year's expenditure) to R&D to sales ratios between 1984 and 1988 times current year's sales (up to a ratio of 0.16).

Table A2 presents the first stage results of regressing *Rival citations* against $Rival\hat{P}AT_{it-2}$ for market value (Column 1) and publication (Column 2) equations. Second stage results are

Table A1: Main Variables Definition

Variable	Description	Data Source
Market value	Following Griliches (1981), market value per firm-year is defined as the sum of the values of common stock, preferred stock, and total debt net of current assets. Tobin's-Q is defined as the ratio of market value to assets.	U.S. Compustat
Tobin's-Q	Tobin's-Q is the ratio of market value to assets.	U.S. Compustat
Assets	The book value of capital includes net plant, property and equipment, inventories, investments in unconsolidated subsidiaries, and intangibles other than R&D.	U.S. Compustat
Publication stock	Publication stock per firm-year is calculated using a perpetual inventory method with a 0.15% growth rate and 15 percent depreciation rate (Hall et al., 2005), such that the Publication stock in year t for firm i is calculated by: $Publication\_stock_t = Pub_t + (1-\delta)Publications\_stock_{t-1}$, where Pub_t is the focal firm's publication count in year t. $\delta = 0.15$.	Web of Science articles, covered in "Science Citation Index" and "Conference Proceedings Citation Index-Science", 1980-2016
Patent Stock	Patent stock per firm-year is calculated using a perpetual inventory method with a 0.15% growth rate and 15 percent depreciation rate (Hall et al., 2005), such that the Patent stock in year t for firm i is calculated by: $Patent\_stock_t = Patent_t + (1-\delta)Patent\_stock_{t-1}$, where $Patent_t$ is the focal firm's patent count in year t. $_=$0.15.	United States Patent and Trademark Office (USPTO) patents granted 1980-2015 from PatStat database
R&D stock	R&D stock per firm-year is calculated using a perpetual inventory method with a 0.15% growth rate and 15 percent depreciation rate (Hall et al., 2005), such that the R&D stock, GRD, in year t is $GRD_t = R_t + (1-\delta)GRD_{t-1}$ where Rt is the focal firm's R&D expenditure in year t based on Compustat data and $\delta = 0.15$.	U.S. Compustat

presented in Column 4 in Table 4 and Column 5 in Table 5, respectively.

Table A2: Instrumental Variable Estimation: Federal and State R&D Tax Credit

	(1)	(2)
Specification:	Market Value (Column 4, Table 4)	Publications (Column 4, Table 4)
Dependent variable:	ln(Cumulative RIVAL citations)$_{t-2}$	ln(External citation, RIVAL)$_{t-2}$
	First Stage	First Stage
$Predicted\ RIVAL\ patents_{t-2}$	0.001**	0.001**
	(0.001)	(0.001)
$ln(Cumulative\ internal\ citations)_{t-2}$	0.538**	
	(0.009)	
$Publication\ stock_{t-2}/Assets_{t-2}$	0.010**	
	(0.003)	
$Patent\ stock_{t-2}/Assets_{t-2}$	-0.001	
	(0.002)	
$ln(Internal\ citation\ to\ own\ publications)_{t-2}$		0.313**
		(0.009)
$ln(R\&D\ stock)_{t-2}$		0.008**
		(0.002)
$ln(Patent\ stock)_{t-2}$		0.016**
		(0.002)
$ln(Publication\ stock)_{t-2}$		0.016**
		(0.002)
Firm fixed-effects	Yes	Yes
Year dummies	Yes	Yes
Dependent variable sample average	4.1	5.47
Number of firms	3,374	3,521
Observations	39,667	45,188

Notes: This table presents the first stage of an instrumental variable estimation for the effect of RIVAL citations on publications, Tobin's Q, and R&D productivity. Data on Federal and State R&D tax credit is based on Lucking, Bloom, Van Reenen (2018). Standard errors in brackets are robust to arbitrary heteroscedasticity.

Table A3 presents the estimation results of regressing rival patents, which generates our instrument $Rival\hat{P}AT_{it-2}$, against rivals' cost of R&D (Column 1). Column 2 replicates the original BSV specification for completeness, which regresses R&D expenditures against the cost of R&D.

Table A3: Predicting Patents and R&D using Federal and State R&D Tax Credit

	(1)	(2)
Dependent variable:	ln(1+Number of patents)	ln(R&D)
$ln(Federal\ tax\ credit\ component\ of\ R\&Duser\ cost)$	-2.202**	-4.557**
	(0.450)	(0.335)
$ln(State\ tax\ credit\ component\ of\ R\&D\ user\ cost)$	-0.474**	-0.389**
	(0.128)	(0.101)
Firm fixed-effects	Yes	Yes
Year dummies	Yes	Yes
Joint F-test of the tax credits	F=19.10	F=101.16
Dependent variable sample average	30.20	109.57
Number of firms	3,451	3,451
Observations	42,642	42,642
R-squared	0.83	0.92

Notes: Data on Federal and State R&D tax credit is based on Lucking, Bloom, Van Reenen (2018). Restricted to firm-years with available data. Standard errors in brackets are robust to arbitrary heteroscedasticity.

ONLINE DATA APPENDIX FOR "KNOWLEDGE SPILLOVERS AND CORPORATE INVESTMENT IN SCIENTIFIC RESEARCH" / A. ARORA, S. BELENZON & L. SHEER[1]

This appendix describes the methodology used to construct our database of publicly listed U.S. headquartered firms matched to assignees of patents from the United States Patent and Trademark Office (USPTO) and scientific publications from the Web of Science for the period 1980-2015. Data users should cite the NBER version of the paper (Working Paper 23187).

We introduce a major data extension and improvement to the historical NBER patent dataset (Hall, Jaffe, and Trajtenberg and others, 2001; Bessen, 2006), which should be valuable for all researchers working with patent and publication data. In updating the data to match between Compustat and patents to 2015, we address two major challenges: name changes and ownership changes. These challenges are central to how patents are assigned to firms over time. To be consistent over the sample period, we reconstruct the complete historical data covered in the NBER data files. About 30% of the Compustat firms in our sample change their name at least once. Accounting for name changes improves the accuracy and scope of matches to patents (and other assets), ownership structure, and dynamic reassignments of GVKEY codes to companies. Dynamic reassignment means that, for instance, if a sample firm merges with another firm, the patents of the merged firm are included in the stock of patents linked to the Compustat record from that point onward, but not before. For ownership and subsidiary data, we rely on a wide range of M&A data, including SDC, historical snapshots of ORBIS files for 2002-2015, 10-K SEC filings, and NBER2006 as well as perform extensive manual checks that help us uncover firms' structure and ownership changes before proceeding to the patent match. Thus, we have extended and improved the NBER patent data. In this Appendix, we document our data construction work, present several examples ("case studies"), and outline the improvements we made to existing NBER historical patent data.

We combine data from six main sources: (i) company and accounting information from U.S. Compustat 2018, (ii) scientific publications from Web of Science, (iii) patents and their non-patent literature (NPL) citations from PatStat; (iv) subsidiary data from historical snapshots of ORBIS files for 2002-2015; (v) mergers and acquisition data from SDC Platinum and (vi) company name changes from WRDS's "CRSP Monthly Stock".

We match (i) corporate subsidiaries to Compustat ultimate owner (UO) firms; (ii) acquisition data to Compustat companies and their related subsidiaries; (iii) patent data to Compustat companies and their related subsidiaries; (iv) scientific publications to Compustat companies and their related subsidiaries; and (v) patent citations to scientific articles. We discuss the details of our methodology below.

A. ACCOUNTING DATA PANEL

Our methodology builds and improves on the NBER patent database (Hall et al., 2001; Bessen, 2006), by extending the time period by a decade (now from 1980 to 2015) and implementing several methodological improvements for the complete sample period.

We start with all North American Compustat records obtained through WRDS in August 2018 and select companies with active records and positive R&D expenses for at least one year during our sample period, 1980-2015[2]. We exclude firms that are not headquartered in the United States based on their current headquarter location. After matching the remaining firms to patent assignees from the USPTO, we further restrict our sample to ultimate-owner[3] (UO) Compustat firms with at least one patent during our sample period. A UO firm enters the sample once it is publicly traded and remains in our data until the end of the sample period unless it is acquired, dissolved, or taken private. All UO firms in our final sample have at least 3 consecutive years of active records in Compustat. Our final estimation sample consists

---

[1] Arora: Duke University, Fuqua School of Business and NBER, ashish.arora@duke.edu; Belenzon: Duke University, Fuqua School of Business and NBER, sharon.belenzon@duke.edu; Sheer: Duke University, Fuqua School of Business, lia.sheer@duke.edu;
We thank Jim Bessen, Nick Bloom, Wesley Cohen, Alfonso Gambardella, Bronwyn Hall, David Hounshell, Adam Jaffe, Brian Lucking, David Mowery, Mark Schankerman, Scott Stern, Manuel Trajtenberg, John Van Reenen, and seminar participants at NBER summer institute and NBER Innovation Information Initiative for helpful comments. We thank Bernardo Dionisi, Honggi Lee, Dror Shvadron, and JK Suh for excellent research assistance. All remaining errors are ours.

[2] We define an active record as a year with positive common shares traded (CSHTR_F). We do this to avoid including years with data based on prospectus submitted by the focal company as part of the filing process before the firm became publicly traded.

[3] Compustat database does not link parent companies to subsidiaries, however we supplement the data with subsidiary level data. Following NBER 2006, we aggregate the data to the parent company level which we call ultimate owner (UO).

of an unbalanced panel of 4,420 UO firms and 58,245 firm-year observations.[4] The process of defining a UO firm and its related subsidiaries is explained below.

We face several challenges when working with Compustat data, as following.

1) **Unique company identifier over time.** Compustat uses GVKEY to track companies over time[5]. However, a single company may correspond to multiple GVKEYs within the Compustat database due to changes in ownership and other accounting changes over the sample period (e.g., the pet food company Ralston Purina is listed under two different GVKEYs: (i) 1980-1993 under "RALSTON PURINA-CONSOLIDATED" (GVKEY 008935) and (ii) 1993-2000 under "RALSTON PURINA CO" (GVKEY 028701)). The Compustat database does not link related company identifiers, making it difficult to track companies over time only based on GVKEY.

2) **Name changes.** While scientific publications and patent records contain the owner's name at the time of their publication, companies appear in the Compustat file under their most current name with no records of previous names. Company names may change over the course of our sample period due to general name changes[6] and M&As[7], including reverse takeovers[8]. About 30% of the Compustat firms in our sample change their name at least once. A company with a name change (which might have been accompanied by an ownership change) without a corresponding change in its GVKEY in Compustat may lead us to assign the record incorrectly to its most recent owner for the complete sample period. Without historical information on the record's ownership, we cannot correctly link patents and scientific publications to their relevant financial records.

3) **Ownership structure.** A parent company and a majority-owned subsidiary may have different identification numbers and records within Compustat. While innovative activities typically take place inside numerous subsidiaries, we aggregate the data to the UO level. Since the Compustat database does not link parent companies and majority-owned publicly traded subsidiaries, comprehensive manual checks and investigations are required.[9] We further link non-publicly traded subsidiaries to their UO firm based on historical snapshots of ORBIS files.

4) **Changes in ownership.** Ownership of a firm can change throughout the sample period due to mergers, acquisitions, and spinoffs[10]. While firms typically stop being traded independently after an M&A, their existing stock of publications and patents must be reassigned to the new owner. Moreover, in many cases, the acquiring entities continue to file patents and produce scientific publications post-acquisition. Compustat data do not provide information on ownership changes. Thus, we rely on SDC Platinum's M&A data and ORBIS to track ownership changes at the UO level as well as at the subsidiary level. Using historical snapshots of ORBIS files for 2002-2015, we are able not only to identify ownership changes at the subsidiary level but also new subsidiaries and changes in subsidiary names.

We implement the following procedures to manage these challenges.

---

[4] See "panel_do.do" file for exact details on the construction of the final panel file.

[5] GVKEY code remains the same, regardless of changes in TICKER, CUSIP, and firm names and thus is preferred on the later as a firm identifier for Compustat records. Compustat database only provides the most recent TICKER, CUSIP and name for each security with no historical info available.

[6] e.g., name abbreviations (for example, "MINNESOTA MINING AND MANUFACTURING" changed its name in 2002 to "3M"),

[7] e.g., "WESTINGHOUSE ELECTRIC CORP" (GVKEY 011436) purchased "CBS INC" in 1995 and changed its own name to "CBS CORPORATION" in 1997 keeping the same GVKEY Compustat firm identifier.

[8] e.g., in 1993 the private company Dentsply International Inc acquired the public company GENDEX CORPORATION (GVKEY 013700) in a reverse takeover and became publicly traded under the "DENTSPLY INTERNATIONAL INC" name and the original GVKEY.

[9] e.g., Thermo Electron's publicly traded majority-owned spun-out subsidiaries (all of which returned to be privately owned after 1999) need to be accounted under the parent company THERMO ELECTRON CORP (GVKEY 010530) for the complete period.

[10] e.g., "AT&T CORP" (GVKEY 001581) stopped being traded independently in 2005 after it was acquired by "SBC COMMUNICATIONS INC" (GVKEY 009899) which in turn changed its own name to "AT&T INC".

I. NAME CHANGES

One of our key contributions is identifying name changes of Compustat firms over the sample years 1980-2015. To the best of our knowledge, this has not been done consistently for a broad range of companies across many industries over a third of a century. Past research mainly considers the name that appears for each record in the most recent Compustat file (CONM variable) as the relevant name for the complete period the security was traded. The variable CONM, however, is the current name of the Compustat record as of the date the file was downloaded with no historical name information provided by Compustat. As shown above, company name changes may not be accompanied by changes in the original GVKEY firm identifier on Compustat, leading to assigning a record to its most recent holder for the complete sample period. Matching the original assignee name to a current Compustat file can result in misallocation of patents and publications. As companies change names, we wish to carry forward past patents and publications assigned to the original name as well as make sure that new patents and publications are assigned to the correct UO firm. Instead of building on the most recent Compustat name, we link our Compustat records to WRDS's "CRSP Monthly Stock" file, which records historical names for each month the security was traded and perform extensive manual checks using SEC filings to validate all related names for our sample period. We find that in our sample, 30 percent of Compustat records have more than one related name[11]. Accounting for all historical names significantly improves the accuracy and scope of the matches we perform across various databases as well as the linkage to relevant financial data. We elaborate on our name change methodology below, using several examples.

