Abstract

We investigate the impact of the introduction of patent rights for software inventions in the U.S. on university computer scientists’ scientific output. Difference-in-difference estimations reveal that the regime shift towards stronger commercialization options causes software scientists at U.S. universities to produce fewer (both count and quality weighted) publications than their European counterparts. The effect is largest for scientists with the lowest pre-shock citation-weighted publication output. This pattern suggests a sorting of U.S. scientists of a low pre-shock citation-weighted productivity into commercialization activities and is consistent with a simple model of time allocation between patenting and publishing.
The Effects of Patent Rights on University Science

February 25, 2019

Abstract

We investigate the impact of the introduction of patent rights for software inventions in the U.S. on university computer scientists’ scientific output. Difference-in-difference estimations reveal that the regime shift towards stronger commercialization options causes software scientists at U.S. universities to produce fewer (both count and quality weighted) publications than their European counterparts. The effect is largest for scientists with the lowest pre-shock citation-weighted publication output. This pattern suggests a sorting of U.S. scientists of a low pre-shock citation-weighted productivity into commercialization activities and is consistent with a simple model of time allocation between patenting and publishing.

JEL codes: O3; K2; L3

Keywords: patent rights; publications; economics of science; difference-in-difference estimation
1 Introduction

In the past decades, numerous policy initiatives strengthened universities’ institutional ownership rights of intellectual property (IP) such as the Bayh-Dole Act that granted universities the right to patent faculty inventions (e.g., Henderson et al., 1998; Mowery and Ziedonis, 2002; Sampat et al., 2003 on the U.S. Bayh-Dole Act, Czarnitzki et al., 2015, 2016, for Germany, Hvide and Jones, 2018, for Norway). These law changes set incentives for universities to increase their engagement in commercialization, in line with the modern university’s emphasis on technology transfer as the third mission (Etzkowitz and Leydesdorff, 2000). They also fueled long-standing concerns that the growing involvement of individual scientists in commercialization of their research may have negative implications for the traditional research process (e.g. Blumenthal et al., 1996; Campbell et al., 2002; Krimsy, 2003; Murray and Stern, 2007; Fabrizio and Di Minin, 2008; Azoulay et al., 2009; Czarnitzki et al., 2009, 2011).

The effects of such IP regime shifts for universities are difficult to assess because regime shifts affect the entire university landscape in an economy so that a counterfactual situation is missing. Moreover, these changes are typically not exogenous, but happen in response to an unsatisfactory level of technology transfer from universities to the private sector.1 Also, such policy interventions are typically designed as technology transfer packages (e.g. Czarnitzki et al., 2016)2 which makes it impossible to disentangle the effects of the individual means of the package.

Here, we focus on the broadening of the patentable subject matter to include software

---

1 The German transfer of ownership rights for university inventions from the professor to the university in 2002, was, for instance, a reaction to the “European paradox” (European Commission, 1995). At that time, policymakers believed that Germany had one of the world’s leading scientific research enterprises, but was lagging the U.S. in terms of technology transfer and commercialization. Similarly, was the introduction of Bayh-Dole Act types of policies in Europe largely inspired by the European fear to fall behind (Genna and Rossi, 2011; Lissoni et al., 2008; Czarnitzki et al., 2011).

2 The transfer of ownership rights of university inventions from the professor to the university in Germany was, for instance, part of a policy program called “Knowledge Creates Markets” which aimed at stimulating technology transfer from universities to private industry for innovation and economic growth (Bundesministerium für Bildung und Forschung and Bundesministerium für Wirtschaft und Technologie, 2001). The program addressed a wide spectrum of science-industry interactions including processes and guidelines governing knowledge transfer, science-based spin-offs, collaboration, and the exploitation of scientific knowledge in the private sector.
inventions in the U.S. in 1996 in order to investigate the effects of patent rights on faculty science. The introduction of patent rights for software inventions represents a regime shift towards stronger commercialization options and, hence, operated the same mechanism as Bayh Dole Act-type of policies. At the same time, the IP law change did not aim at universities or at enhancing technology transfer. This renders the introduction of software patents an attractive setting for studying the impact of a regime shift towards stronger IP ownership rights for universities on faculty science.

Focusing on computer scientists in the field of software employed at U.S. universities, we use difference-in-difference estimations with software scientists affiliated with European universities as a control group. Prior to 1996, software was not patentable per se in the U.S., only when tied to physical or mechanical processes (Graham and Mowery, 2003). This situation still corresponds to the status quo in Europe (Guntersdorfer, 2003). Our empirical results show that the regime shift led to a lower quantity of (citation-weighted) publications. The effect is stronger for scientists at the left-hand side of the quality-weighted publication distribution which hints at a sorting of scientists of low quality-weighted productivity into commercial activities. This pattern is consistent with a simple model of time allocation between patenting and publishing which predicts that scientists at the lower tail of the scientist ability distribution with the, hence, lowest citation-weighted pre-shock publication output have the highest incentives to allocate time to commercialization efforts. They have the lowest opportunity cost. Empirical results confirm: the decline in publications is 31% for U.S. scientists at the bottom of the ability distribution and 12% for those at the top of the productivity distribution.

Our results cannot rule out concerns about negative implications for science of a regime shift towards stronger commercialization options due to strengthened IP ownership rights for universities. Therewith our study makes a contribution to technology transfer policy evaluation (e.g. Henderson et al., 1998; Sampat et al., 2003; Mowery and Sampat, 2005; Czarnitzki et al., 2015, 2016; Hvide and Jones, 2018). The finding of negative effects for the domain of software science, an applied science field where the value of patents remains disputed,\textsuperscript{3}

\textsuperscript{3}Legal scholars have disputed the breadth of software patent claims (Burk and Lemley, 2003; Rai, 2003)
suggests stronger effects for other disciplines characterized by a higher degree of basicness of science and by less fuzzy patent rights. Our results further reveal that the introduction of patent rights leads to a sorting of scientists of low quality-weighted productivity into the commercialization track. Exploiting a better match of skills and tasks of faculty is essential for the success of a modern university that emphasizes technology transfer as the third mission of the university.

2 Background

This section provides an overview on the patentability of software in the U.S. and Europe. For further details on the legal background we refer to Graham and Mowery (2003) for the U.S. case, Bakels et al. (2008) for the case of Europe and Guntersdorfer (2003) for a comparison.

2.1 The U.S. case

In the 1970s, the predominant way to protect software in the U.S. was copyright. Algorithms were deemed to not be patentable at the United States Patent and Trademark Office (USPTO) which was confirmed by a number of Supreme Court decisions in the 1970s (Graham and Mowery, 2003). The case Gottschalk versus Benson (409 U.S. 63 (1972)) explicitly rejected software as a patentable subject matter and the 1976 Copyright Act explicitly endorsed copyright as an appropriate protection regime for software. As noted by Hall and MacGarvie (2010), patents and copyrights protect very different aspects of software though. While copyright is awarded to creators of original works and protects a specific computer code as an “original expression”, it does not protect the functions performed by the code. The function of a software program may be protected by a patent.4

In the 1980s, patent law started to slowly change in favor of software patents with the ruling that software can be patented if tied to physical or mechanical processes. The change was and the allegedly poor quality of prior art documentation (Lunney Jr, 2000) questioning the validity of software patents per se.

4See Graham and Mowery (2003) for a detailed discussion of the usage of different IP protection means for software inventions.
initiated by the case decision Diamond versus Diehr (450 U.S. 175 (1981)) where the Supreme Court decided on the patentability of a rubber-curing process that used software to calculate cure time. The physical transformation of rubber “into a different state or thing” took the invention out of the realm of abstraction. The subject matter was declared patentable even though software implementation represented the only novel feature of the invention. During the first half of the 1990s, a number of court decisions which are described in more detail by e.g. Graham and Mowery (2003) and Hall and MacGarvie (2010) spurred a discussion about the broadening of the patentable subject matter with important implications for software. An important step towards a new legislation was taken in 1994 by the Court of Appeals of the Federal Circuit (CAFC) that distinguished between patentable software as “rather a specific machine to produce a useful, concrete, and tangible result” and unpatentable software as a disembodied mathematical concept such as a law of nature, natural phenomenon, or an abstract idea (Sterne and Bugaisky, 2004). After a series of further cases in 1994\(^5\) - with the last one being that the CAFC ruled that the rejection of a software patent application at the USPTO by IBM was erroneous in 1995\(^6\) - the U.S. Commissioner of Patents issued new patentability guidelines in 1996 which allowed inventors to patent any software embodied in physical media (Sterne and Bugaisky, 2004)\(^7\).