Example 1: SEALED POWER and GENERAL SIGNAL

The following example underscores the mismatching consequences of not accounting properly for name and ownership changes and how it affects the existing NBER patent data.

Up to the year 1998, SEALED POWER and GENERAL SIGNAL are two distinct entities. Historical Compustat records include the following records for these companies up to 1998:

1) GVKEY 9556, related names:

    i. SEALED POWER CORP (1962-1988) – original name
    ii. SPX CORP (1988-1997) -name changes retroactively in Compustat

2) GVKEY 5087, related name: GENERAL SIGNAL CORP (1950-1997)

In 1998, SPX Corp acquired General Signal Corp in a reverse merger transaction, and General's GVKEY (5087) became the new security of SPX traded retroactively under the new name "SPX CORP". At the same time, the original SPX records are renamed retroactively in Compustat as "SPX CORP-OLD" and stopped being traded. Current Compustat records include the following records for these companies for the complete period they are traded:

1) GVKEY 9556, related name: SPX CORP-OLD

2) GVKEY 5087, related name: SPX CORP

Our approach is to treat these GVKEYs as two separate companies up to 1997 accounting for all relevant names (SEALED POWER CORP, SPX CORP for GVKEY 9556 and GENERAL SIGNAL CORP for GVKEY 5087) in our matches and to connect the SPX CORP name to General's original GVKEY (5087) only from 1998.

When we examine the NBER 2006 patent dataset, we find that the two companies are collapsed under the same company (same PDPCO id) and that for the purpose of Compustat accounting information General's original GVKEY (5087) is used for the complete period while the original SPX GVKEY (9556) is disregarded:

[11] This is comparable to the findings of Wu (2010), who finds that during 1925-2000 over 30% of CRSP-listed firms changed their names at some point after going public. For name changes occurring between 1980-2000 the paper finds that the top 3 reason for name changes are: (i) M&As & restructure activity (36%); (ii) change in focus of operation (17%); (iii) brand or subsidiary name adoption (12%)

Table 1. Data for SPX Corp in NBER 2006

current name	gvkey	firstyr	lastyr	pdpco	pdpseq	begyr	endyr
SPX CORP	5087	1950	2006	5087	1	1950	2006
SPX CORP-OLD	9556	1962	1997	5087	-1		

Practically, this means that all the patents of SPX CORP are matched to General's financial data up to 1998. To verify, we tracked the NBER files and confirmed that indeed SPX patents pre-1998 are matched to General's GVKEY. Moreover, patents related to "GENERAL SIGNAL CORP" (757 patents without considering related subsidiaries) as well as "SEALED POWER CORP" (36 patents without considering related subsidiaries) are located in the 2006 NBER raw patent match but are not assigned to any Compustat record.

The NBER patent data file does not track ownership and name changes of GVKEYs over time. However, as shown in this example, using the current Compustat name can be misleading. The availability of data on historical name changes enables us to have a better understanding of the firms included in our sample and their origin. We are able to improve the accuracy of their match to the different databases (by using the complete history of firm names) and their linkage to relevant financial data. To be consistent over the sample period, we reconstruct the complete historical data covered in the NBER data files.

Compiling historical names

To locate historical names, we use the WRDS's "CRSP Monthly Stock" file, which includes historical monthly information on names for each security alongside its historical CUSIP code and a unique permanent security identification number assigned by CRSP, the PERMNO code, which is kept constant throughout the trading period regardless of changes in name or capital structure.[12] We compute for each name the starting and end years based on their trading dates in the "CRSP Monthly Stock" file.

Using WRDS "CRSP/Compustat Merged Database - Linking Table", we link each PERMNO to Compustat GVKEY code. The crosswalk between CRSP and Compustat is not obvious as it first seems. As shown above, a PERMNO can have multiple GVKEYs related to it- in such case, we apply a dynamic match between a PERMNO and Compustat accounting data. However, CRSP also includes cases where under the same GVKEY there are several PERMNO codes. This is mainly due to significant M&As, including reverse acquisition, that occurred during the years when the firm was not listed. For example, in some cases, the merge between CRSP to Compustat results in a firm name related to more than one GVKEY identifier. For those cases, we manually checked using 10K-SEC fillings the years that the name was relevant for each GVKEY. Also, there is a difference in coverage between CRSP and Compustat for the early sample years[13] – we added missing information from Compustat and manually checked for historical names wherever possible.

Our main firm identifier PERMNO_ADJ builds on the original CRSP PERMNO id with several adjustments[14]. (i) In cases where under the same GVKEY, we find several PERMNO codes we replace it with one main PERMNO code[15] – for example, OWENS Corning GVKEY (008214) was split to two PERMNO codes 24811 and 91531 due to it being unlisted between 2003-2005. However, we keep PERMNO_ADJ the same for the complete period (24811). (ii) We manually add a PERMNO_ADJ code for firms in our Compustat sample that did not appear in the "CRSP Monthly Stock" file due to coverage differences.

[12] For example, while SPHERIX INC is related to 2 different GVKEYs (002237 for 1980-2013 and 018738 for 2013-current) it has a unique PERMNO code for the entire period (18148). Similarly, Google Inc PERMNO code is 90319 and it remains the same after the company reorganized as ALPHABET INC in 2015.

[13] There are differences between CRSP and Compustat coverage- for example, CRSP only includes firms listed in USA major exchanges and specifically excludes regional exchanges, while Compustat includes all 10-K filer firms in North America. Moreover, CRSP coverage for major exchanges has expanded gradually over the years (e.g., ARCA was only added from 2006).

[14] It is consistent with NBER2006's PDPCO firm id.

[15] In the final accounting data panel, we split firms based on big jumps in sales, patents or publications. For example, we split PERMNO_ADJ 66093 to the period before and after SBC Communications Inc acquired AT&T Corp and became AT&T Inc. PERMNO_ADJ_LONG is the final UO identifier in the accounting data panel after the split.

We further perform extensive manual checks on the name list, including identifying and distinguishing companies with similar names[16]. Finally, we cleaned and standardized firm names as CRSP tends to abbreviate long words in the company name that it provides. We located those cases and manually corrected them to avoid mismatches.[17]

Standardizing firm names

Prior to matching, we standardize firm names to reconcile company names that may be spelled differently across databases. We compose a standardization code used on both the source and the target names to increase the number of exact matches.

Each company name was first standardized by converting all strings to uppercase characters and cleaning all non-alphabetic characters as well as Compustat related indicators (e.g., -OLD, -NEW, -CL A) and other common words (e.g., THE).

Additionally, an important step in standardizing the company names is standardizing abbreviations. We formed a list that includes over 80 abbreviated words matched to their various original words. For example, LABORATORIES, LABORATORY, LABS, LABO, LABORATORIE, LABORATARI, LABORATARIO, LABORATARIA, LABORATORIET, LABORATORYS, and LABORATORIUM were all abbreviated to "LAB". The list was compiled from the most frequently abbreviated words in WOS affiliation field (accordingly, the list is targeted to our sample). This list is presented in Table 2.

Table 2. Most frequent abbreviated words

ADV	AEROSP	AGR	AMER	ANAL	ANALYT	ANIM	APPL	APPLICAT
ASSOC	AUTOMAT	BIOL	BIOMED	BIOPHARM	BIOSCI	BIOSURG	BIOSYS	BIOTECH
BIOTHERAPEUT	CHEM	CLIN	COMMUN	COMP	CORP	CTR	DEV	DIAGNOST
DYNAM	EDUC	ELECTR	ENGN	ENVIRONM	FAVORS	GEN	GENET	GRAPH
GRP	HLDG	HLTHCR	HOSP	INC	IND	INFO	INNOVAT	INST
INSTR	INTERACT	INTL	INVEST	LAB	LTD	MAT	MED	MFG
MICROELECTR	MICROSYS	MOLEC	NATL	NAVIGAT	NEUROSCI	NUTR	ONCOL	ORTHOPAED
PHARM	PHOTON	PHYS	PROD	RES	SCI	SECUR	SEMICOND	SERV
SFTWR	SOLUT	SURG	SYS	TECH	TEL	TELECOM	THERAPEUT	TRANSPORTAT

For each standardized name, we create a cleaner, fully-standardized name by omitting the legal entity endings and other general words (e.g., INC, CORP, LTD, PLC, LAB, PHARMACEUTICAL), where possible, to maximize match rates (e.g., "XEROX CORP" was standardized to "XEROX", "ABBOTT LABORATORIES" to "ABBOTT"). However, in cases where the company name is too short, generic or can match to other strings within the affiliation field, we preserved the original standardized name to avoid mismatches and extensive manual checks on the match results. For example, omitting the legal entity from "QUANTUM CORP" would result in a potential mismatch between "QUANTUM" and "TEXAS STATE UNIV CTR APPL QUANTUM ELECTR DEPT".

The last step in name standardization is to locate abbreviations that are commonly used by companies instead of their official names. For example, "INTERNATIONAL BUSINESS MACHINES CORP", will also appear under its common abbreviation "IBM" and ''GENERAL ELECTRIC CO" under "GE". We also add the names of prominent R&D laboratories affiliated with companies, such as the T.J. Watson Research Center (IBM) and Bell Labs (initially AT&T and later under Lucent technologies), as authors often omit the name of the company when the address of the laboratory is stated as the publication address.

[16] For instance, RACKABLE SYSTEMS INC (GVKEY 162907) changed its name to SILICON GRAPHICS INTL CORP after it acquired the public company SILICON GRAPHICS INC (GVKEY 012679) in 2009 – we need to make sure that we count SILICON GRAPHICS related publications and patents under RACKABLE's GVKEY only from 2009. Similarly, we need to distinguish between the original BIOGEN INC (GVKEY 002226) and the new BIOGEN INC (GVKEY 024468) that was formed only after the merger with IDEC PHARMACEUTICALS CORP in 2003.

[17] It is also worth mentioning that the "CRSP Monthly Stock" file reports acronym firm names with extra space between the initial letters (e.g., E G & G INC and not EG&G INC). This has to be taken into consideration when performing matches to other databases that do not use this format.

Constructing the name list

All our matching is done at the firm name level. We assign each firm name a unique identifier ID_NAME and indicate the first, and last year the name is relevant for a PERMNO_ADJ. We then perform dynamic matching of names to PERMNO_ADJ based on SDC's M&A data. M&A reassignment includes up to five reassignments per name over the sample period (explained in further details below). PERMNO_ADJs are then dynamically linked to GVKEYs[18]. We further link non-publicly traded subsidiaries to their UO firm. Related subsidiary names are reassigned accordingly up to five times to UO firms. For further details on the ownership methodology, see Section B below.

Our UO and subsidiary historical standardized name lists ("DISCERN_UO_name_list.dta" and "DISCERN_SUB_name_list.dta", respectively), including the dynamic reassignment, will become publicly available for researches to match to their database of interest. Main variables of the name list file are described below:

Variable name	Description
NAME_STD	Historical standardized UO firm names (1980-2015) for firms that were included in our initial Compustat sample[19] and their related subsidiaries.
ID_NAME	Name ID unique at name_std-permno_adj1
$\text{PERMNO\_ADJ}_{0\text{-}5}$	Owner firm id: up to 5 owners + "0" is usually the pre-IPO owner if applicable.
$\text{NAME\_ACQ}_{0\text{-}5}$	Owner name
$\text{FYEAR}_{0\text{-}5}$	First-year assigned to the owner
$\text{NYEAR}_{0\text{-}5}$	Last-year assigned to the owner

I. DYNAMIC REASSIGNMENT

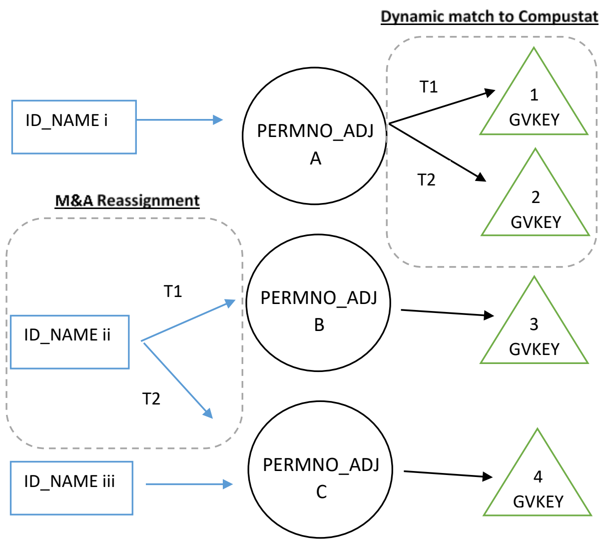
We build on the strategy used by NBER patent match (2006) to perform a dynamic reassignment for our subset of UO Compustat firms (see Figure 1). The dynamic reassignment accounts for: (i) changes in Compustat identification numbers (challenge 1 above) - dynamically matching Compustat accounting information for firms that are related to more than one GVKEY record, and (ii) M&A reassignment based on SDC data and construction of a complete name history for the period 1980-2015 (Challenge 4 above). For M&A reassignment, we include up to five ownership reassignments for each firm name that appears in our initial Compustat subsample and acquired by another firm in our sample. Unless a name is reassigned to another PERMNO_ADJ, it stays with the focal firm until the end of the sample (or the firm's trading period). We dynamically reassign related patents and scientific publications of the acquired UO firm and its related subsidiaries to acquirer firms accordingly (will be discussed in more detail below).

Each PERMNO_ADJ is then linked to Compustat GVKEYs. For cases where there are changes in Compustat identification numbers over the sample period, we dynamically match PERMNO_ADJ to GVKEYs. In the final accounting data panel, we further split firms based on big jumps in sales, patents, or publications. PERMNO_ADJ_LONG is the final UO identifier in the accounting data panel after the split.

[18] For the link between PERMNO_ADJ and GVKEYs see "permno_gvkey.dta" file. In the final panel file, we further split UO firms based on big jumps in sales, patents, or publications and our unique UO firm identifier in the accounting data panel is labeled as PERMNO_ADJ_LONG.