Although their announcement was perceived negatively by the stock market, the new patentability guidelines were followed by a surge in software patenting in the private sector during the 1996–1999 (Hall and MacGarvie, 2010) which has been attributed to a large extent to strategic considerations rather than to an increase in innovation (Bessen and Hunt, 2007; Noel and Schankerman, 2013). As patents for general purpose technologies finding application in complex product industries, software patents are strongly associated with legal disputes (Bessen and Meurer, 2005) and market entry barriers (Cockburn and MacGarvie, 2011). Legal scholars dispute the breadth of software patent claims (Burk and Lemley, 2003; Rai, 2005).

---

\(^5\)In re Alappat (5440676), In re Warmerdam (6089742), In re Lowry, In re Trovato.

\(^6\)In re Beauregard (5710578)

\(^7\)The guidelines specified that it should be distinguished between (a) “a computer or a programmable apparatus controlled by software as a statutory ‘machine’”, (b) a computer-readable memory used to direct a computer such as a memory device, a compact disc or a floppy disk as a statutory ‘article of manufacture’ and (c) a series of steps to be performed on or with the aid of a computer as a statutory’s process.” (USPTO guidelines, 1996, https://www.uspto.gov/web/offices/com/sol/og/con/files/cons093.htm)
and the poor quality of prior art documentation (Lunney Jr, 2000). Overall, the value of software patents for the private sector remains disputed (see Gallini, 2002, for a summary of the arguments).

With focus on the university landscape, software was already one of the fields in which universities had licensing agreements before the introduction of software patent rights (Mowery et al., 2001). After 1996, the number of university-held software patents decupled over the period 1982-2002 from 37 patents in 1982 to 396 patents in 2002 which corresponded to an increase in the share of software patents among university patents of 4% (Rai et al., 2009). The disproportionate increase in university software patents has been attributed to economies of scale realized through technology transfer offices (TTOs) (Rai et al., 2009; Graham and Mowery, 2003). TTOs started to play a more active role in the field of computer science after the introduction of patent rights (Rai et al., 2009). Thereby, TTOs took a “one size fits all” approach in the sense that the propensity to apply for patent protection for a software invention was predominantly determined by the TTO’s tendency to seek for patent protection in other disciplines (Rai et al., 2009). This implied that computer scientists faced a closer scrutiny of their inventions by the TTO. Moreover, patents became important for tenure, promotion and annual salary raises (Love, 2013). Due to the significant changes for computer scientists with regard to the role of the TTO, output expectations and career requirements we refer to the introduction of software patents as a regime shift.

2.2 The European case

Whereas the U.S. Patent Act of 1952 laid the foundation for expansion of the patentable subject matter (Sterne and Bugaisky, 2004), Article 52(2) of the European Patent Convention (EPC) explicitly excludes specific categories of inventions such as business methods and software. These inventions do not fulfill the technical contribution requirement. Article 52(2) specifies that software is not patentable “as such”. Further guidelines are provided by case decisions of the Technical Board of Appeal of the EPO.\textsuperscript{8} According to those, software

\textsuperscript{8}The “as such” clause leaves some room for interpretation. A decision of the Technical Board of Appeal of the EPO from 1988 holds, for instance, that “even if the basic idea underlying an invention may be considered
may, for instance, be patented if tied to physical or mechanical processes. A proposal for a Directive on the Patentability of Computer-implemented Inventions (known as the CII Directive) which was made to improve clarity on the treatment of software inventions under European patent law was rejected in 2005 (González, 2006). Hence, the legal situation in Europe corresponds to the legal situation in the U.S. before the introduction of the new patent guidelines and after the Diamond versus Diehr case (Guntersdorfer, 2003).

3 A simple model of scientific production

In this section, we introduce a simple model of science production to formalize the consequence of the introduction of patent rights for software inventions. In this model, scientists can engage in science or commercialization activities and accordingly produce two different kinds of output: publications and patents. In a given period, the production of both outcomes is based on a Poisson process that depends on the scientist’s ability with decreasing returns to scale:

\[
\text{Pub}_{it} \sim \text{Poisson} \left( \sqrt{\gamma_{it}^{\text{pub}} \delta_{it}} \right), \quad \text{Pat}_{it} \sim \text{Poisson} \left( \sqrt{\gamma_{it}^{\text{pat}} (1 - \delta_{it})} \right),
\]

with \(\text{Pub}_{it}\) (\(\text{Pat}_{it}\)) representing the number of publications (patents) produced by scientist \(i\) in period \(t\). The variables \(\gamma_{it}^{\text{pub}}\) and \(\gamma_{it}^{\text{pat}}\) reflect the scientist’s ability to work on science resulting in publications or commercialization activities resulting in patents respectively. Finally, \(\delta_{it} \in [0, 1]\) is the fraction of time allocated to working on publications. Ability levels of scientists \(\gamma_{it}^{\text{pub}}\) and \(\gamma_{it}^{\text{pat}}\) are taken as given and the allocation of time \(\delta\) is the only decision variable of the scientists.

For simplicity, we assume that both patents and publications confer the same utility to

to reside in a computer program a claim directed to its use in the solution of a technical problem cannot be regarded as seeking protection for the program as such within the meaning of Article 52(2)(c) and (3) EPC” (T 0115/85, 1988). However, “a computer program product is not excluded from patentability under Article 52(2) and (3) EPC if, when it is run on a computer, it produces a further technical effect which goes beyond the “normal” physical interactions between program (software) and computer (hardware)” (T 1173/97, 1998). In other words, if the invention covers a “trick” of how to do something rather than a program code is can be patentable (Bakels et al., 2008).
the scientist, normalized to unity. This can be thought of as a scientist striving for tenure or a promotion at a university that gives credit for both publications and patents.\textsuperscript{9} The utility can then be written as:

\[ U_{it} = Pub_{it} + Pat_{it}. \]

Assuming that scientists are risk neutral, maximizing their utility leads to the following optimal allocation of time:

\[ \delta_{it}^* = \frac{\gamma_{it}^{pub}}{\gamma_{it}^{pub} + \gamma_{it}^{pat}}. \]

As we can see, the higher the ability for science leading to publications the more time will be allocated to it.

At the optimum, the expected number of publications depends positively on $\gamma^{pub}$ and negatively on $\gamma^{pat}$ and can be written as:

\[ \hat{Pub}_{it} \equiv E(Pub_{it}) = \sqrt{\frac{\gamma_{it}^{pub} \delta_{it}^*}{\gamma_{it}^{pub} + \gamma_{it}^{pat}}}. \]

With this simple model at hand, we can analyze the consequence of the law change. The introduction of patent rights for software inventions translates to an increase in $\gamma^{pat}$ after the shock. Say now $\gamma_{it}^{pat}$ increases by an amount $\theta$. To analyze the consequence for the production of publications, consider an individual $i$ and her counterfactual $k$ who has the same characteristics but is not affected by the law change. The log of the new production of

\textsuperscript{9}Note that assuming different utilities would lead qualitatively to the same results.
The publication of individual $i$ corresponds to:

$$
\log \left( \hat{P}_{\text{ub}it} \right) = \log \left( \gamma_{\text{pub}it} \right) - \frac{1}{2} \log \left( \gamma_{\text{pub}it} + \gamma_{\text{pat}it} + \theta \right)
$$

$$
= \log \left( \gamma_{\text{pub}it} \right) - \frac{1}{2} \log \left( \gamma_{\text{pub}it} + \gamma_{\text{pat}it} \right) - \frac{1}{2} \log \left( 1 + \frac{\theta}{\gamma_{\text{pub}it} + \gamma_{\text{pat}it}} \right)
$$

$$
= \underbrace{\log \left( \hat{P}_{\text{ub}kt} \right)}_{\text{publication outcome before the law change}} - \underbrace{\frac{1}{2} \log \left( 1 + \frac{\theta}{\gamma_{\text{pub}it} + \gamma_{\text{pat}it}} \right)}_{\text{decrease in publication outcome after the law change}}.
$$

We clearly see that the introduction of patent rights decreases the production of publications since the incentive to allocate time to working on patents is increased. Furthermore, we observe that the magnitude of the shock depends on the initial characteristics of the scientists. The higher the ability for science ($\gamma_{\text{pub}}$), the lower the decrease in publication outcome. The empirical part of the paper will test these hypotheses.

## 4 Data, variables and descriptive statistics

### 4.1 Data

The population of interest includes all scientists in computer science working on software related topics at U.S. and European universities. We defined the study period to extend from 1989 through 2004 so that we observe enough years before and after the policy change. Based on publication records retrieved from the Web of Science (WoS), we constructed a panel dataset of academic scientists in the field of computer science following a multistep procedure, which is detailed in Appendix A. Computer science is defined by the WoS subfields Artificial Intelligence, Information Systems, Theory & Methods, Software Engineering, Interdisciplinary Applications, Hardware & Architecture, Cybernetics. Since the field of computer science is broad, it occurs frequently that scientists coming from other university departments publish in this field. We therefore only kept scientists with a minimum of three computer science publications in three different years to qualify for our sample. Two of those publications have to be from before 1996. Using other thresholds does not alter the findings.
Scientists working on hardware related topics have been excluded from the treatment group as well as scientists that were switching affiliations between the U.S. and Europe. For each computer scientist, our data contains the individual’s history of publishing between 1989 and 2004 supplemented by the individual’s history of patents between 1979 and 2004. Computer science publications include articles published in scientific journals and in conference proceedings. Conference proceedings are an important outlet for computer scientists which is used for a speedy dissemination of results (Patterson et al., 1999). Our final sample consists of 8,133 computer scientists, among them 4,437 European, working on software related topics. In total, these scientists produced 86,756 unique publications.

4.2 Measuring science production

Quantity. The quantity measure is the number of publications produced in a given year by a scientist. This measure includes all publications contained in the WoS data base in the field “computer science”, both journal articles and conference proceedings. We only exclude publications in non peer-reviewed magazines (such as Datamation and Dr Dobbs journal).  

Quality. We create several measures reflecting publication quality. The first measure is the citation-weighted number of publications. This is simply the sum of the citations received by all publications of a scientist in a given year. The number of citations is defined as the number of citations received from all other WoS publications which were published up to 2016, the year of extraction of the data. We are aware of the fact that older articles receive more citations than younger citations. This should not bias our results as i) we use a control group with a similar distribution of publications over time and ii) we include year fixed-effects. Citation data is however highly skewed: in our sample the 3% most cited publications receive 50% of all citations. A consequence of such skewed variable is that a few

---

10 The complete list of excluded journals is: Datamation, Byte, Dr Dobbs Journal, Computer Design, Sharp Technical Journal and Hewlett-Packard Journal. The excluded journals are U.S.-centered and most of them stop publishing around 1998. Therefore including them would induce a positive bias in magnitude on our treatment effect. In practice, including these journals does not affect the estimates since they are not the core targets of university scientists. The estimates including these journal end up being similar to the main results up to the 3-digit level.
scientist will have an excessive weight on the estimates. To counterbalance this effect, we complement this citations measure with alternative quality measures.

The second quality measure is therefore the number of top cited publications, \( TOP5\% \). To construct this variable, we consider all worldwide publications in computer science in a given year, to define the publications in the 95th percentile of the citation distribution. For each scientist we count the number of such top 5\% publications per year.

The third quality measure, \( JIF10\% \), is the number of publications in highly recognized journals and conference proceedings, such as the ACM Computing Surveys or the IEEE Conference on Computer Vision and Pattern Recognition. To identify top publication outlets, we use the journal impact factors (JIF) from Scopus for the year 2015.\(^{11}\) In total, 13\% of the publications appeared to be published in a top 10\% JIF outlet.

### 4.3 Descriptive statistics

Table 1 reports some descriptive statistics for U.S. and European software computer scientists. Disregarding any temporal effect, these statistics depict two samples looking very similar in terms of scientific output. The yearly output of the average scientist is 0.58 articles and 0.38 conference proceedings per year. Taking publication quality into account a gap between the U.S. and European scientists opens up. U.S. scientists have almost twice the number of citation-weighted publications of European scientists, but with a higher variance too. The pattern is similar with regard to the two alternative quality measures, with U.S. scientists producing on average 0.25 publications in top journal/conference proceedings per year as opposed to 0.16 for European computer scientists. These numbers are respectively 0.11 and 0.068 for the yearly production of top cited publication. For both groups patenting is a rare event as their ninth decile in patents per year is 0. However, the patenting rate is almost four times higher for U.S. scientists.

\(^{11}\) About 53\% of WoS publications’ could not be matched to the Scopus database, the main causes being i) the change of outlet names over time and ii) the presence of many unreferenced conference proceedings. We, however, ensured that the most cited journals/conferences could be matched to the Scopus database by manually matching them and correcting the name changes.
Table 1: Descriptive statistics of the production of U.S. and European university software scientists – across all years (1989–2004).

<table>
<thead>
<tr>
<th>Variables</th>
<th>U.S. scientists</th>
<th>European scientists</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Min</td>
<td>Median</td>
</tr>
<tr>
<td>Yearly # Publications</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Yearly # Articles</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Yearly # Proceedings</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Yearly # of citation-weighted publications</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Yearly # Publications in top 10% JIF journal/conference</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Yearly # top 5% cited publication</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Yearly # Patents</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td># Scientist-year</td>
<td>50,191</td>
<td></td>
</tr>
<tr>
<td># Scientists</td>
<td>3696</td>
<td></td>
</tr>
</tbody>
</table>

Notes: The data consists of active computer scientists working in software. An active computer scientist is defined by having at least two publication before 1996 (with one before 1994), and at least one publication in the period 1997-2004. Observations correspond to scientist × year. Source: authors’ own calculations based on publication data from Web of Science.
5 Descriptive evidence

Before we move to the multivariate analysis, we provide some descriptive evidence depicting the evolution of publications over time for the treated group of U.S. software scientists. Panel A of Figure 1 compares the scientific output of U.S. and European software scientists. It appears that, before 1996, the trends for both European and U.S. scientists are similar. After the introduction of software patents, however, the trends differ. European scientists experience an upwards trend while the U.S. scientists face a downwards trend. This simple graphical analysis suggests that the policy change had a clear negative effect on U.S. publication volume.

In order to be sure that the pattern illustrated in Panel A is software specific, Panel B of Figure 1 shows the same figure for a placebo group of university scientists working on hardware related topics. Hardware scientists are a valid placebo group since they work in a subfield within computer science, come from the same institutions, but were not affected by the law change since they could already patent their hardware inventions since the Bayh Dole Act of 1980. We identify a total of 4,158 hardware computer scientists based on the WoS subfield “Hardware & Architecture”, 2,347 of which are affiliated with U.S. and 1,811 with European universities. Panel B shows that for hardware scientists, the publication outcome curves of U.S. and European scientists are roughly parallel across the entire period. This evidence confirms that the drop in publications is specific to software scientists, which strongly speaks in favor of a causal relationship between the introduction of software patent rights and the decrease in publications.

Further, Panel A suggests that software scientists reallocate part of their time from scientific to commercialization activities. This is in line with Figure 2 which depicts the patent outcomes of U.S. university software scientists in our sample over time. The patenting rate is around 0.02 patent per year before 1996, while it greatly increases afterwards now hovering around 0.035.
Figure 1: Publication trends of U.S. and European scientists.
Notes: Sample of U.S. and European university software scientists. Each point represents the average yearly production which was smoothed using a 3 years window \(y_{t}^{yw} = (y_{t-1} + y_t + y_{t+1})/3\).
Evolution of U.S. university software scientists patenting

Figure 2: Patenting trends of U.S. software scientists.