[19] The UO list, "DISCERN_UO_name_list.dta", includes only names of UO parent firms included in our initial Compustat sample. Exceptional are names of top laboratories and names of majority-owned publicly traded subsidiaries that appeared in our initial Compustat sample and were collapsed under the UO parent firm. The subsidiary name list, "DISCERN_SUB_name_list.dta", includes all related subsidiaries as explained in Section B below. The standardization code that was used to standardize the names is available under NAME_STD.do file. Standardized names include legal entity and other common words - in cases where users want to match to a cleaner version of the name, they should apply their own script to clean the names further. When matching the name list to other databases, users should include extensive manual inspection to matched results. Special care should be given to companies with similar names and to generic company names.

Figure 1. Description of dynamic changes

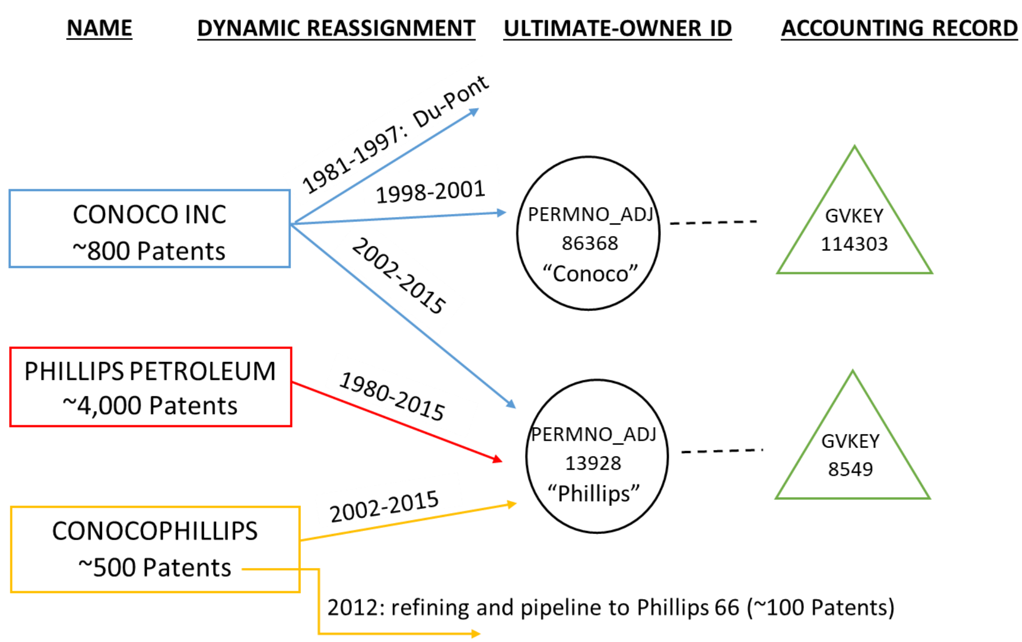


Note: This figure illustrates the dynamic structure of the data. The dynamic reassignment accounts for: (i) changes in Compustat identification numbers (GVKEY), and (ii) M&A reassignment. Each name (ID_NAME) can be assigned throughout the sample period to more than one firm (PERMNO_ADJ) and each firm can be linked to more than one Compustat record (GVKEY).

Example 2: CONOCO and PHILLIPS PETROLEUM

In 1981, Conoco was acquired by Dupont, which has later spun it off as a publicly traded company, which was eventually acquired by the publicly traded company, Phillips Petroleum, in 2002. The merged entity was renamed ConocoPhillips. When we examine current Compustat records, we would only locate the name ConocoPhillips with no record of Philips Petroleum. Compustat does not provide any info on the owner of the record prior to the merger. We use the CRSP monthly stock file to locate all historical names of related securities.

Figure 2. Conoco-Phillips historical names and related patents, 1980-2015

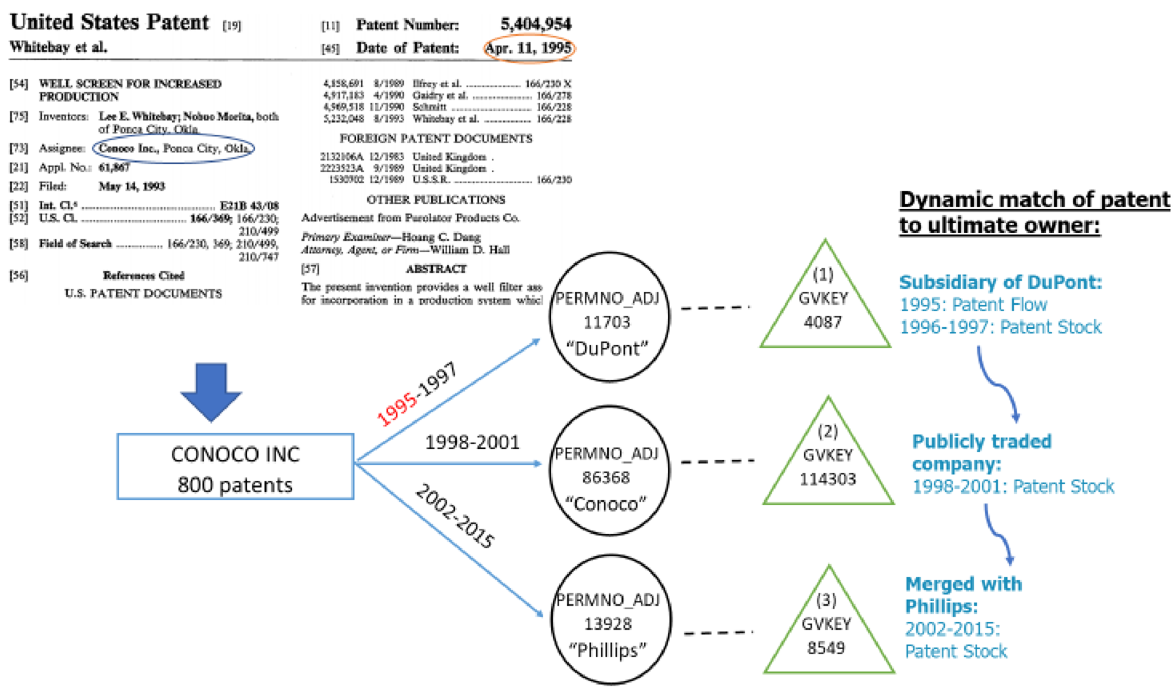


Note: This figure illustrates the historical names and the dynamic structure of the data related to Conoco-Phillips. Each name can be assigned throughout the sample period to more than one firm (PERMNO_ADJ)

Historical names are important for matching patents (and other assets) for the following reasons. (i) They allow us to account for patents assigned to firms in our sample that earlier data missed because the focal firm operated under a different name: under Phillips's name, we locate the majority of granted patents. Four thousand patents that were issued to Phillips Petroleum that were not matched previously without the historical name info. (ii) For ownership changes- we match the merged firm's patents only after the M&A and not before. In this case, Conoco is matched to ConocoPhillips only after the merger in 2002. (iii) Historical names also help match subsidiary data as UO names appear in ORBIS files as of the year the file was recorded (e.g., Chevron-Phillips JV formed in 2000 that we match at the subsidiary level).

In addition to locating historical firm names, we do extensive work on ownership, which enables us to match firm names dynamically to more than one UO-Firm.

Figure 3. Conoco-Phillips dynamic match



Note: This figure illustrates the dynamic match of patents to UO firms. Each patent can be assigned throughout the sample period to more than one firm (PERMNO_ADJ)

Figure 3 illustrates the process of dynamic matching. Patent "5404954" was granted to Conoco Inc in 1995. At that time, Conoco was a subsidiary of Dupont. In our data, this patent would be included in Dupont's patent flow for 1995. It will also be counted under Dupont's patent-stock for 1996-1997. However, from 1998- when Conoco is spun-off as an independent publicly traded company, this patent would be transferred dynamically from Dupont to Conoco's patent stock. Similarly, in 2002 the patent would move on to ConocoPhilips patent stock.

A different patent, which is issued to Phillips Petroleum in 1999, for instance, would be part of the patent flow assigned to Phillips in 1999 and be counted under the patent stock for Phillips Petroleum till 2002, and then would move on to become part of ConocoPhilips patent stock. The dynamic reassignments are based on our dynamic name list, as shown in Figure 4. We put much effort into tracking these ownership changes. We will elaborate on our ownership methodology below.

Figure 4. Example of dynamic name list for Conoco-Phillips:

ID Name	Name std	fyear 0	nyear 0	Permno Adj_0	Name ACQ_0	Fyear 1	Nyear 1	Permno Adj_1	Name ACQ_1	fyear 2	nyear 2	Permno Adj_2	Name ACQ_2
2384	CONOCO INC	1981	1997	11703	DU PONT E I DE NEMOURS & CO	1998	2001	86368	CONOCO INC	2002	2015	13928	PHILLIPS PETR CO
7325	PHILLIPS PETR CO					1980	2002	13928	PHILLIPS PETR CO	2003	2015	13928	CONOCO PHILLIPS
2385	CONOCO PHILLIPS					2002	2015	13928	CONOCO PHILLIPS				
7324	PHILLIPS 66	1980	2011	13928	CONOCO PHILLIPS	2012	2015	13356	PHILLIPS 66				

Example 3: TIME-WARNER and AMERICAN ONLINE

This example illustrates how properly accounting for name and ownership changes improve the accuracy of patent flow as well as the dynamic reassignment of patents.

Warner Communication and its subsidiaries were independent and publicly traded companies until their merger with Time Inc in 1989 when Time-Warner Inc was formed. In the second half of 2000, Time-Warner was merged with American Online to form AOL Time Warner. In 2003 the company dropped the "AOL" from its name and was renamed Time-Warner Inc. AOL remained a subsidiary until it was spun-out in 2009.

The NBER 2006 patent match reveals:

1) Warner Communication and its related subsidiary patents are correctly matched to WARNER COMMUNICATIONS INC (GVKEY 11284) up to the merger with Time Inc. However, they are not dynamically assigned after 1988 to Time Warner or any other company, implying that the patent stock and patent flow of Time-Warner (and later AOL Time-Warner) from patents related to Warner communication and its subsidiaries (e.g., Warner Bros, WEA Manufacturing (before it was acquired) – above 60 patents up to 2006) are below the true value after the acquisition in1989.

2) TIME-WARNER related patents from 1991 to 2000 (before the merger with American-Online Inc in late 2000) are matched incorrectly to GVKEY 25056, which during those years was solely AMERICAN-ONLINE INC original Compustat financial records. The current name of GVKEY 25056, TIME WARNER INC, which is likely to have misled NBER to link the Time Warner patents to it, was only adopted retroactively in 2003 when the “AOL” was dropped from the official name. Moreover, AMERICAN ONLINE INC and AOL related patents (152 patents up to 2006 based on NBER raw patent match) are not linked to any Compustat record. AOL-TIME WARNER related patents, on the other hand, are matched to a “Pro-Form” Compustat record that is active for only two years 1999-2000: AOL TIME WARNER INC-PRO FORM (GVKEY 142022). All of which implies that AOL Time Warner’s flow of patents is below the true level throughout the period.

Having a complete history of names enables us to correctly identify each Compustat record and its origin and dynamically match each firm name in our sample to the correct financial records accordingly: (i) AMER ONLINE INC (and later AOL) is matched from 1980 until its spinout in 2009 to GVKEY 25056 and after to AOL INC (GVKEY 183920). (ii) Warner Communication is matched up to the merger with Time Inc to WARNER COMMUNICATIONS INC (GVKEY 11284) and later dynamically transferred ending up in AOL -Time Warner GVKEY (25056) starting 2001. (iii) AOL -Time Warner is matched to AOL -TIME WARNER (GVKEY 25056) starting 2001 after the merger was approved. (iv) As a side note- Time Inc is not included as an UO in our sample as it did not have R&D expenses, but it is included as a subsidiary name under the Time-Warner UO company.

Example 4: PHARMACIA & UPJOHN and MONSANTO

This example demonstrates that having a complete history of names enables us to correctly identify each Compustat record's historical ownership and dynamically match each firm name in our sample to its relevant financial records in each period. For instance, linking each patent to its correct financial record can be a concern for papers that link patents to market value, specifically those distinguishing different types (e.g., high vs. low cited patents), which rely on the specific patent that was matched and not only the quantity.[20]

In 1995 original Pharmacia merged with Upjohn to form Pharmacia & Upjohn. In 2000, original Monsanto merged with Pharmacia & Upjohn to form Pharmacia Corporation (New Pharmacia). Between 2000-2002 the new Pharmacia gradually spun off its agricultural operations to a newly created subsidiary, Monsanto Company (New Monsanto). In 2003 the new Pharmacia was acquired by Pfizer and is now a wholly-owned subsidiary of Pfizer. Table 3 illustrates how our methodology allows us to compute patent stock and flow for each GVKEY record correctly.

[20] The following are additional examples: (I) Patents of Honeywell before the merger with Allied Signal (3,112 patents) are incorrectly linked to Allied Signal's GVKEY (001300) up to 1999, while the financial records of the original Honeywell Inc are disregarded (GVKEY 5693). (II) Patents of TELEDYNE INC (GVKEY 10405) pre-merger with the publicly traded ALLEGHENY LUDLUM CORP in 1996 (to form ALLEGHENY TELEDYNE INC, which in 1999 was renamed ALLEGHENY TECHNOLOGIES INC after TELEDYNE was spun-off as free-standing public company) are not linked GVKEY 10405 (634 patents up to 1999, of which 597 patents are pre-1996 merger). In addition, ALLEGHENY LUDLUM CORP's (GVKEY 13708) patents (254 patents, of which 240 patents pre-1996 merger) were not dynamically moved to TELEDYNE INC post-merger. This means that in 1996 (post-merger) the patent stock of GVKEY 10405 is missing at least 789 patents (not including related subsidiary patents). (III) For the new Biogen Inc (GVKEY 24468) NBER does not include patents of IDEC pharmaceuticals, who was the owner of the security before Biogen and IDEC merged in 2003 (40 patents).