6 Econometric approach and empirical results

6.1 Effect on scientific output

6.1.1 Econometric approach and identification

Our analysis considers difference-in-difference (DiD) regressions where the introduction of the new patentability guidelines defines the pre- and post-change period. The effect of the legal change can be obtained by the following regression:

\[ E(y_{it}) = \exp(\alpha_i + \gamma_t + a \times Treat_i + b \times Treat_i \times Post_t), \]  

where \( y_{it} \) represents the (quality-weighted) outcome of scientist \( i \) in year \( t \), \( \alpha_i \) are scientist fixed-effects, \( \gamma_t \) are time fixed effects, \( Treat_i \) is a dummy taking value 1 if \( i \) is a U.S. scientist and \( Post_t \equiv 1 \{ t > 1996 \} \) is a dummy variable identifying the post-shock period. The coefficient of interest is \( b \). It represents the total impact of the legal change on scientists’
productivity over the post-shock period. Since the outcome variables are count variables, we estimate Equation (1) using Poisson models (Santos Silva and Tenreyro, 2006). Poisson estimations provide consistent estimates as long as the conditional mean is properly specified. Further, we cluster the standard-errors at the scientist level in order to avoid the issues raised by Bertrand et al. (2004) due to possible serial correlation.

In addition, we estimate the following equation to obtain yearly treatment effects:

\[
E(y_{it}) = \exp\left(\alpha_i + \gamma_t + \sum_{t'=1989, t' \neq 1995}^{2004} b_{t'} \times Treat_i \times 1_{t' = t}\right).
\] (2)

In this setup, we estimate the effect of the treatment for each of the sample years, i.e. each \(b_t\). We use the year before the policy change, 1995, as a reference category. If the policy implies a change in the science production of U.S. scientists, there should be a shift in the coefficients \(b_t\) for the period after 1995. This model allows to have a clear visual representation of the policy change over time.

Note that we do not focus on the direct involvement of scientists in patenting as has been done in previous studies (e.g. Azoulay et al., 2009; Czarnitzki et al., 2009). We do so since the introduction of software patent rights represents a regime shift that affects all scientists with regard to their career requirement, outcome expectations and interest for the TTO, independently on whether an individual scientist undertakes commercialization efforts or not and whether the scientist is successful at patenting or not. The number of patenting software scientists corresponds to a very selected group of scientists that engage in commercialization efforts, focus on patenting and are successful at taking out patents. The regime change however affects all scientists. Therefore focusing only on patenting scientists would lead to a very incomplete picture of the consequence of the change: we would miss all the scientists that invest time in obtaining a patent but fail to obtain it eventually.

6.1.2 Effects on publication volume and quality-weighted publication volume

Table 2 reports the regression results. Column 1 displays the estimation of Equation (1), in which the overall effect of the regime shift is estimated without time dummies and scientist
fixed effects. We find a significant negative effect of -0.20, meaning that the IP law change reduced the production of journal publications and publications in conference proceedings of U.S. scientists by 18% (i.e. $100 \times [1 - \exp(-0.20)]$) as compared to European scientists.

In Column 2, we include time fixed effects and, in Column 3, we add scientist fixed effects. These inclusions do not change the magnitude of the estimate.

Turning to publication quality, Table 3 reports the fixed effects DiD estimates of the three quality variables. The estimated coefficients for $b$ are equally negative, with a similar order of magnitude of -0.22 for the citation-weighted number of publications, -0.24 for the number of articles in top JIF journals and -0.30 for the number of top cited publications.

To get an idea of the temporal dynamics, yearly treatment effects for the quantity and the three quality variables are reported in Figure 3. The estimation follows Equation (2) and includes both year and scientist fixed-effects. The upper left panel represents the number

---

### Table 2: Effect of the legal change on scientists’ production.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Column: (1)</th>
<th>(2)</th>
<th>(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-0.0915***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0118)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treat</td>
<td>0.0881***</td>
<td>0.0849***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0183)</td>
<td>(0.0183)</td>
<td></td>
</tr>
<tr>
<td>Post</td>
<td>0.0876***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0174)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treat $\times$ Post</td>
<td>-0.2026***</td>
<td>-0.1994***</td>
<td>-0.2018***</td>
</tr>
<tr>
<td></td>
<td>(0.0268)</td>
<td>(0.0267)</td>
<td>(0.0272)</td>
</tr>
</tbody>
</table>

**Fixed-Effects**
- Year: No, Yes, Yes
- Scientist: No, Yes, Yes

**Fit statistics**
- Observations: 109,029, 109,029, 109,029
- Adj-pseudo $R^2$: 0.00084, 0.00374, 0.16729
- Log-Likelihood: -163,630.16, -163,140.17, -128,224.22

*Clustered (Scientist) standard-errors in parenthesis.*

*Signif Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: The coefficients correspond to maximum likelihood Poisson estimates. The sample consists of active computer scientists working in software. An active computer scientist is defined by having at least two publications before 1996 (with one before 1994), and at least one publication in the period 1997-2004. Observations correspond to scientist $\times$ year.

Source: authors' own calculations based on publication data from Web of Science.
Table 3: Effect of the legal change on scientists’ quality-weighted production.

<table>
<thead>
<tr>
<th>Column:</th>
<th>Yearly # of Citations-Weighted Pub.</th>
<th>Yearly # of Top 10% Ranked Articles (JIF)</th>
<th>Yearly # of Top 5% Cited Articles (Worldwide)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variables</td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
</tr>
<tr>
<td>Treat \times Post</td>
<td>-0.2183*** (0.0765)</td>
<td>-0.2397*** (0.0425)</td>
<td>-0.3001*** (0.0543)</td>
</tr>
</tbody>
</table>

Fixed-Effects

<table>
<thead>
<tr>
<th></th>
<th>Year</th>
<th>Scientist</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Fit statistics</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>106,907</td>
<td>71,384</td>
<td>46,649</td>
<td></td>
</tr>
<tr>
<td># Scientist</td>
<td>7,959</td>
<td>5,255</td>
<td>3,406</td>
<td></td>
</tr>
<tr>
<td>Adj-pseudo $R^2$</td>
<td>0.47799</td>
<td>0.08356</td>
<td>0.01702</td>
<td></td>
</tr>
<tr>
<td>Log-Likelihood</td>
<td>-2,120,336.89</td>
<td>-43,603.66</td>
<td>-22,089.23</td>
<td></td>
</tr>
</tbody>
</table>

Clustered standard-errors in parenthesis. Signif Codes: ***: 0.01, **: 0.05, *: 0.1

Notes: The coefficients correspond to maximum likelihood Poisson estimates. The sample consists of active computer scientists working in software. An active computer scientist is defined by having at least two publications before 1996 (with one before 1994), and at least one publication in the period 1997-2004. Observations correspond to scientist \times year.

Although the same sample is used across all regressions, the number of observations vary because, due to the Poisson fixed-effects setup, all scientists whose dependent variable is equal to 0 across all periods are dropped.

Source: authors’ own calculations based on publication data from Web of Science.

of publications for which the pattern is the clearest. We observe that before the shock, the estimates fluctuate around 0. After 1996, the coefficients become strongly negative. As could be expected, a few years elapsed between the introduction of software patent rights in 1996 and a sharp drop in publications. From 1999 onwards, the estimated effect fluctuates around -0.3 which hints at a long-term decrease of publications of 25%.

Regarding the citations-weighted publication measure, before the shock the coefficients hover around 0.5, while a few years after the shock the estimated coefficients approach 0. This shows a decrease of the U.S. scientists’ citation advantage due to the IP law change. The pattern is similar for the number of top cited publications. For the number of publications in top JIF journals, there is no significant treatment effect before the IP law change, while the treatment becomes significantly negative afterwards.

The results presented in this section are in line with the model of section 3: the introduction of software patents reduced significantly the propensity of U.S. software scientists to
Figure 3: Yearly treatment effects.
publish. The result holds for both pure and quality-adjusted quantity measures.