Table 3. PHARMACIA & UPJOHN and MONSANTO dynamic match

Period	related GVKEY	Relevant Compustat name for period	Most recent Compustat name	Comments	Patent flow per period per our strategy (based on NBER raw patent match, w/o subsidiaries)	Original NBER match
1950-1994	11040	UPJOHN CO	PHARMACIA & UPJOHN INC	Original Upjohn before merger with Pharmacia	**2,091** Upjohn related patents	N/A
1995-1999	11040	PHARMACIA & UPJOHN INC	PHARMACIA & UPJOHN INC	**1995**: Upjohn merged with original Pharmacia to form Pharmacia & Upjohn	**479** Pharmacia &/ Upjohn related patents	N/A
1950-1999	7536	MONSANTO CO	PHARMACIA CORP	Original Monsanto before merger with Pharmacia & Upjohn	**3,228** Monsanto related patents	**2,733** Pharmacia &/ Upjohn related patents (including patents of Pharmacia before it merged with Upjohn). **While Monsanto's 3,228 patents are not linked.**
2000-2002	7536	PHARMACIA CORP ("new Pharmacia")	PHARMACIA CORP	**2000:** original Monsanto merged with Pharmacia & Upjohn to form Pharmacia Corporation (New Pharmacia). All of PHARMACIA, UPJOHN and PHARMACIA & UPJOHN patents are transferred here from 2000. Monsanto's patents are redirected to the new Monsanto spin-off company.	**304** Pharmacia &/ Upjohn related patents	**304** Pharmacia &/ Upjohn related patents
2000-2015	140760	MONSANTO CO ("new Monsanto")	MONSANTO CO	**2000-2002:** Pharmacia Corporation (New Pharmacia)gradually spun-off its agricultural operations to a new publicly traded company, Monsanto Co (New Monsanto). All Monsanto related patents are transferred here from 2000.	**553** Monsanto related patents (2000-2006)	**553** Monsanto related patents (2000-2006). *NBER links Monsanto's patents to GVKEY 140760 from 1997 - while records for 1997-1999 are available on Compustat, they are based on prospective filings when Monsanto was still traded under GVKEY 140760.
2003-2015	8530	PFIZER INC	PFIZER INC	**2003:** Pharmacia Corporation (New Pharmacia) was acquired by Pfizer and is now a wholly owned subsidiary of Pfizer. All of PHARMACIA, UPJOHN and PHARMACIA & UPJOHN patents are transferred here from 2003.	**472** Pharmacia &/ Upjohn related patents(up to 2006)	**472** Pharmacia &/ Upjohn related patents(up to 2006)

II. AGGREGATING DATA TO THE UO FIRM LEVEL

To merge parent Compustat companies and their independent majority-owned publicly traded Compustat subsidiaries (Challenge 3 above), we locate related firms in our initial Compustat subsample based on name similarity as well as by matching the firm names to ORBIS subsidiary data. Where needed, we perform manual checks to confirm majority ownership using SEC 10-K filings. We aggregate the data to the UO parent-company level, accordingly.[21] We further link private subsidiaries to their UO firm based on ORBIS data (will be explained separately below). Accordingly, if a firm's subsidiary publishes scientific articles while the parent company is the assignee registered on the firm's patents, we record both at the UO level and a citation from a patent to a publication would be considered as an internal citation.

B. OWNERSHIP STRUCTURE

Dealing with ownership changes has been a major effort of this project, especially in regard to reconstructing and improving the NBER patent database. We unpack firms' ownership structure by constructing firm-level data before proceeding to patent match. Ownership may change over the years of our sample due to changes at the UO Compustat firm level as well as at the subsidiary level. We rely on two main sources to construct ownership data: (i) SDC Platinum and (ii) historical snapshots of ORBIS files.

I. SDC M&A MATCH

Ownership changes of the UO Compustat firms in our sample are tracked through the SDC Platinum database with each firm name dynamically matched to up to five PERMNO_ADJ between the years 1980 and 2015. Based on M&A deals available in SDC Platinum from 1980 to 2015, we downloaded detailed information on the acquirer and target firm names, acquirer and target firm CUSIPs, types of deals, execution dates, and percentage of shares owned after each transaction. We exclude deals that we identify as asset or business unit acquisitions.

We restrict the sample to deals involving a change in ownership that resulted in majority ownership (more than 50% of shares) for the acquirer. Execution dates are used to define the years a target firm begins or ends (in case of several acquisitions during the sample period) being owned by an acquirer. We then standardized both target and acquirer names similar to the standardization done for Compustat firm names. We match each deal's target and acquirer firm to our list of Compustat firms using both CUSIP numbers and all standardized historical names. It is important to use historical data as the information is recorded on SDC at the time of acquisition. We retain deals where both acquirer and target firms are matched to a Compustat firm in our sample. We track up to five ownership changes for each target firm name after it enters Compustat and one additional reassignment before it became publicly traded if relevant (i.e., if it was a subsidiary of another Compustat firm in our sample prior to its IPO)[22].

We perform extensive manual checks, including identifying and distinguishing companies with similar names (e.g., old vs. new Pharmacia). We Assume that if a firm is acquired, all its patents and publications are transferred to the acquirer firm.

[21] For example, GENZYME CORP (GVKEY 12233) - after verifying ownership on SEC filings: GENZYME MOLECULAR ONCOLOGY (GVKEY 117298), GENZYME TISSUE REPAIR (GVKEY 118653), GENZYME SURGICAL PRODUCTS (GVKEY 121742) and GENZYME BIOSURGERY (GVKEY 143176) are all accounted under their parent company GENZYME CORP (GVKEY 12233). While, GENZYME TRANSGENICS CORP (a.k.a. GTC BIOTHERAPEUTICS, GVKEY 028563) is a standalone alone company in our data as it was not majority-owned by GENZYME CORP after it spun-off.

[22] For example, Vysis Inc first enters our sample as a subsidiary of Amoco (1991-1997) and is then spun-off and becomes an UO firm in our sample as an independent publicly traded company in 1998 and eventually acquired and becomes a subsidiary of Abbott in 2001.

Example 5: NABISCO

This example illustrates how we account for ownership changes in our data. During our sample period, Nabisco has changed ownership four times. In 1981 Nabisco merged with the publicly traded company Standard Brands to form Nabisco Brands. Then, in 1985 R.J. Reynolds merged with Nabisco Brands to create RJR Nabisco, which eventually became Nabisco Group holding after the tobacco business was spun out in 1999. In 2000, Nabisco was acquired by Phillip Morris, which combined Nabisco with its Kraft brand. Finally, in 2001 Kraft (together with Nabisco) was spun out as a publicly traded company that later on became Mondelez International Inc. In our dataset all Nabisco related patents and publications are dynamically transferred between Compustat records and UO firms based on its ownership throughout the years:

Table 4. Nabisco dynamic match

Years	related GVKEY	Original owner	Current Compustat name	Comments
1981-1985	7674	STANDARD BRANDS INC	NABISCO BRANDS INC	1981: Standard Brands company merged with Nabisco Inc to form Nabisco Brands Inc.
1986-1999	9113	R J REYNOLDS IND INC	NABISCO GROUP HOLDINGS CORP	1985: R.J. Reynolds Industries merged with Nabisco Brands to form R J R Nabisco Inc
2000	8543	PHILIP MORRIS COS INC	ALTRIA GROUP INC	2000: Nabisco was acquired by Phillip Morris
2001-2015	142953	KRAFT FOODS INC	MONDELEZ INTERNATIONAL INC	2001: Kraft together with Nabisco split from Phillip Morris

Examining NBER 2006, we find that for the purpose of Compustat accounting information, all Nabisco related patents are linked to GVKEY 9113 from 1950 to 1999. Though the current name related to GVKEY 9113 is "Nabisco Group Holding Corp", based on the historical name information, we know that up to the merger of R.J. Reynolds with Nabisco it belonged solely to R.J. Reynolds. Reynold's patents, on the other hand (Over 419 patents for the period before it spun-out of RJR Nabisco and not including patents of acquired companies such as Heublein Inc), are not assigned by NBER to GVKEY 9113 and they are only being linked to Compustat records after the tobacco business spun-out of RJR Nabisco and became independently traded again under GVKEY 120877 (eventually merging with U.S. operations of British American Tobacco to form Reynolds American Inc). As a result, in 1998, the patent stock in NBER for GVKEY 9113 ("Nabisco Group Holding Corp") is 495 (consisting solely of Nabisco matched patents), whereas it should be 914 if it included R.J. Reynolds related patents. Furthermore, NBER does not dynamically move Nabisco's patent-stock or account for its patent flow after 1999 when it was bought by Philip Morris and eventually became part of Kraft (a total of 529 Nabisco related patents up to 2006).

Table 5. Data for Nabisco in NBER 2006

Current compustat record name	gvkey	firstyr	lastyr	pdpco	pdpseq	begyr	endyr
NABISCO GROUP HOLDINGS CORP	**9113**	**1950**	**1999**	**9113**	**1**	**1950**	**1999**
NABISCO INC	7675	1950	1980	9113	-1		
NABISCO BRANDS INC	7674	1950	1984	9113	-1		
NABISCO HLDGS CORP -CL A	31427	1993	1999	9113	-1		

Example 6: CHEMTURA CORPORATION

An example that illustrates how having historical names helps account for ownership changes in our data and accurately compute the patent stock. Chemtura Corporation traces back to the chemical corporation Crompton & Knowles that was founded in the 19th century. In 1996, Uniroyal Chemical Corporation merged with Crompton & Knowles. In 1999, Crompton & Knowles merged with the publicly traded company Witco to form Crompton Corporation. In 2005, Crompton acquired the publicly traded company Great Lakes Chemical Company, Inc., to form Chemtura Corporation, while Great Lakes Chemical Corporation continued to exist as a subsidiary company of Chemtura.

Based on our strategy, we consider all historical names of the current Chemtura Corporation (PERMNO_ADJ 38420) including:

1) CROMPTON & KNOWLES CORP starting 1980
2) CK WITCO CORP starting 1999
3) CROMPTON CORP starting 2000
4) CHEMTURA CORP starting 2005

Most importantly, because we consider the complete set of historical names, we are able to locate all the relevant M&As throughout the years of the publicly traded firms that exist as an independently traded company in our data prior to an acquisition. Accordingly, we dynamically transfer them post-acquisition to PERMNO_ADJ 38420:

1) Uniroyal Chemical Corporation (acquired 1996)
2) Witco Corp (acquired 1999)
3) Great Lakes Chemical (acquired 2005)

When we examine NBER 2006 patent dataset, we find that the only name that was matched to CHEMTURA CORP (GVKEY 3607) is "CHEMTURA CORP" (PDPASS 13245038). As the Chemtura name was adopted in 2005, only one patent was matched for that name. In addition, none of the acquired publicly traded companies were dynamically transferred to CHEMTURA CORP post-acquisition. It is likely that a lack of information on historical names led NBER to rely on post-acquisition name (Chemtura) and thus prevented it from accounting for the M&A activities.

By considering all previous names (without their subsidiaries and the acquired companies) related to GVKEY 3607: (i) Crompton & Knowles Corp; (ii) CK Witco Corp and (iii) Crompton Corp - based on the NBER raw patent match, we locate 220 additional patents up to 2006 that were not linked to any Compustat record that should be assigned to *Crompton & Knowles* (77 patents), *CK Witco* (26 patents)), and *Crompton* (117 patents). In addition, the acquired *Uniroyal Chemical Corp* has a patent stock of 379 patents in 2006 (out of which 185 patents are post-acquisition), and the acquired Witco company has a patent stock of 405 in 2006 (out of which 62 patents are from post-acquisition period), and *Great Lake Chemicals* has a patent stock of 183 in 2006 (out of which three patents are in 2006, the year after the company was acquired).

Overall, applying our strategy to the raw NBER patent match, we find a patent stock of **1,187** patents in 2006 for GVKEY 3607 as opposed to **1** patent in NBER.

II. ORBIS SUBSIDIARY MATCH

Due to the complexity of measuring large firms' innovative activities, which typically take place inside numerous subsidiaries, we aggregate the data to the ultimate-owner-parent-company level based on majority ownership. There are several challenges in keeping track of subsidiaries owned by UO Compustat firms, which may publish and patent in their own name. First, many of these subsidiaries are private, and manual checks are sometimes required to verify which of the several similarly named companies was acquired by the firm. Furthermore, subsidiary ownership may change over the years. Companies may spin out their subsidiaries, some of which might go public or sold to other firms, where they are maintained as stand-alone subsidiaries and continue to patent or publish. Tracking subsidiary ownership is the main challenge we deal with and is explained below.

For firms with at least 50 patents[23] over the sample years at the PERMNO_ADJ UO level, we collect all related domestic and international subsidiary names using ORBIS and SEC filings, as explained below.

We obtained historical ORBIS files for years 2002 to 2015, which provide us with snapshots of ownership structures for each of the years. Using historical snapshots of ORBIS files, we are able not only to identify ownership changes at the subsidiary level but also new established subsidiaries.[24]

[23] At the UO level we match for all subsample firms related subsidiaries with organic names.
[24] One caveat is that the coverage of subsidiaries in the first few years of data files is incomplete.

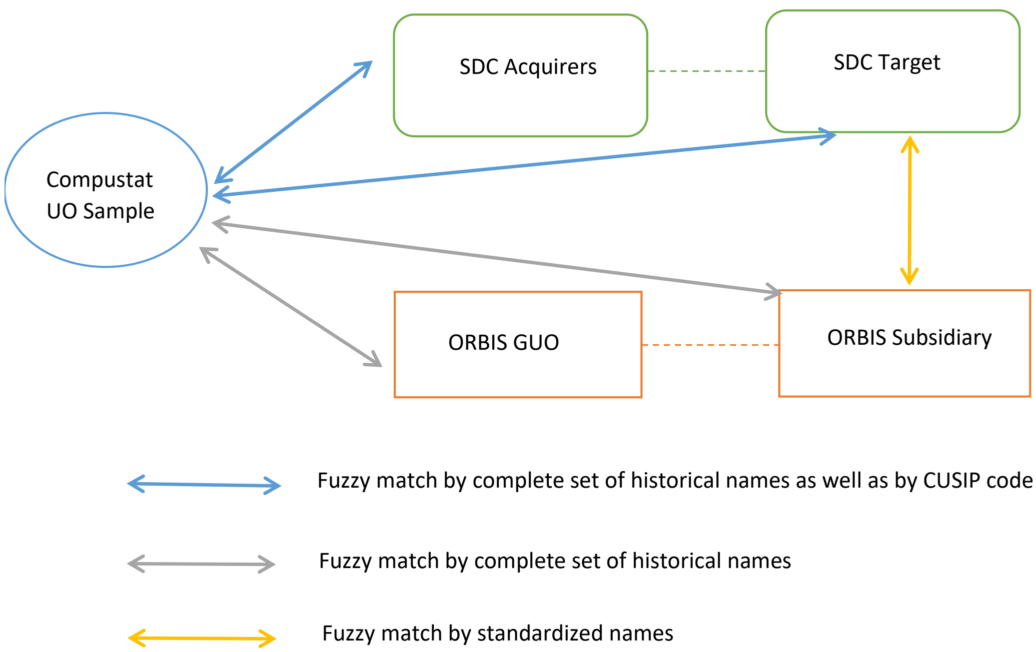
We start by standardizing the names of "Global Ultimate Owner" (GUO) firms and match the names to standardized historical Compustat names of firms with more than 50 patents at the UO level. Once again, it is important to use historical names for this match as the names in each of our ORBIS files appear as of the year the file was recorded.