6.1.3 Scientist heterogeneity

According to the model of section 3, we expect that the change in incentives would be highest for the scientists having the lowest publication ability before the IP law change. We expect that those scientists have the highest propensity to switch to commercialization activities. To test this hypothesis, we approximate the publication ability of the scientists’ by citation-weighted publications. We split the sample into five groups of scientists, based on their positions in the ex ante distribution of the average number of citations per publication. Reflecting the skewness of the citation distribution, the first three groups are the first three quartiles, the fourth group is composed of scientists in the [75;90] percentile while the last group is formed of the most cited scientists ([90;100] percentile). The distribution for the two regions is given in Table 4. It appears that, except for the top group, there are stark differences between European and U.S. scientists, the latter having almost twice the number of citations along the distribution. This is why we form the groups of U.S. and European scientists according to the the distribution of their respective region.

Table 4: Distribution of average citations per article before 1996, separated between European and U.S. scientists.

<table>
<thead>
<tr>
<th>Percentile:</th>
<th>Ex ante average # of citations per publication</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0%</td>
</tr>
<tr>
<td>Europe</td>
<td>0</td>
</tr>
<tr>
<td>U.S.</td>
<td>0.25</td>
</tr>
</tbody>
</table>

We estimate DiD models with scientist and time fixed effects for the five groups. The results (coefficient estimates and standard-errors) are given in Figure 4 where the dependent variable is the number of publications. In line with the model predictions, we observe that the effect is largest for the first, followed by the second quartile. The associated coefficient estimates are -0.32 and -0.30 respectively. We note that the effect size of the law change decreases along the quality distribution, reaching a non-significant coefficient estimate of
ATE on US scientists number of publications
The sample is split by ex ante citation level

Estimate and 95% CI of Average Treatment Effect
Dep. var. is: # of publications

-0.4  -0.3  -0.2  -0.1  0.0
Q1  Q2  Q3  [75%; 90%] top 10%

Figure 4: Consequence of software patent introduction on the production of U.S. scientists. Estimates from several regressions where the sample is split by ex ante (1989-1996) average number of citations received per publication.

-0.06 for the highly cited scientists.

Following our model, we expect that the reason for scientists in the lowest quartile to experience the largest publication decrease is the relatively largest increase in commercialization efforts. Figure 5 where we approximate commercialization efforts by patent output supports this mechanism. The average number of patents per year produced by the low publication ability scientists rises from 0.012 before 1996 to 0.029 afterwards. This is the highest increase across all citation categories both in absolute value and in relative terms (+139%).

The graph further illustrates a positive correlation (especially \textit{ex ante}) between the position in the citation distribution and the number of patents produced. While our empirical framework only considers potential substitution between publications and patents, introducing complementarity between the two activities would not alter our empirical results. The
The total negative effect of the IP law change means that the substitution effect trumps the benefits of a potentially present complementarity effect.

### 6.2 Quantifying the consequences

From Table 1 we find an estimate of -0.20. To evaluate the counterfactual scientific output of U.S. scientists, had software patents not been introduced, we compute the total number of publications of the software scientists of our sample in the period 1997-2004, \( \text{prod}_{US}^{1997-2004} \).

The counterfactual situation can be written as \( \frac{\text{prod}_{US}^{1997-2004}}{\exp(-0.20)} \). We find a total loss of 5,246 publications (95% confidence interval: [3743;6830]). This drop is significant: MIT, for instance, had 3,735 publications in computer science between 1997 and 2004. It represents about 1% of the worldwide WoS computer science publications in that period.\(^{12}\)

Using the same logic, we find a loss of 1,693 publications in top 10% JIF journals and of

\(^{12}\)It is worth mentioning that this is a very conservative estimate, since we applied the counterfactual only to the U.S. computer scientists that: i) are software scientists and ii) have at least one publication before 1994. Therefore in computing the counterfactual, we neglect the effect the shock had on scientists entering after 1994.
1,076 publications that are in the top 5% most cited worldwide.

Turning to patents, we use estimates from this simple model with fixed-effect and time trend:

\[ E(y_{it}) = \exp(\alpha_i + \beta \times Post_t + \gamma \ln Trend_t), \]

where index \( i \) represents U.S. university software scientists. In the absence of a good counterfactual situation regarding patenting behavior, \( \beta \) provides an approximation of the gain in patents due to the law change. From the estimates, reported in Appendix C, we obtain a gain of 320 patents. The two results put together, this leads to a total “price” of 16 publications per patent or 3.35 top cited publication per patent.

### 6.3 Further analysis: Switching publication patterns

In this section, we investigate whether the introduction of software patents changed the publication pattern of U.S. scientists towards more applied contributions. Computer scientists publish in academic journals and in conference proceedings, whereby the content of conference proceedings tends to be more applied than the content of articles in traditional scientific journals. In words of the Computer Research Association: “experimental research is at variance with conventional academic publication tradition” and “experimentalists [prefer] conference publications” (Patterson et al., 1999, page A). If scientists devote more time to projects of commercial interest after the introduction of patent rights for software inventions we shall expect a larger drop in the production of journal articles than in conference proceedings.

We test this hypothesis using DiD estimation to compare the drop in the number of articles as compared to the drop in the number of proceedings. Figure 6 reports the estimated treatment effects for both dependent variables. We observe that across all five groups of scientists along their productivity distribution, the estimated treatment effect for the number of proceedings is lower in magnitude than the one for journal articles. As expected, the largest gap between the coefficients of proceedings and articles is for the lowest quartile of the citation distribution. This group suffers a decrease of 38% of their number of journal articles produced while this number is only 20% for the number of proceedings. The coefficient for the change
Figure 6: Estimate and 95 confidence intervals of average treatment effects on the number of articles and proceedings produced. Sample split by ex ante average number of citation per publication.

in proceedings is even almost not-significant for the top half of the distribution.

7 Discussion

This paper investigates the implications of the introduction of patent rights on university science. The introduction of formal IP rights is likely to change university scientists’ incentives to contribute to public science with the result that patents crowd out scientific publications.

To understand the impact of the change in IPR regime, we model scientists' behavior with a simple model of scientific production. The law change facilitates patenting and thus, through a substitution effect, scientists reduce the time devoted to publications. Further, the most affected scientists are the ones with the lowest scientific productivity or utility.
from publications. They face the highest opportunity costs of engaging in commercialization activities and are therefore more prone to switching to patenting.

We evidence a drop in publications following the introduction of software patents, this drop hovers around 20% both for volume and various quality measures. We are confident that our estimates are causal since U.S. hardware computer scientists, working in a close yet unaffected field,\textsuperscript{13} did not alter their publication pattern after the change. Then, in line with the model’s prediction, we show that scientists who suffer the biggest drop in publications are the ones having the lowest level of citation per publication \textit{ex ante}. Scientists in the lowest citation quartile suffer a 27% drop in publications, while the ones in the top 10% quantile are not significantly affected. Further, we also notice that they shift their publication pattern towards more applied research since their number of journal articles is reduced by 38% while the number of conference proceedings, more applied in nature, falls only from 20%. The consequence of the shock is heterogenous and overall the decrease is more important for journal articles than for conference proceedings suggesting a change in content.

The cost of patent rights on science is not trivial: for each new patent due to the law change, there is a loss of 16 publications or 3.3 top cited publications. Whether the “price” of the new patents is high or low depends on the university’s strategy. The modern university that emphasizes technology transfer as its third mission (Etzkowitz and Leydesdorff, 2000) is certainly interested in trading a certain amount of basic research for commercialization efforts of its faculty. From a policy perspective, this negative effect should be factored in when we assess the overall benefit/cost of the patent legislation on the economy.

The here presented evidence makes an important contribution to the technology transfer policy evaluation literature (e.g. Henderson et al., 1998; Sampat et al., 2003; Mowery and Sampat, 2005; Czarnitzki et al., 2015, 2016; Hvide and Jones, 2018). The analyzed setting operates the same mechanism as Bayh Dole Act-type of policies that strengthen universities IP ownership rights. The advantage of our scenario is that the introduction of software patents was not a response to unsatisfactory technology transfer, nor was it accompanied by other technology transfer means.

\textsuperscript{13}Hardware scientists could already patent before the change.
There are multiple avenues for future research. In this paper, we do not investigate the effect scientists’ environment on their behavior. For example, one source of influence could be the patenting behavior of colleagues in other fields at same university, or the network of collaboration. Therefore, looking at the change on publication schemes conditional on the environment can lead to interesting insights. We leave these questions for future work.

References


Bundesministerium für Bildung und Forschung and Bundesministerium für Wirtschaft und Technologie, 2001. Wissen schafft märkte - aktionsprogramm der bundesregierung, berlin. 2


A Data appendix

A.1 Data

Our starting point is the Web of Science (WoS) database provided by Thomson Reuters from which we retrieve all publications in the WoS field “computer science” from 1989 to 2004. Computer science is defined by the WoS subfields Artificial Intelligence, Information Systems, Theory & Methods, Software Engineering, Interdisciplinary Applications, Hardware & Architecture, Cybernetics. The WoS database contains both publications in scientific journals and publications in conference proceedings which are frequent in computer science and used for a speedy dissemination of results (Patterson et al., 1999). Proceedings form a growing share of computer scientists’ output over time. One main feature of a conference proceeding is that its content tends to be more applied than the content of articles in traditional scientific journals. Our initial sample consisted of 655,441 unique publications. The WoS records contain various information such as authors’ names and affiliations, the subfields within computer science, the number of citations, etc.

A.2 Unique identifiers

To track scientists’ publications over time, we created unique author identifiers mapping the publications to individual authors. We employed a novel disambiguation approach developed by Doherr (2017). In a nutshell, this method uses a “Google like” search algorithm for correcting spelling issues. It uses network analysis to disambiguate namesakes where a network is created in which two articles are connected if their shared features (such as journal name, affiliation, keywords, co-author, etc) are far enough from “chance” in terms of relative probabilities. Network analysis tools are applied to create coherent clusters of articles which are confidently identified as being written by the same author. For further details we refer

---

14 There are seven different subfields in computer science. In the full sample of publications, the number of publications per subfield is as follows: Artificial Intelligence: 203,467; Information Systems: 172,991; Theory & Methods: 164,370; Software Engineering: 138,165; Interdisciplinary Applications: 137,959; Hardware & Architecture: 97,506; Cybernetics: 33,751.

15 The algorithm has been used for disambiguation for the articles Czarnitki et al. (2015), Czarnitki et al. (2016), Czarnitzki et al. (forthcoming).
In our data set which consisted of 1,889,740 author-article pairs, the algorithm identified 691,061 unique scientists. From those, we dropped all scientists that had only one publication in computer science. Those can be PhD students that left academia or scientists in related field that ended up on a scientific publication in computer science by collaboration or coincidence. This led to a sample of 199,010 unique scientists.

In the next step, we eliminated all scientists who did not have 100% of their affiliations in our sample years of software patents within the U.S. or Europe. Researchers who change continent in our time window of interest are limited. A total of 5.98% of the European scientists and 8.03% of the U.S. scientists were dropped leading to a sample of 123,509 scientists from both regions. In the next step, we kept only scientists affiliated to universities across all the years, excluding scientists employed at any point by a governmental institution or a private firm. This left us with 81,377 university computer scientists.

A.3 Assigning affiliations

In order to select university scientists, a complex procedure to clean the affiliations was implemented before. We retrieved from WoS the addresses of each scientist in a given year. An address is a formatted character string, such as “Harvard Univ, Aiken Comp Lab, Cambridge, MA 02138, USA”, from which we extract the institution name from the first item before the first comma (here “Harvard Univ”). We define an institution as a combination of an institution name and a country. WoS unfortunately does not link the addresses/institutions to the individual scientists on the publications. Instead it provides a list of all addresses per publication, as well as an indication of the address of the reprint author. Moreover, 4.2% of the publications report no address and 18% do not provide reprint author information.

In order to assign an institution to each scientist-year, we applied an algorithm consisting of several steps. We started with the simplest case of scientists that have only one publication

---

16Our definition of Europe encompasses the EU28 countries plus Norway and Switzerland.
17E.g. AT&T based in Seattle in New Jersey and AT&T based in Denver in Colorado are considered as the same institution since both are in the USA
per year which contains only one institution (which then is valid for all authors on the publication). In this case, the scientist can be unambiguously assigned to that institution for that year. We moved stepwise forward to the more complex cases involving scientists with several publications per year, each with multiple different institutions. In these cases, we infer the institution based on the previous information we obtained and on the frequency of occurrences of the institutions per author. In the end, we were able to assign an institution to 97% of the scientist-years.

**A.3.1 Details of the algorithm**

Using the reprint and regular addresses, we create an heuristic to assign each *author-year* to an unique institution. These heuristics can be split in two categories: 1) unequivocal cases, and 2) equivocal cases.

**Unequivocal cases.** We apply these three heuristics successively:

1. The reprint author information is the most reliable source of information. In a given year, if:
   
   (a) an author is at least once a reprint author
   
   (b) all her reprint institutions are the same
   
   (c) this institution is strictly included in the set of institutions of her non-reprint-author articles

   Then we assign the author-year to the reprint institution.

   This leads to an identification of 31.9% of the sample. With these author-years identified, we can take them out of the sample and update the institutions for each article. This means concretely that if an article contains *N* authors and *K* institutions, and if *N* − 1 authors are identified and attributed to *K* − 1 institutions, then we “know” that the *N*th author is affiliated to the *K*th institution. We then use this information as follows:
2. In a given year, if:

   (a) all of an author’s articles display only a unique institution (using the updated article-institution information)

   Then we assign the author-year to that institution. This methods identifies 76.9% of the sample.

3. In a given year, if,

   (a) an institution appears in all the articles of an author

   (b) that institution appears strictly more times than the other institutions

   Then we assign the author-year to that majority institution. Now 79.9% of all author-years are identified.

An illustration of these methods are given in Table A.1.

Now a scientist may change institutions in a given year, and have two articles in the same year from these two institutions: This is an equivocal case.

**Equivocal cases.** Here the simple cases described previously do not occur anymore: authors have different institutions in a given year and we cannot directly discriminate which one is really her affiliation – maybe simply because she is indeed affiliated to several institutions. Thus we proceed as follows:

1. If in a given year,

   (a) an author’s institution appears strictly more times than other institutions

   Then we assign the author-year to that institution. 83.5% of the author-years becomes identified.

Note that this case differs from the 3\(^{rd}\) case of the “unequivocal cases” because the requirement of that institution appearing in all articles of the year is not here anymore.
Table A.1: Example of unequivocal identification of author-year institutions.

**Case 1.** Assume that in 2000 Jane Doe has only two articles:

<table>
<thead>
<tr>
<th>Author</th>
<th>Article ID</th>
<th>Institution</th>
<th>Is reprint author</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jane Doe</td>
<td>1</td>
<td>HARVARD UNIV</td>
<td>Yes</td>
</tr>
<tr>
<td>Jane Doe</td>
<td>1</td>
<td>UNIV ILLINOIS</td>
<td>No</td>
</tr>
<tr>
<td>Jane Doe</td>
<td>2</td>
<td>AT&amp;T BELL LABS</td>
<td>No</td>
</tr>
<tr>
<td>Jane Doe</td>
<td>2</td>
<td>HARVARD UNIV</td>
<td>No</td>
</tr>
</tbody>
</table>

Only looking at the addresses we cannot assign her to an institution. However, as she is a reprint author in the first article, in which her institution is Harvard, and Harvard also appears in her second article, then we can assign her to Harvard.

**Case 2.** In 2000, assume that John Smith has two articles. These articles contain several institutions, but some institutions were identified to other author-years via the previous method and John Smith is the last unidentified author:

<table>
<thead>
<tr>
<th>Author</th>
<th>Article ID</th>
<th>Institution</th>
<th>Is reprint author</th>
</tr>
</thead>
<tbody>
<tr>
<td>John Smith</td>
<td>1</td>
<td>HARVARD UNIV</td>
<td>No</td>
</tr>
<tr>
<td>John Smith</td>
<td>1</td>
<td>UNIV ILLINOIS</td>
<td>No</td>
</tr>
<tr>
<td>John Smith</td>
<td>3</td>
<td>UNIV ILLINOIS</td>
<td>No</td>
</tr>
<tr>
<td>John Smith</td>
<td>3</td>
<td>UNIV CAROLINA</td>
<td>No</td>
</tr>
</tbody>
</table>

Then we assign John Smith to UNIV ILLINOIS.