Next, we link the subsidiaries of the successfully matched ORBIS owners to the PERMENO_ADJ of the corresponding parent firms. We restrict our sample to subsidiaries that are majority-owned by the parent firm. After standardizing each subsidiary name similar to the standardization done for Compustat names, we obtain the first and last year it appears under a PERMENO_ADJ during 2002-2015. To avoid duplicated matching efforts, in many cases, we drop subsidiaries that have the same organic name as the parent UO firm as they were already matched at the UO Compustat level. Some subsidiary names appear under more than one PERMNO_ADJ due to acquisitions throughout the years. Because we use yearly snapshots of ownership structure from ORBIS, we are able to account for name changes of subsidiaries over the period.

For firms that exit Compustat before 2002, we manually collect subsidiary names based on their latest available 10-K SEC filing[25] as well as rely on the NBER patent database for pre-2002 ownership data.

Since our sample starts from 1980 and the ORBIS files are only from 2002, we try our best to account for ownership changes of the subsidiaries for the years preceding 2002 using SDC and Compustat databases. We elaborate on our approach below.

Figure 5. Subsidiary matching Description



1) Fuzzy match between standardized subsidiary names and standardized SDC target name. For the matched result, we locate:

a) Cases where the acquirer firm is a UO Compustat firm in our sample, which include:

(i) Cases where the acquirer firm has the same PERMNO_ADJ as the parent firm of the subsidiaries. These cases confirm the direct acquisition of the subsidiary by the parent firm and provide us with the start date of the subsidiary (the year of acquisition) under the parent firm.

[25] We do so for top 100 firms based on R&D spending.

(ii) Cases where the acquirer firm is a UO Compustat firm in our sample that was acquired by the parent firm of the subsidiary (i.e., the PERMNO_ADJ of the acquirer and the PERMNO_ADJ of the parent of the subsidiary are related through acquisition). These cases confirm an indirect acquisition of the subsidiary by the parent firm and provide us with the start year of the subsidiary under the parent firm – i.e., year the ORBIS parent firm acquired the Compustat acquirer firm or the year of acquisition of the subsidiary (the latest).

b) cases where the acquirer firm is not a UO Compustat firm in our sample:

(i) if the CUSIP code of the UO parent firm related to the target firm (as indicated in SDC file) is the same as a CUSIP code related to the PERMNO_ADJ of the ORBIS parent of the subsidiary, it indicates that the subsidiary was acquired from the parent firm by the acquirer and provides us with the end date for the subsidiary under the parent firm – the year of acquisition.

(ii) For each acquirer firm's direct CUSIP code, we search the complete SDC file for a deal where it was acquired by a firm with a CUSIP code related to the PERMNO_ADJ of the ORBIS parent of the subsidiary. These cases indicate indirect acquisitions, in which the subsidiary was acquired by a non-Compustat sample firm that was itself acquired by the subsidiary's ORBIS parent firm. Such cases provide us the start year of the subsidiary under the parent firm –i.e., the year the ORBIS parent firm acquired the non-Compustat acquirer firm or the date of acquisition of the subsidiary (the latest).

2) Fuzzy Match of cleaned subsidiary names

As the subsidiary name list includes closely related firm names with different legal entity, we use a clean version of the names that omits legal entity and other common words and we fuzzy match it to both clean Compustat names and the list of clean subsidiary names we found relevant acquisitions for in (1) above.

The fuzzy match to Compustat enables us to link each matched subsidiary name to the dynamic year sequence we constructed for UO Compustat firms. For the fuzzy match to the list of acquired subsidiaries, we adopt the relevant start &/end year we located in (1) above to all related subsidiaries.

3) As an additional check, we manually go over subsidiaries that did not match under 1) or 2) above and appear under more than one parent firm in our ORBIS sample or have more than 100 matched publications or patents.[26] For these cases, we check online sources and manually adjust their start and end date. Finally, for subsidiaries, we were not able to identify the start or end year- we assume that they belong to the UO firm from its start date until the end date. However, if the UO firm appeared in ORBIS files for more than three years before the subsidiary was first linked to it, we adopt the first year the subsidiary is connected to the parent ORBIS firm as the start date of the subsidiary, under the assumption that it was acquired during that year by the parent firm.

All subsidiaries are assumed to move with their parent firm in cases where the parent firm is acquired unless a subsidiary has a different end date from its parent firm, or it is related to the Compustat dynamic year sequence. Moreover, we do not account for reassignment of patents that are not part of the ownership changes that we document.

[26] When matching the subsidiary name list to other databases users should include extensive manual inspection to matched results, including manually verifying the start and end year for top matched result that differ from the top 100 matches that we manually verified.

C. MATCHING

We perform several matches to construct our data, including (1) matching patent data to Compustat companies and their related subsidiaries; (2) matching scientific publications to Compustat companies and their related subsidiaries; (3) mapping patent citations to publications. We discuss each of these procedures below.

I. MATCHING PATENT DATA TO COMPUSTAT COMPANIES AND THEIR SUBSIDIARIES

After obtaining our initial subsample of firms and the various firm names, we proceed to match our firm sample to assignees of the patents granted by USPTO[27] using PatStat, which includes approximately 5.3 million patents for years 1980 through 2015.

We first remove published patent applications (i.e., publication numbers longer than 7 characters), non-utility patents, including Design, Reissue, Plant and T documents, and reexamination certificates. Next, we remove patents assigned to individuals or government entities (for example, an assignee that includes the string "DECEASED" or "U.S. DEPARTMENT"). We are then left with 4.97 million granted utility patents.

To compare assignee names to the standardized firm names in our sample, we standardize assignee names similar to the firm name standardization explained above. Assignee name standardization includes converting names to upper case, removing excess spaces, cleaning non-alphanumeric characters, and replacing legal entity endings, including commonly abbreviated terms (for example, "CORPORATION" is replaced with "CORP"; "LABORATORIES" and "LABS" with "LAB"). At the end of this process, we are left with 897K unique standardized assignee names.

The matching strategy includes several distinct steps. We begin by matching firm names to assignees using an exact match. We then perform several fuzzy matching techniques to account for names that are slightly different but are in fact, the same entities. Extensive manual checks at the assignee name and patent level were performed to ensure the quality of the matches.

UO Level Matching

(1) Exact Matching

Exact matching was conducted by comparing assignee names to firm names. The matching was carried out twice, both for standardized and for original names. An additional match was conducted after dropping legal entities. The latter step was performed to account for firms whose names differ only by the legal entity. Extensive manual checks are performed to verify the matches. Special care was taken in cases where firm or assignee names are generic, when several different firms share a common portion of a name, or when firm names contain a common given or family name. To resolve ambiguities, we performed web searches and examined the actual patent documents.

(2) Fuzzy Matching

For the remaining assignee names not matched during the exact matching process, fuzzy matching was performed to find each of the assignee names from the firm names to catch cases where assignee and firm names do not match exactly but are, in fact, the same firm. Some names are misspelled or include additional letters that prevent an exact match. In other cases, patent assignee names include a specific division title ("ROCKWELL BODY AND CHASSIS SYSTEMS", "ROCKWELL SOFTWARE"), a licensing unit ("MICROSOFT TECHNOLOGY LICENSING LTD", "RCA LICENSING"), or a geographic branch or firm location ("BIOSENSE WEBSTER ISRAEL LTD").

Fuzzy matching was performed using the FuzzyWuzzy library in Python (i.e., Token Set function), and using term frequency-inverse document frequency (TF-IDF).

FuzzyWuzzy uses a slightly modified Levenshtein distance to calculate similarities between two strings. More specifically, a vector is created for each assignee name using the words contained in it and then compared to the entire list of firm names (that are also vectorized) to find potential matches. When comparing two vectors, the same

[27] We limited our data sources to USPTO data to make the project manageable in terms of matching. Since our firm sample is limited to U.S. HQ firms, we believe it is reasonable to focus only on USPTO data.

elements (i.e., words) contained in both vectors are marked as "matched", and the similarity between the remaining, different elements are calculated using the Levenshtein distance algorithm after sorting the elements alphabetically. The similarity score between the two strings is higher when the elements that match exactly make up a larger portion of the strings and when the remaining (unmatched) part has a small distance based on the Levenshtein distance. To account for multiple scores that indicate a strong match, the top ten potential matches with the highest scores are examined manually to identify the most appropriate match.

An additional fuzzy match was done by converting the assignee and firm names into a term frequency-inverse document frequency (TF-IDF) matrix and calculating a cosine similarity score for each pair of assignee and firm name. This method is widely used to take care of typos and variations of spelling in textual string matching. By increasing weights of unique words and reducing the weights of common words in the corpus, the TF-IDF algorithm improves the relevancy of cosine similarity measures that are calculated between each pair of names.

An additional search of the top 300 patenting firm names was conducted to find matching assignee names that were not matched through the initial fuzzy match process. In this step, we search for assignee names with at least five related patents that contain any of the fully standardized firm names after the removal of legal entities. Through this process, we include subsidiaries that have the same organic name as the parent UO firm (For example, "EMERSON" firm name matched with "EMERSON CLIMATE TECH", a division within the firm). The search was conducted through a script that receives the list of assignee names and fully standardized firm names and automatically produces all matching pairs. In each search result pair, a firm name is contained within the assignee name string. Following the search, a complete manual check was conducted among all search results to mark the legitimate matches.

As a final check, we employed RAs to verify that the assignees with more than 100 patents were correctly matched by the fuzzy matching algorithm. The RAs went through the fuzzy matched names to confirm that they are in fact, the right match. Existing matches were invalidated when they were not the right match, and new matches were added when more appropriate matches were found.

Subsidiary Level Matching

(1) Exact Matching

Exact matching was conducted in a similar fashion to the UO level matching process. Original and standardized versions of the assignee names were compared to the list of standardized subsidiary names, and manual checks were performed in cases where the name was generic.

(2) Fuzzy Matching

The fuzzy match for subsidiaries was done by converting the assignee and firm names into a term frequency-inverse document frequency (TF-IDF) matrix and calculating a cosine similarity score for each pair of assignee and standardized firm name. To reduce the size of the task, results were limited to assignees with at least 30 patents, and identification of matches was conducted by manually comparing the top-scoring assignee-firm pairs for each assignee.

Overall, this process yields 1.3 million patents mapped to 4,420 U.S. headquartered Compustat firms and their subsidiaries via patent number and NAME_ID. These patents account for about 50% of all utility patent grants from U.S. Origin. When a patent has several assignees, we match the patent to multiple firms and assign fractional patent ownership to each assignee (i.e., 1/number of assignees). Patents enter our sample once the related UO firm is publicly traded and not before. Any patent that enters the data remains until the end of the sample period unless the related firm it is acquired by an out of sample firm, dissolved, or taken private. In case of ownership change within the sample, patents are dynamically matched to up to five UO firms. Moreover, we do not account for reassignment of patents that are not part of the ownership changes that we document.[28]

[28] Specific details on construction of patent flow and patent stock variables are provided under "patent_do.do" file. The main patent output file is "DISCERN_patent_database_1980_2015_final1.dta"

II. MATCHING SCIENTIFIC PUBLICATIONS TO COMPUSTAT FIRMS AND THEIR SUBSIDIARIES

We procced by matching our firm names to publication data to capture their investment in science. We obtain publications data from the Web of Science database (previously known as ISI Web of Knowledge). We include articles from journals covered in the "Science Citation Index" and "Conference Proceedings Citation Index - Science," while excluding social sciences, arts, and humanities articles.

Each publication record contains detailed information including the title of the publication, authors, journal, and our primary variable of interest, an affiliation field with name and address of the publishing institute or company in case of a corporate publication. This field can include more than one listing in case of a collaborative publication, for example, "*TEXAS INSTRUMENTS INC*, DEPT DATAPATH VLSI PROD SEMICOND GRP 8330 LBJ FREEWAY, POB 655303, DALLAS, TX 75265 USA | *SUN MICROSYST INC*, MT VIEW, CA USA".

We apply a many-to-many fuzzy matching algorithm between each standardized name and the affiliation field for each publication (approximately 47 million publications, 8 million conference proceedings and 60 thousand names) while allowing for more than one firm to be matched to each publication (to allow for collaborative publications).

We first standardize the affiliation string of each Web of Science publication similar to the name standardization process explained above. The standardization removes special characters such as ampersands and words that indicate legal entities such as "INC" or "CORP". It also ensures that common words such as "technology" and "chemicals" that frequently appear in company names are abbreviated in the same manner[29].

Second, we perform exact matching on company names and publication affiliation string using regular expressions. In addition, we calculate Levenshtein edit distances between company name-publication affiliation pairs. This step is necessary because misspellings are common (e.g., BRISTOL MYERS SQUIBB misspelled as BRISTOL MEYERS SQUIBB). Since the company name in a publication affiliation is typically embedded in a longer string, which includes buildings, street names, cities, zip-codes, and country names, even correct matches will incur large distances. Therefore, we use a "partial" Levenshtein distance, which calculates the edit distance between the shortest common segment between two strings. That is our "partial" edit distance for the company name "IBM" and affiliation "IBM Corp, SSD, San Jose, CA 951953 USA" will be zero, whereas a raw Levenshtein distance would be 35.