**Case 3.** In 2000, Julien Dupond also has two articles but his co-authors were not identified via Case 1:

<table>
<thead>
<tr>
<th>Author</th>
<th>Article ID</th>
<th>Institution</th>
<th>Is reprint author</th>
</tr>
</thead>
<tbody>
<tr>
<td>Julien Dupond</td>
<td>4</td>
<td>ECOLE POLYTECH</td>
<td>No</td>
</tr>
<tr>
<td>Julien Dupond</td>
<td>4</td>
<td>UNIV PARIS</td>
<td>No</td>
</tr>
<tr>
<td>Julien Dupond</td>
<td>5</td>
<td>ECOLE POLYTECH</td>
<td>No</td>
</tr>
<tr>
<td>Julien Dupond</td>
<td>5</td>
<td>UNIV POLITECN MADRID</td>
<td>No</td>
</tr>
</tbody>
</table>

We assign him to ECOLE POLYTECH as it appears in all his publications and strictly more times than the other institutions.

Thus we simply consider that if an author has more publications in one institution, it is likely that it is the place where she was the most active.

Now we deal with the case where there is no majority of an institution in a given year.

2. In a given year, if,

   (a) an author’s institution has already been identified in other years via the previous methods

   (b) that institution has been assigned to other author-years a *strict* majority of times
Then we assign this author-year to this institution, leading to a 89.6% of the sample identified.

3. In a given year, if,

(a) an author’s institution also appears in other years
(b) that institution is the most frequent institution across all years

We assign the author-year to institution. We end with 90.8% of the sample being identified.

Table A.2 illustrates these cases.

**Ex post modifications.** After these two classifications are done, we correct the sequences of institutions for possible mistakes. We consider two causes of problems:

1. Timing of publications: a scientist changes institutions, but the publications timing is not appropriate.

2. Missing information in our raw data due to formatting issues or mere misreporting.

To cure these two problems, we apply a simple rule: when an institution is surrounded by two identical institutions, we replace it by the surrounding institution, as illustrated by Table A.3. We apply 21,303 such modifications (in the initial full sample of 1,208,600 scientist-years, it is a 1.7% frequency).

**Filling the gaps.** Finally when a scientist doesn’t have a publication in a given year, we recursively assign her to the identified institution of the previous year. If no previous year information is provided, we assign them recursively to the identified institution of the next year. This simple process is illustrated in Table A.4.
Table A.2: Example of equivocal identification of author-year institutions.

**Case 1.** In 2000, Jane Doe has three publications with the following institutions:

<table>
<thead>
<tr>
<th>Author</th>
<th>Article ID</th>
<th>Institution</th>
<th>Is reprint author</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jane Doe</td>
<td>1</td>
<td>HARVARD UNIV</td>
<td>No</td>
</tr>
<tr>
<td>Jane Doe</td>
<td>1</td>
<td>UNIV ILLINOIS</td>
<td>No</td>
</tr>
<tr>
<td>Jane Doe</td>
<td>2</td>
<td>AT&amp;T BELL LABS</td>
<td>No</td>
</tr>
<tr>
<td>Jane Doe</td>
<td>2</td>
<td>HARVARD UNIV</td>
<td>No</td>
</tr>
<tr>
<td>Jane Doe</td>
<td>3</td>
<td>UNIV CAROLINA</td>
<td>No</td>
</tr>
</tbody>
</table>

Although HARVARD UNIV does not appear in all publications, we assign Jane Doe to it since it appears a strict majority of times.

**Case 2.** In 2000, John Smith has two articles:

<table>
<thead>
<tr>
<th>Author</th>
<th>Article ID</th>
<th>Institution</th>
<th>Is reprint author</th>
</tr>
</thead>
<tbody>
<tr>
<td>John Smith</td>
<td>4</td>
<td>HARVARD UNIV</td>
<td>No</td>
</tr>
<tr>
<td>John Smith</td>
<td>5</td>
<td>UNIV ILLINOIS</td>
<td>No</td>
</tr>
</tbody>
</table>

Based on this information we cannot assign them to any institution. Yet, we identified John Smith’s institutions in some other years via the previous methods:

<table>
<thead>
<tr>
<th>Author</th>
<th>Year</th>
<th>Identified Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>John Smith</td>
<td>1997</td>
<td>HARVARD UNIV</td>
</tr>
<tr>
<td>John Smith</td>
<td>1999</td>
<td>HARVARD UNIV</td>
</tr>
<tr>
<td>John Smith</td>
<td>2000</td>
<td>?</td>
</tr>
<tr>
<td>John Smith</td>
<td>2002</td>
<td>UNIV ILLINOIS</td>
</tr>
</tbody>
</table>

Since HARVARD UNIV has already been identified in other years and so a strict majority of times, we assign John Smith to HARVARD UNIV in 2000.

**Case 3.** In 2000, Julien Dupond has the following publications:

<table>
<thead>
<tr>
<th>Author</th>
<th>Article ID</th>
<th>Institution</th>
<th>Is reprint author</th>
</tr>
</thead>
<tbody>
<tr>
<td>Julien Dupond</td>
<td>6</td>
<td>ECOLE POLYTECH</td>
<td>No</td>
</tr>
<tr>
<td>Julien Dupond</td>
<td>6</td>
<td>UNIV PARIS</td>
<td>No</td>
</tr>
<tr>
<td>Julien Dupond</td>
<td>7</td>
<td>ST MICROELECTRONICS</td>
<td>No</td>
</tr>
<tr>
<td>Julien Dupond</td>
<td>7</td>
<td>UNIV POLITECN MADRID</td>
<td>No</td>
</tr>
</tbody>
</table>

There are four different institutions for two publications. We cannot assign Julien Dupond to an institution with this information. Further, ECOLE POLYTECH and UNIV PARIS have been identified as Julien Dupond’s institutions in other years but an equal number of times.

To discriminate, we then look at his publication records across all years, we have the following count:

<table>
<thead>
<tr>
<th>Institution</th>
<th># of appearances (all years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ECOLE POLYTECH</td>
<td>6</td>
</tr>
<tr>
<td>UNIV PARIS</td>
<td>3</td>
</tr>
<tr>
<td>ST MICROELECTRONICS</td>
<td>1</td>
</tr>
<tr>
<td>UNIV POLITECN MADRID</td>
<td>2</td>
</tr>
</tbody>
</table>

Since across Julien Dupond’s publications, ECOLE POLYTECH is the most frequent, we assign Julien Dupond to ECOLE POLYTECH in 2000.
Table A.3: Corrections of institutional spells.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Case 1</td>
<td>A</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>Case 1 Corrected</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>B</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>Case 2</td>
<td>A</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>Case 2 Corrected</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
</tbody>
</table>

Notes: A and B represent two different institutions.  
In case 1, it is very likely that the scientist was in institution A in 1993 (and left to institution B during 1993) and is in institution B from 1994 on.  
In case 2, it is very likely that the institution reported in 1993 is a mistake (usually it is the institution of a coauthor). For example, in the article “Simultaneous Fitting Of Several Planes To Point Sets Using Neural Networks” published in “Computer Vision Graphics Image Processing” in 1990, the author Behrooz Kamgar-Parsi is affiliated to “Computer Vision Laboratory, Center for Automation Research, University of Maryland, College Park, Maryland 20742, USA”. However, in our data, the only two addresses showing up in that publication record are: “George Mason Univ, Dept Comp Sci, Fairfax, VA 22030”, and “USN, Res Lab, Washington, DC 20375” which are the two institutions of his two co-authors.