Third, we conduct manual checks on fuzzy-matched company name-publication affiliation pairs. In particular, we exclude matches from company names to eponymous buildings (e.g., Gillette Hall), schools (e.g., Heinz College), hospitals (e.g., Du Pont Children's Hospital), charitable foundations, and endowed chairs. We also conduct manual checks on company-publication pairs with zero edit distances (exact matches) if the company names overlap with a common last name (e.g., ABBOTT), a geographic/historical location (e.g., BABYLON, BRISTOL), or branch of science & engineering such as "APPLIED MATERIALS" or "SEMICONDUCTOR", as these are especially prone to being false positive matches. We also ensure that similar but distinct company names do not match to the same affiliation field (e.g., NORTHROP and GRUMMAN before their merger in 1994 are treated as separate companies and will not match to NORTHROP GRUMANN). In cases where company names are the same, we verify matches by comparing the address listed within Compustat to the address in the publication data. For example, to distinguish between "THERATECH INC / UTAH" and "THERATECH INC", we verify that the address of the firm under the affiliation field is in Salt Lake City.

At the end of this procedure, we obtain a match between a WOS record ID and our NAME_ID. We find approximately 800 thousand unique articles from more than 10 thousand different journals that were published from 1980 through 2015, with at least one author employed by our sample of Compustat firms and their subsidiaries. For the sample of patenting firms, publications enter our sample once the related UO firm is publicly traded and not before. Any publication that enters the data remains until the end of the sample period unless the related firm it is acquired by an out of sample firm, dissolved, or taken private. In case of ownership change within the sample, publications are dynamically matched to up to five UO firms.[30]

[29] For instance, the word "technology" in a company name can be plural ("technologies") or abbreviated ("technol", "tech"). These special cases are abbreviated to "TECH" in our standardization code.

[30] Specific details on construction of publication flow and publication stock variables are provided under "pub_do.do" file. The main publication output file is "DISCERN_pub_database_1980_2015_final1.dta".

III. MATCHING NPL PATENT CITATIONS TO WEB OF SCIENCE ARTICLES

Patent citations to science are obtained from the Non-Patent Literature (NPL) citations section located at the front page of patents taken from the PatStat database. An example of a front-page patent citation to non-patent literature is provided in Figure 7. We obtain all NPLs related to patents granted in the period 1980-2015 (including corporate sample firm patents and non-corporate patents). We first remove NPL citations that we identify as non-publication references (e.g., reference that includes the string "PATENT ABSTRACT", "U.S. APPLICATION NO.", "US COURT", "PRODUCT INFORMATION", "DATA SHEET", "WHITEPAPER"). We then proceed to match NPLs to corporate publications from Web of Science (approximately 10M citations and 800K corporate publications). This step presents a significant challenge due to differences in structure between NPL and publication string text-NPL patent citations to publications are highly non-standardized (see Table 7 for examples). We begin with a many-to-many match, allowing more than one publication to be matched to each NPL. For each possible records pair, we construct a score that captures the degree of textual overlap between the title, journal, authors, and publication year. To exclude mismatches, we use a more detailed matching algorithm that is based on different sources of publication information: standardized authors' names, number of authors, article title, journal name, and year of publication. The matching algorithm accounts for misspelling, unstructured text, incomplete references, and other issues that may cause mismatches.

We will use the example below to illustrate the complication of the match and the algorithm we applied to detect a match.[31]

The first step is to match the publication's "Title" field and the title that is located within the citation string. There are two main problems: (i) the position of the title within the citation is not fixed and (ii) there may be a small variation in the title (e.g., "GIVE" vs. "GIVES") and thus an exact match may not perform well. To overcome these problems, we implement a fuzzy matching algorithm. After we standardize and clean the different strings, we measure the length-difference between the citation string and the publication title string. Then, using STATA's "STRDIST" command, we calculated the distance between the two strings. We use the difference between the length difference and distance as a measure of proximity of the titles. We supplement this measure with an exact match of the first part of the title. In some cases, the title is missing from the citation string. In such cases, we rely more on other available features to determine the final match.

Second, we match between the publication's "Authors" field and the authors listed within the citation string. As with the title, we cannot identify the exact location where the authors are contained within the citation string since the location varies from one citation to another. In addition, there are several differences in how names are written: (i) Last name only vs. full names; (ii) name vs. initials (e.g., LIN KS vs. LIN KUN SHAN); (iii) listing of all authors vs. one author followed (or not) by "et al."; (iv) order of last and first names within the string. To verify a match by authors, we first count the number of authors listed in the publication record. We then check whether the citation string contains "et al.". To mitigate the name variation problem, we implement an algorithm that matches different variations of the authors' name to the citation (including the transformation of last and/or first and/or middle name to initials and changes in the order listed). In cases where several authors are listed under the publication and "et al." does not appear within the citation, we perform a one-to-many match between the citation and each author and impose that at least 80% of the authors must be matched to the citation to determine a match. For cases where several authors are listed in the publication and only one is matched within the citation while "et al." is omitted, we rely more on match results in other features to determine the final match.

Next, we match journal information including standardized journal's name, publication year, page numbers and volume, while accounting for typos, abbreviations (e.g., "INTERNATIONAL ELECTRONICS" vs. "ELECTRONICS") and differences in format of the string between the datasets (e.g., "VOL. 53, NO. 3" vs. "53(3)").

[31] The following example (first line in Table 7) illustrates the matching challenge. NPL citation: LIN, KUN SHAN, ET AL., SOFTWARE RULES GIVES PERSONAL COMPUTER REAL WORD POWER, INTERNATIONAL ELECTRONICS, VOL. 53, NO. 3, FEB. 10, 1981, PP. 122 125.
Matched Publication: Title: SOFTWARE RULES GIVE PERSONAL-COMPUTER REAL WORD POWER, Authors: LIN KS, FRANTZ GA, GOUDIE K, Journal information: ELECTRONICS 54 (3): 122-125 1981.

Finally, we use different combinations of the match results for the various features (title, authors, and journal information) according to their relative importance to determine a final match[32]. We perform extensive manual checks to confirm matches[33]. At the end of this procedure, we obtain unique identification numbers for the citation, the citing patent, and the cited publication.

We then focus on citations made by corporate sample patents. We further differentiate between internal citations (patent citation by the focal firm's patent to its own publication) and external corporate citations (patent citation to the focal firm's publication by other corporate patents). The Dynamic match of patents and publications allows us to classify an internal or external citation based on the owners of the citing patent and the cited paper at the time the paper is published. For the purpose of classifying internal or external citation, we rely on the original UO firm the publication was affiliated with at its publication year[34]. For external citations from the corporate sample firms, we further construct a segment proximity measures between the cited and the citing firms as explained in the main text.

Following the above procedures, we obtain 71 thousand unique corporate cited publications (9 percent of corporate publications), by 142 thousand unique corporate citing patents. Of the cited publications, 61 percent receive only external corporate citations, and the remaining receive at least one internal citation[35]. The temporal structure of citations and publications are illustrated in Figure 8.

[32] A sample algorithm is provided under "NPL_cleaning_exp.do" file

[33] There are several cases where the NPL reference is a citation to a working paper and we are able to match it to the final published paper that appears on WOS database – we consider those as matches.

[34] i.e., if Company B acquires Company A (let's assume A is a Compustat firm in our sample pre-acquisition): Citations by B's patents post-acquisition to A's publications that were published pre-acquisition are classified as external citations. However, citation from B's patents to A's publications published post-acquisition are classified as internal citations. Moreover, as opposed to publication and patent stock variables, citations do not move dynamically between firms in case of acquisition.

[35] Specific details on construction of NPL citation variables are provided under "npl_do.do" file. The main NPL output file is "DISCERN_corp_NPL_output_80_15_final.dta".

D. COMPARISON OF OUR DATA TO NBER PATENT DATA, FOR 1980-2006

We match 780 thousand patents for 1980-2006 (Figure 6). We compare our sample for 1980-2006 to NBER 2006 patent data for U.S. headquarter firms and their related subsidiaries looking at a specific patent assigned to a GVKEY at a grant year (Table 6).

Figure 6. Patents assigned to U.S. HQ public corporations and their related subsidiaries

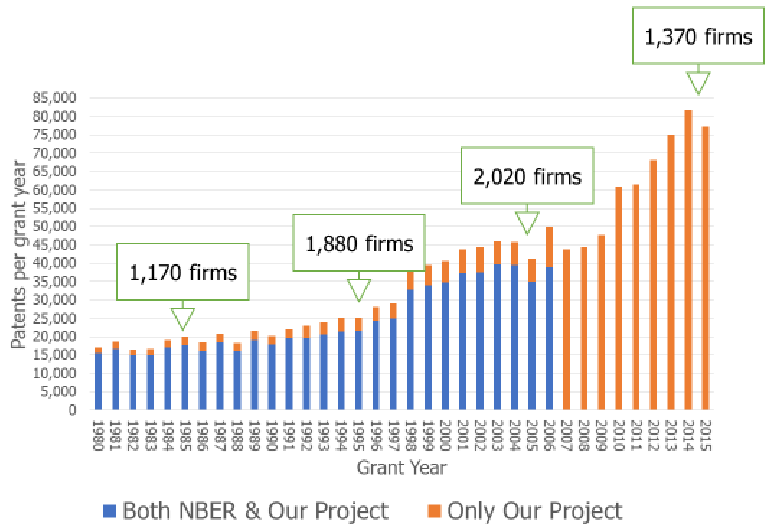


Table 6 presents the comparison results. For this period, we match about 80% of the patent-GVKEY matches as in NBER. We find an additional 17% patents due to: (i) improved dynamic linkage of patents to GVKEYs (e.g., Pharmacia), and (ii) linkage of additional patents based on historical name information, wider M&A coverage, and improved matching techniques (e.g., Phillips). In 1% of the cases, we find the same assignment as NBER, but these matches are irrelevant for our sample (e.g., Rhone-Poulenc). Lastly, in about 1% of the cases, we are unable to include the NBER matches for a variety of reasons, including possible mistakes on our end.

Table 6. Comparison to with NBER for 1980-2006: Patent-GVKEY Assignments, U.S. HQ Firms

Comparison 1980-2006	% Patents	Examples
Agreement	79	
Matched to different GVKEY	4	Improved dynamic matching to Compustat records using historical name >>> Patents of the merged company included under the GVKEY from acquisition, but not before. Example: PHARMACIA: we matched to PHARMACIA & UPJOHN's GVKEY pre-2000 instead to MONSANTO.
Only our Sample	13	Newly matched patents due to (i) availability of historical names; (ii) better M&A data; and (iii) Improved matching. e.g., PHILLIPS PETROLEUM CO: 4000+ patents pre-merger with Conoco Inc in 2002; MONSANTO: 2000+ patents pre-merger with Pharmacia;
Only NBER- we matched but irrelevant gvkey-year	1	(i) NBER match (incorrectly) based on 2006 Compstat name: e.g., ~1000 patents of RHÔNE-POULENC patents matched to RORER's GVKEY pre-merger in 1990; (ii) Improved subsidiary coverage: e.g., ~450 patents of HUGHES AIRCRAFT are incorrectly linked to GM's GVKEY pre-1985 acquisition;
Only NBER	1	(i) Withdrawn patents: ~600 patents (ii) Misc. couldn't verify connection, typos, and possible mistakes by us

REFERENCES

[1] Hall, B.H., Jaffe, A.B., and Trajtenberg, M., 2001. The NBER patent citation data file: Lessons, insights and methodological tools (No. w8498). National Bureau of Economic Research.
[2] Bessen, J., 2009. NBER PDP Project User Documentation. National Bureau of Economic Research.
[3] Wu, Y., 2010. What's in a name? What leads a firm to change its name and what the new name foreshadows. Journal of Banking & Finance, 34(6), pp.1344-1359.

Table 7. Matching Citations to Scientific Publications - Examples

Citation	Publication info			Comment
	Title	Authors	Journal information	
LIN, KUN SHAN, ET AL., SOFTWARE RULES ***GIVES*** PERSONAL COMPUTER REAL WORD POWER , **INTERNATIONAL ELECTRONICS**, VOL. **53,** NO. 3, FEB. 10, **1981**, PP. **122 125**.	"SOFTWARE RULES ***GIVE*** PERSONAL-COMPUTER REAL WORD POWER"	LIN KS, FRANTZ GA, GOUDIE K	**ELECTRONICS 54** (3): 122-125 1981	Typo in title and journal Vol.; initials vs. full name
U. WACHSMANN, R. F. H. FISCHER AND J.B. HUBER, MULTILEVEL CODES: THEORETICAL CONCEPTS AND PRACTICAL DESIGN RULES, **IEEE TRANS INFORM. THEORY**, VOL. 45, NO. 5, PP. **1361-1391**, JUL. **1999**.	"MULTILEVEL CODES: THEORETICAL CONCEPTS AND PRACTICAL DESIGN RULES"	WACHSMANN U, FISCHER RFH, HUBER JB	**IEEE TRANSACTIONS ON INFORMATION THEORY** 45 (5): 1361-1391 JUL 1999	Several names listed; variation in journal name
DESIGN CHARACTERISTICS OF GAS JET GENERATORS, **BORISOV, 1979**, PP. 21 25.	"DESIGN CHARACTERISTICS OF GAS-JET GENERATORS"	BORISOV YY	SOVIET PHYSICS ACOUSTICS-USSR 26 (1): 21-25 **1980**	Typo in year; diff in location of title within the citation
KERNS, SHERRA E., THE DESIGN OF RADIATION HARDENED ICS FOR SPACE: A COMPENDIUM OF APPROACHES, PROCEEDINGS OF THE IEEE, NOV. 1988, PP. 1470 1509.	"THE DESIGN OF RADIATION-HARDENED ICS FOR SPACE - A COMPENDIUM OF APPROACHES"	**KERNS SE**, SHAFER BD, ROCKETT LR, PRIDMORE JS, BERNDT DF, VANVONNO N, BARBER FE	PROCEEDINGS OF THE IEEE 76 (11): 1470-1509 NOV 1988	Several authors w/o "et al."
GENESTIER ET AL (BLOOD, 1997, VOL. 90, PP. 3629-3639).	"FAS-INDEPENDENT APOPTOSIS OF ACTIVATED T CELLS INDUCED BY ANTIBODIES TO THE HLA CLASS I ALPHA 1 DOMAIN"	GENESTIER L, PAILLOT R, BONNEFOYBERARD N, MEFFRE G, FLACHER M, FEVRE D, LIU YJ, LEBOUTEILLER P, WALDMANN H, ENGELHARD VH, BANCHEREAU J, REVILLARD JP	BLOOD 90 (9): 3629-3639 NOV 1 1997	No title within citation- however, perfect match in all other features
STEPHEN M. BEBGE, LYLE D. BIGHLEY AND DONALD C. MONKHOUSE PHARMACEUTICAL SALTS JOURNAL OF PHARMACEUTICAL SCIENCES, 1977, 66, 1-19.	"PHARMACEUTICAL SALTS"	**BERGE SM, BIGHLEY LD, MONKHOUSE DC**	JOURNAL OF PHARMACEUTICAL SCIENCES 66 (1): 1-19 1977	Several names listed; variation of names
L. YOUNG AND D. SHEENA, **METHODS & DESIGNS:** SURVEY OF EYE MOVEMENT RECORDING METHODS, BEHAV. RES. METHODS INSTRUM., VOL. 5, PP. 397-429, 1975.	"SURVEY OF EYE-MOVEMENT RECORDING METHODS"	YOUNG LR, SHEENA D	BEHAVIOR RESEARCH METHODS &INSTRUMENTATION 7 (5): 397-429 1975	diff in title
MICROWAVE JOURNAL, VOL. 22, NO. 2, FEB. 1979, DEDAHAM US PP. 51 52, **H. C. CHAPPELL.**	"DESIGNING IMPEDANCE MATCHED IN-PHASE POWER DIVIDERS"	CHAPPELL HC	MICROWAVE JOURNAL 22 (2): 51-52 1979	no title - however, perfect match in all other features; diff position of author's name within citation

Figure 7. External and Internal citation, matching process

(i) Example of an external citation to IBM's publication : the patent owner and cited corporate publication are different

(12) **United States Patent**
Liu et al.

(10) **Patent No.:** **US 6,468,596 B1**
(45) **Date of Patent:** **Oct. 22, 2002**

(54) **LASER-ASSISTED IN-SITU FRACTIONATED LUBRICANT AND A NEW PROCESS FOR SURFACE OF MAGNETIC RECORDING MEDIA**

(75) Inventors: **Youming Liu**, Palo Alto; **Jialuo Jack Xuan**, Milpitas; **Xiaohua Shel Yang**, Fremont; **Chung-Yuang Shih**, Cupertino; **Vidya K. Gubbi**, Milpitas, all of CA (US)

(73) Assignee: **Seagate Technology LLC**, Scotts Valley, CA (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

(21) Appl. No.: **09/577,674**

(22) Filed: **May 25, 2000**

Related U.S. Application Data

(60) Provisional application No. 60/144,357, filed on Jul. 15, 1999.

(51) **Int. Cl.**[7] **C08F 2/48**; C08J 7/18; C23C 14/30

(52) **U.S. Cl.** **427/508**; 427/554; 427/596

(58) **Field of Search** 427/510, 554, 427/555, 556, 597, 127, 226, 258, 261, 264, 270, 271, 402, 508

(56) **References Cited**

U.S. PATENT DOCUMENTS

3,674,340 A	7/1972	Jacob et al.	....	350/157
3,764,218 A	10/1973	Schedewie	....	356/118
3,938,878 A	2/1976	Fox	....	350/150

(List continued on next page.)

FOREIGN PATENT DOCUMENTS

OTHER PUBLICATIONS

P. Baumgart et al., "A New Laser Texturing Technique For High Performance Magnetic Disk Drives" IBM storage Systems Division and IBM Almadon Research Center, San Jose, CA.
D. Kuo et al., "Laser Zone Texturing on Glass and Glass–Ceramic Substrates" Seagate Recording Media, Fremont, CA.
P. Baumgart et al., "Safe Landings: Laser Texturing of High–Density Magnetic Disks" IBM Corp., *Data Storage* 1996.
A. Tam et al., "Laser Cleaning Techniques for Removal of Surface Particulates" IBM Research Division, San Jose, *Journal of Applied Physics* 71 (7), Apr. 1, 1992, pp. 3515–3523.
K. Johnson et al., "In–Plane Anisotropy in Thin–Film Physical Origins of Orientation Ratio (Invited)" IBM Storage Systems Division, San Jose, CA, *IEEE Transactions on Magnetics* vol. 31, No. 6, Nov. 1995, pp. 2721–2727.
J. Miles et al., "Micromagnetic Simulation of Textured Induced Orientation in Thin Film Media" the University of Manchester, Manchester, M13 9PL, U.K., *IEEE Transactions on Magnetics* vol. 31, No. 6, Nov. 1995, pp. 2770–2772.
C. Kissinger et al., "Fiber Optic Probe Measures Runout of Stacked Disks" B.W. Brennan Associates, *Data Storage* Jul./Aug. 1997.

Primary Examiner—Shrive P. Beck
Assistant Examiner—Eric B. Fuller
(74) *Attorney, Agent, or Firm*—McDermott, Will & Emery

(57) **ABSTRACT**

A magnetic recording medium is formed with enhanced tribological performance by applying a raw, unfractionated lubricant having a wide molecular weight distribution over a disk surface and treating the deposited lubricant with a laser light beam to effect in-situ fractionation of the lubricant to a very narrow molecular weight distribution. Embodiments of the present invention also include laser treating a deposited lubricant to increase the thickness of the bonded lube layer.

IEEE TRANSACTIONS ON MAGNETICS, VOL. 31, NO. 6, NOVEMBER 1995 2721

In-Plane Anisotropy in Thin-Film Media: Physical Origins of Orientation Ratio (Invited)

Kenneth E. Johnson, Mohammad Mirzamaani, and Mary F. Doerner
IBM Storage Systems Division, San Jose, CA 95193

(ii) Example of an internal citation to IBM's publication : the patent owner and cited corporate publication are the same

(12) **United States Patent**
Cabral, Jr. et al.

(10) **Patent No.:** **US 7,193,323 B2**
(45) **Date of Patent:** **Mar. 20, 2007**

(54) **ELECTROPLATED COWP COMPOSITE STRUCTURES AS COPPER BARRIER LAYERS**

(75) Inventors: **Cyril Cabral, Jr.**, Ossining, NY (US); **Stefanie R. Chiras**, Peekskill, NY (US); **Emanuel Cooper**, Scarsdale, NY (US); **Hariklia Deligianni**, Tenafly, NY (US); **Andrew J. Kellock**, Sunnyvale, CA (US); **Judith M. Rubino**, Ossining, NY (US); **Roger Y. Tsai**, Yorkstown Heights, NY (US)

(73) Assignee: **International Business Machines Corporation**, Armonk, NY (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

(21) Appl. No.: **10/714,966**

(22) Filed: **Nov. 18, 2003**

(65) **Prior Publication Data**

US 2005/0104216 A1 May 19, 2005

(51) **Int. Cl.**
H01L 23/48 (2006.01)
H01L 23/52 (2006.01)

(52) **U.S. Cl.** **257/751**; 257/752; 257/762

(58) **Field of Classification Search** 257/751–753, 257/758, 759, 761–763

See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

5,695,810 A	12/1997	Dubin et al.		
6,168,991 B1 *	1/2001	Choi et al.	....	438/254
6,323,128 B1	11/2001	Sambucetti et al.		
6,342,733 B1	1/2002	Hu et al.		
6,528,409 B1	3/2003	Lopatin et al.		
6,573,606 B2 *	6/2003	Sambucetti et al.	....	257/762
2003/0010645 A1	1/2003	Ting et al.		
2003/0075808 A1 *	4/2003	Inoue et al.	....	257/774

OTHER PUBLICATIONS

A. Kohn, et al., "Characterization of electroless deposited Co(W,P) thin films for encapsulation of copper metallization" Materials Science and Engineering A302 (2001) pp. 18-25.
C.-K. Hu, et al., "Reduced electromigration of Cu wires by surface coating" IBM T.J. Watson Research Center, Yorktown Heights, New York, 2002.

(Continued)

Primary Examiner—Hung Vu
(74) *Attorney, Agent, or Firm*—Connolly Bove Lodge & Hutz, LLP; Robert M. Trepp

(57) **ABSTRACT**

A composite material comprising a layer containing copper, and an electrodeposited CoWP film on the copper layer. The CoWP film contains from 11 atom percent to 25 atom percent phosphorus and has a thickness from 5 nm to 200 nm. The invention is also directed to a method of making an interconnect structure comprising: providing a trench or via within a dielectric material, and a conducting metal containing copper within the trench or the via; and forming a CoWP film by electrodeposition on the copper layer. The CoWP film contains from 10 atom percent to 25 atom percent phosphorus and has a thickness from 5 nm to 200 nm. The invention is also directed to a interconnect structure comprising a dielectric layer in contact with a metal layer; an electrodeposited CoWP film on the metal layer, and a copper layer on the CoWP film.

18 Claims, 6 Drawing Sheets

APPLIED PHYSICS LETTERS VOLUME 81, NUMBER 10 2 SEPTEMBER 2002

Reduced electromigration of Cu wires by surface coating

C.-K. Hu,[a)] L. Gignac, R. Rosenberg, E. Liniger, J. Rubino, and C. Sambucetti
IBM T. J. Watson Research Center, Yorktown Heights, New York 10598

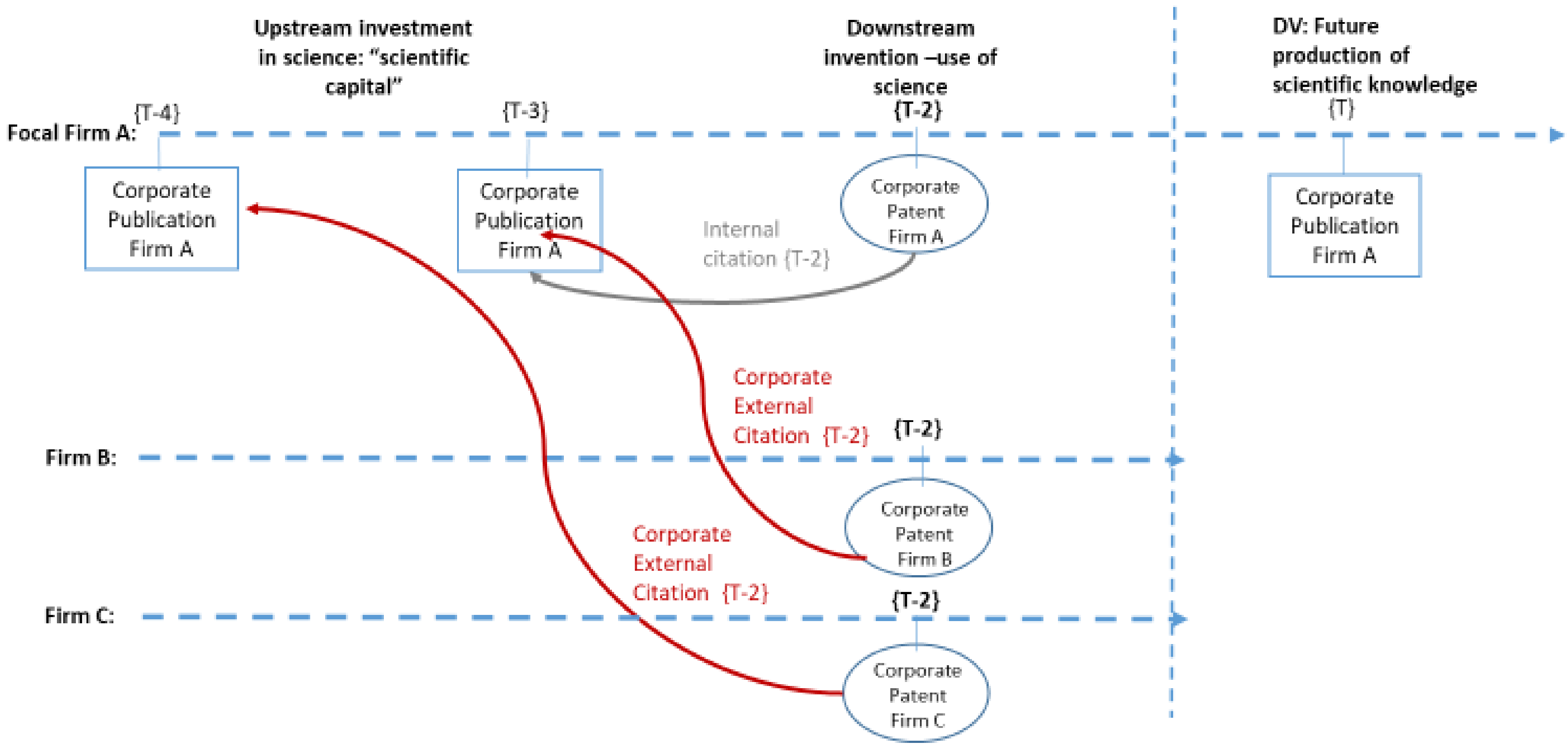
A. Domenicucci and X. Chen
IBM Microelectronics Division, Hopewell Junction, New York 12533

A. K. Stamper
IBM Microelectronics Division, Essex Junction, Vermont 05452

(Received 28 May 2002; accepted for publication 11 July 2002)

Note: this figure presents examples of front-page patent reference to non-patent literature. Below each patent reference is the related scientific publication that is being cited. Example (i) is an external patent citation to IBM's publication and example (ii) is an internal patent citation to IBM's publication.

Figure 8. Timeline- Production and Use of Research



Note: this figure illustrates the temporal structure of citations and publications. At time T-2 the focal firm (Firm A) has: (i) one Internal citation and (ii) two corporate external citations from patents filed by sample Compustat firms (Firm B and C).

Supplementary analysis not for publication: Trends in corporate publications

March 9, 2020

We provide econometric evidence supporting the trend in publications presented in Figure 2. This analysis builds on and updates the analysis in (Arora et al., 2018). We estimate time trends in the rate of corporate publications as follows:[1]

$$\ln(Publications_{it}) = \alpha_0 + \alpha_1 Trend + \alpha_2 R\&D\ stock_{it-2} + \eta_i + \epsilon_{it} \tag{1}$$

$Publications_{it}$ is the annual flow of publications by firm i in year t, $Trend$ is the time trend computed as year t minus 1980 and is presented in decennial units (i.e., per decade). η_i is a complete set of firm dummies. ϵ_{it} is an iid error term.

Table 1 presents the estimation results. Consistent with Figure 2, we expect a falling publication rate, $\hat{\alpha_1} < 0$. This is confirmed by our estimates. Column 1 presents results from a pooled specification with a complete set of 4-digit SIC industry dummies, without firm fixed-effects. $\hat{\alpha_1}$ implied that over our sample period, covering 3.5 decades, publication rate has fallen by more than 44% (-0.125 .5). Column 2 adds firm fixed effects, which lowers $\hat{\alpha_1}$ considerably, implying an overall decline of about 20%. A possible explanation for the large fall in the time trend estimate is a substantial entry of low-publishing firms over the sample years. That is, firms entering the public equity markets over time are less likely to publish scientific research. Columns 3 presents similar estimates when weighting publications by the number of citations they receive from other publications and normalizing each citation by the average number of

[1]Unless stated otherwise, one is added to number of publications and all specifications include a dummy variable for firm-year observations with zero publications.

citations received by all other WoS publications published in the same year to account for truncation.

Because the number of journals is rising over time, comparing early to late publication rates might underestimate the fall in corporate publications when not accounting for the rise in available journal space. Column 4 presents time trend estimates when excluding new journals (journals established post-1990). As expected, the time trend effect rises in absolute value when holding journal space constant throughout the later sample period. Publications fall by about 35%.[2]

Columns 5 and 6 split the sample by firm size using median sales value. Large firms account for ten times the publications, and they exhibit the greatest shift in R&D composition, as reflected in the much sharper decline in publication rates.[3]

Columns 7 and 8 break up the time trends into eight periods and includes separate dummies for each period (the base period is 1980-1985) to account for non-linear time effects. The magnitude of the decline is similar to that captured by $Trend$.

Columns 9-12 examine the robustness of our results to having zeros in our dependent variables. Column 8 excludes firm-year observations with zero publications. The trend estimate increases substantially and indicates a fall in publications of about 66% over our sample period. To reduce the prevalence of observations with zero publications, Column 10 restructures our panel to firm-5-year cohorts (instead of firm-1-year) using firm-year averages (hence, instead of having 35 periods, this specification includes only 7 periods. $Trend$ is defined accordingly, with the value of 0 for period 1 and the value of 7 for the last period). The trend effect increases as well, indicating about 35% decline in publications over our sample period.

Column 11 estimates our original panel using a negative binomial specification with firm fixed effects accounted for using pre-sample means (Blundell et al., 1999). For each firm in our

[2] We also experimented with removing only the low quality new journals (journals with impact factor below unity) The decline in publications is about 30% over the complete sample period.

[3] In an unreported specification, we estimate trend for a subsample of very large firms (90th percentile of sales value). For these firms, the trend estimate is substantially larger (0.203 with a standard error of 0.033, and a sample average publication of 63), indicating a fall of 70% in publication rate over the complete sample period, or a loss of 45 publications per firm over the same period.

sample, we calculate the 4-year average value of publications and exclude these years from our sample. We refer to these average values as pre-sample means– our firm fixed effects control in the regression. The implied total decline in publication rate is similar to the within-firm estimates obtained from OLS. Column 12 generates similar estimates as Column 2 using an Inverse hyperbolic sine.[4]

From Table 1, we conclude that firms are withdrawing from research, and that later entrants to the sample are less engaged in research. Further, these results are not artifacts of how the dependent variable is defined, nor of the estimation method.

Table 2 presents time trends across industries. We focus on three main industry groups[5]: (i) life sciences, (ii) IT & Software and Communication, and (iii) Chemicals & Energy. There is substantial heterogeneity in the behavior of corporate publications over time by industry. While there is a decline in the publication rate in the latter two groups (namely in ICT, and Chemicals & Energy), the pattern for life science is less clear. For life sciences, there is an increase in publication rate followed by a gradual decline until the end of the sample period, where publication rate is statistically the same as at the beginning of the sample period.

Several factors may be responsible for the different time trends in life sciences. Insofar as patents are more effective in protecting innovations in life sciences relative to other industries, the returns from investments in research may be higher in the pharmaceutical sector than in

[4]In auxiliary unreported analysis we address the concern that some of the decline in publication output may reflect greater secrecy about scientific research rather than a decline in scientific research itself. A variety of direct and indirect evidence suggests that this is unlikely to be the entire story. For one, there are well documented cases of firms such as Xerox, HP, IBM, AT&T, and DuPont reducing their scientific research. As well, NSF Science and Engineering surveys show that privately performed basic and applied research, as a share of total private R&D, has fallen steadily over time since the mid-1980s. To further examine the secrecy explanation in our context, we present the following test. If firms are persisting in research but merely keeping it secret instead of publishing, we would expect a larger fall in publication rate for firms in states that extend greater protection to trade secrets. We follow Klasa et al. (2018) and exploit variation in the adoption of Inevitable Disclosure Doctrine (IDD) by U.S. state courts. IDD is a legal doctrine that restricts worker mobility from one organization to another in cases where they might be inevitably disclosed trade secrets. It is applicable even if the employee did not sign a non-compete or non-disclosure agreement, if there is no evidence of actual disclosure, or if the rival is located in another state. We create an IDD dummy variable that receives the value of one if IDD is in effect in the focal firm's state in a given year, and zero otherwise. We add an interaction term between IDD and trend. If secrecy drives the drop in publication rate, we expect a negative and significant interaction effect. That is, the drop in publication rate should be larger for firms operating in states with stronger trade secret protection. Yet, the evidence are inconsistent with the secrecy story. The coefficient estimate on the trend-IDD interaction is positive rather than negative, and small and statistically indistinguishable from zero.

[5]Table 3 includes a list of all four-digit SIC codes that comprise each industry group.

other sectors, making spillovers less harmful. The commercial applicability of upstream research is also much more apparent in the pharmaceutical industry (Li et al. (2017)). Consistent with this, we find that publications by life-science firms receive 2.5 more citations from patents than publications by non-life-science firms do (based on own calculations), underscoring the higher relevance of research to invention in the sector. Finally, corporate research in life sciences may have benefited from biomedical research funded by the National Institutes of Health, which increased dramatically, from US$2.5 billion in 1980 to US$29 billion in 2015.

References

Arora, A., Belenzon, S., & Patacconi, A. (2018). The decline of science in corporate R&D. Strategic Management Journal, 39(1), 3–32.

Blundell, R., Griffith, R., & van Reenen, J. (1999). Market Share, Market Value and Innovation in a Panel of British Manufacturing Firms. The Review of Economic Studies, 66(3), 529–554.

Klasa, S., Ortiz-Molina, H., Serfling, M., & Srinivasan, S. (2018). Protection of trade secrets and capital structure decisions. Journal of Financial Economics, 128(2), 266–286.

Li, D., Azoulay, P., & Sampat, B. N. (2017). The applied value of public investments in biomedical research. Science, 356(6333), 78–81.

Table 1: Corporate Publication Over Time

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Dependent variable	ln(1+Number of publications)						ln(1+Number of publications)		ln(Number of publications)		Number of publications	Inverse hyperbolic sine
	Pooled	Within Firms	Cite-weighted	Excluding new journals	Large firm	Small firms	Period dummies	Excluding new journals	Exclude zeros	5-year cohorts	Negative Binomial	OLS
Time trend	-0.125	-0.057	-0.045	-0.111	-0.109	-0.037			-0.194	-0.090	-0.122	-0.065
	(0.014)	(0.012)	(0.013)	(0.013)	(0.019)	(0.011)			(0.031)	(0.016)	(0.027)	(0.014)
Time dummy for												
2011 Year 2015							-0.201	-0.347				
							(0.036)	(0.038)				
2006 Year 2010							-0.173	-0.304				
							(0.033)	(0.034)				
2001 Year 2005							-0.142	-0.247				
							(0.029)	(0.031)				
1996 Year 2000							-0.080	-0.192				
							(0.025)	(0.026)				
1991 Year 1995							-0.092	-0.160				
							(0.021)	(0.021)				
1986 Year 1990							-0.100	-0.108				
							(0.015)	(0.015)				
1980 Year 1985							Base	Base				
ln(R&D stock)t-2	0.274	0.125	0.116	0.095	0.106	0.029	0.125	0.092	0.277	0.244	0.364	0.141
	(0.017)	(0.014)	(0.015)	(0.012)	(0.016)	(0.012)	(0.014)	(0.012)	(0.034)	(0.036)	(0.029)	(0.016)
ln(Sales)t-2	0.000	0.055	0.047	0.045	0.172	0.014	0.054	0.043	0.146	0.174	0.173	0.062
	(0.000)	(0.006)	(0.006)	(0.005)	(0.016)	(0.005)	(0.006)	(0.005)	(0.017)	(0.025)	(0.016)	(0.007)
Pre-sample FE											0.708**	
											(0.029)	
Firm fixed-effects	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes
Dep Var Avg:	15.907	15.907	20.037	11.464	25.485	2.101	15.907	11.464	35.992	23.174	18.275	15.907
Number of firms	3807	3807	3807	3807	1888	1919	3807	3807	2533	2640	3030	3807
Observations	45,496	45,496	45,496	45,496	26,862	18,634	45,496	45,496	20,108	6,817	34,889	45,496
R-squared	0.65	0.93	0.89	0.94	0.94	0.91	0.94	0.94	0.86	0.92	0.26	0.93

Notes: This table examines time trends in scientific publication for main industries. Industry classification is based on SIC codes (see Table 3 for detailed list). All columns include a dummy variable that equals one for years with zero publications. Standard errors (in brackets) are robust to arbitrary heteroscedasticity.

Table 2: Corporate Publication Time Trend by Industry

	Dependent variable: ln(1+Number of publications)											
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Industry:	Life science			IT, Software & Communication			Chemicals and Energy			Others		
	Within Firms	Excluding new journals	Inverse hyperb. sine	Within Firms	Excluding new journals	Inverse hyperb. sine	Within Firms	Excluding new journals	Inverse hyperb. sine	Within Firms	Excluding new journals	Inverse hyperb. sine
Cohort dummies for:												
2011 Year 2015	0.169	0.073	0.128	-0.313	-0.408	-0.338	-0.490	-0.688	-0.558	-0.081	-0.228	-0.099
	(0.125)	(0.117)	(0.134)	(0.126)	(0.126)	(0.139)	(0.115)	(0.125)	(0.130)	(0.042)	(0.041)	(0.048)
2006 Year 2010	0.219	0.130	0.190	-0.275	-0.375	-0.295	-0.422	-0.592	-0.477	-0.082	-0.211	-0.100
	(0.119)	(0.113)	(0.126)	(0.117)	(0.114)	(0.130)	(0.098)	(0.109)	(0.110)	(0.040)	(0.039)	(0.046)
2001 Year 2005	0.323	0.243	0.310	-0.255	-0.333	-0.273	-0.334	-0.481	-0.373	-0.081	-0.183	-0.098
	(0.109)	(0.105)	(0.113)	(0.114)	(0.112)	(0.125)	(0.089)	(0.097)	(0.100)	(0.035)	(0.036)	(0.041)
1996 Year 2000	0.436	0.362	0.438	-0.179	-0.281	-0.185	-0.189	-0.365	-0.207	-0.055	-0.167	-0.066
	(0.099)	(0.097)	(0.101)	(0.099)	(0.104)	(0.107)	(0.068)	(0.078)	(0.077)	(0.031)	(0.031)	(0.035)
1991 Year 1995	0.474	0.425	0.477	-0.188	-0.241	-0.201	-0.130	-0.231	-0.146	-0.087	-0.155	-0.104
	(0.094)	(0.092)	(0.094)	(0.085)	(0.083)	(0.091)	(0.053)	(0.059)	(0.061)	(0.025)	(0.026)	(0.029)
1986 Year 1990	0.302	0.286	0.307	-0.167	-0.152	-0.182	-0.150	-0.161	-0.169	-0.092	-0.103	-0.110
	(0.069)	(0.068)	(0.070)	(0.060)	(0.057)	(0.065)	(0.042)	(0.043)	(0.049)	(0.018)	(0.018)	(0.021)
1980 Year 1985	Base	Base	Base	Base	Base	Base	Base	Base	Base	Base	Base	Base
ln(R&D stock)t-2	0.266	0.206	0.295	0.125	0.082	0.135	0.161	0.146	0.183	0.066	0.046	0.076
	(0.031)	(0.030)	(0.035)	(0.067)	(0.056)	(0.071)	(0.037)	(0.038)	(0.042)	(0.015)	(0.012)	(0.018)
ln(Sales)t-2	0.041	0.033	0.045	0.045	0.031	0.050	0.089	0.078	0.102	0.053	0.043	0.063
	(0.010)	(0.009)	(0.012)	(0.017)	(0.014)	(0.019)	(0.025)	(0.027)	(0.027)	(0.009)	(0.007)	(0.010)
Firm fixed-effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Dependent variable average:	41.798	31.751	41.798	29.589	20.530	29.589	21.565	17.888	21.565	6.549	4.440	6.549
Number of firms	634	634	634	372	372	372	293	293	293	1780	1780	1780
Observations	6,440	6,440	6,440	4,303	4,303	4,303	4,321	4,321	4,321	22,229	22,229	22,229
R-squared	0.93	0.94	0.93	0.94	0.94	0.94	0.94	0.94	0.94	0.92	0.93	0.92

Notes: This table examines time trends in scientific publication for main industries. Industry classification is based on SIC codes (see Table 3 for detailed list). All columns include a dummy variable that equals one for years with zero publications. Standard errors in brackets are robust to arbitrary heteroscedasticity.

Table 3: SIC Classification by Main Industries

Category	**Description**	**Related 4-digit SIC codes for our firm sample**
Life science (Medical, Pharmaceuticals and Biotechnology)	Drugs, pharmaceuticals, biotech and medical devices- Manufacture, Sale & Services	2833 2834 2835 2836 5122 5912 8060 8071 8082 8090 8093 8731 8734
IT & Software & Communication	IT & Software - Development, Provider, Sale & Services; Telecom, Communication- system, equipment, services;	3661 3663 3669 4812 4813 4822 4832 4833 4841 4888 4899 5040 5045 7370 7371 7372 7373 7374
Chemicals & Energy	Chemicals- Manufacture & Sale. Energy: Electricity, Oil, Gas, Power station- including: utility, exploration, equipment, services, etc.	1000 1040 1220 1311 1381 1382 1389 1400 2800 2810 2820 2821 2840 2842 2844 2851 2860 2870 2890 2891 2911 2950 2990 3320 3330 3334 3341 3350 3357 3360 3390 3460 3470 4923 5051 5160