Table A.4: Filling the gaps in institutional spells.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Case 1 Filled</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td></td>
</tr>
<tr>
<td>Case 2 Filled</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td></td>
</tr>
</tbody>
</table>

Notes: A and B represent two different institutions, the data is the same as in Table A.3.  
The years in bold are the ones where there is no available information on the scientists (e.g. a year without publication).

A.3.2 Identification of universities

To identify which institution is a university, we apply a pattern matching on the institution name. Thanks to manual identification, the following classification guarantees that no institution with more than 150 publications is left unidentified.

An institution is considered as a university if:

- it contains one of the following words: caltech, coll, cuny, ecole, ens, epfl, eth, fac, faculty, harvard, kth, mit, nyu, politecn, polytech, sch, school, scuola, stanford, suny, supelec, tu, univ, university, upmc,
or it contains the following patterns: inst technol, inst sci technol, virginia tech,

or it is equal to: ensieg, enst, enst bretagne, georgia tech, iit, imag, imag lab grenoble,
inst eurecom, itesm, lirmm, rhein westfal th aachen, telecom paris, th darmstadt, tima
lab, ucl, ufrgs, ufrj, umist, unicamp, verimag.

In total, there are 67,305 different institution names. We identify 14,705 of them as universities. Looking at the 859,862 publication-institution pairs, 74% are universities.

A.3.3 Further sample restrictions

Since the field of computer science is broad, it occurs frequently that scientists coming from other university departments publish in this field. We therefore only kept scientist with a minimum of three computer science publications in three different years to qualify for our sample. Two of those publications have to be from before 1996. Using other thresholds does not alter the findings (see Appendix B). Scientists working on hardware related topics have been excluded from the treatment group as well as scientists that were switching affiliations between the U.S. and Europe. Our final sample consists of 8,133 computer scientists, among them 4,437 European scientists, working on software related topics. In total, these scientists produced 86,756 unique publications.

A.4 Patent information

Although the main focus of the paper is on scientists’ publication output, we complement each scientist’s information with her patent records. We extract all granted patents applied at the USPTO between 1989 to 2004 from the Patstat database. We identify inventors using the same disambiguation algorithm described in Section A.2. Then we match the patent data information to the publication data using the disambiguated inventor/scientist names as well as all other complementary information, such as the institutions names contained in the patent/publication documents. We find a total of 680 university software scientists having at least one patent over the period (460 U.S. scientists).
The application of the disambiguation algorithm provided us with three different career realms: the U.S. inventor, the European inventor and the author publishing in the field of computer science. The distinction between the inventor careers stems from the separation of the patent authorities. The term career refers to a sequence of documents produced by an individual with a high probability. A career is labeled with a name, it has an entry and an exit point defined by the first and last published document (patent or publication) and is associated with affiliations respectively applicants accrued on its passage. Of course, there are more properties to a career, but given out specific setup, we refrain from imposing additional restrictions to avoid a bias towards state dependency, i.e. ignoring job changes or the diffusion of the research field engendered by participation in larger projects.

As a first step, for the three realms, we create a table containing all name and country code combinations encountered in the respective data. The country codes stem from the associated affiliations or applicants. By applying the search tool “SearchEngine”\textsuperscript{18} configured for n-grams (see Doherr, 2017), we link the publication names with the two patent name tables. Every author-name-country combination produces a list of inventor name candidates with the same country and a high similarity to the author name. Linking merely by the name and country seems to be a recipe for disaster given the high degree of homonymy in our data exacerbated by the fact, that the publication data does only provide initials instead of proper first names, defining the lowest common denominator for our matching effort. Fortunately, the author names come attached with additional meta information transferred from the disambiguation routine. We can directly exploit the estimated number of namesakes to assess the matching capabilities of a name. Further, by observing the number of careers associated with a name for the three realms, we can exclude one-to-many or many-to-many career intersections, implicitly solving the issue of having multiple matched inventor name variants for an author name. As long as the variants are accrued within one inventor career, the one-on-one tenet is not violated. We can relax this tenet, by including affiliation to applicant linkage. In a separate step, we matched the affiliations to patent applicants using the “SearchEngine” configured for frequency-based heuristics to filter filler words like

\textsuperscript{18}The tool can be downloaded at ftp://ftp.zew.de/tools/searchengine.zip.
legal forms. This linkage introduces additional criteria potentially separating multiple career assignments into one-on-one matches. Of course, not every one-on-one assignment represents a distinctive career switch from author to inventor or vice-versa, but the inclusion of the estimated namesake count and the juxtaposition of the respective career periods represented by the entry and exit points allow for fine-tuning of recall vs. precision. A career switch of an author with a high namesake count, overlapping career periods or an implausible stretch of inactivity is deemed to be dropped during an explorative phase of sample adjustments. Under the assumption of conditional independence of names to career paths, we are confident to introduce not any bias due to our arbitrary decisions.

**B Estimates for different samples**

**Active computer scientists.** The results of this paper are based on active computer scientists. We defined an active computer scientist as a scientist with at least two publications in two different years before 1996 and at least one publication after 1996.

We now make the definition of active scientists vary in terms of production before and after the shock. First, in terms of production before the shock, we consider scientists either active in two, or more restrictively, in three different years. Second, after the shock, we consider scientists: a) without restriction, or having at least b) one, or c) two active years.

Table B.5 reports the results of the 6 estimations. All coefficients are in range of our main result (which is reported in column 2). It is worth noting that the estimates are even larger in magnitude when we don’t make restrictions for the production after the shock.

**Hardware scientists definition.** This section reports estimates for varying definitions of hardware computer scientists. To identify hardware computer scientists, we relied on the information contained in the subfield of each journal. We define a hardware publication as a publication containing the subfield “Hardware & Architecture” and not containing the subfield “Software Engineering”. We then defined a software scientists as someone having no hardware publication.
Table B.5: Main estimates for varying definitions of active scientists.

<table>
<thead>
<tr>
<th># of Active Years</th>
<th>2 Active Years Before 1996</th>
<th>3 Active Years Before 1996</th>
</tr>
</thead>
<tbody>
<tr>
<td>After 1996</td>
<td>≥ 0</td>
<td>≥ 1</td>
</tr>
<tr>
<td>Dep. Variable:</td>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td>Model:</td>
<td>Yearly # Publications</td>
<td></td>
</tr>
<tr>
<td>Treat×Post</td>
<td>-0.2326***</td>
<td>-0.2018***</td>
</tr>
<tr>
<td></td>
<td>(0.0313)</td>
<td>(0.0272)</td>
</tr>
</tbody>
</table>

| Fixed-Effects     |                             |                             |                             |                             |
| Scientist         | Yes                         | Yes                         | Yes                         | Yes                         | Yes                         | Yes                         |
| Year              | Yes                         | Yes                         | Yes                         | Yes                         | Yes                         | Yes                         |

| Fit statistics    |                             |                             |                             |                             |
| Observations      | 196,683                     | 109,029                     | 79,672                      | 90,954                      | 66,324                      | 53,379                      |
| # Scientist       | 14,986                      | 8,133                       | 5,894                       | 6,537                       | 4,744                       | 3,804                       |
| Adj-pseudo $R^2$  | 0.20154                     | 0.16729                     | 0.15652                     | 0.19125                     | 0.16709                     | 0.15949                     |
| Log-Likelihood    | -175,498.22                 | -128,224.22                 | -104,755.01                 | -105,716.59                 | -88,043.85                  | -75,842.28                  |

Clustered (Scientist) standard-errors in parenthesis. Signif Codes: ***: 0.01, **: 0.05, *: 0.1

NOTES: Fixed-effects Poisson estimations. An active year is a year with a publication.
Source: authors’ own calculations based on publication data from Web of Science.

We now replicate the main estimation, varying the threshold to qualify as a hardware scientist. Table B.6 reports the estimates for thresholds ranging from 0 (like in the paper) to 3. The estimates for different thresholds are all in line with the main estimates.

Table B.6: Varying the threshold defining software scientists.

<table>
<thead>
<tr>
<th>Total Hardware Publications (1989-2004)</th>
<th>≤ 0</th>
<th>≤ 1</th>
<th>≤ 2</th>
<th>≤ 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dependent Variable:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yearly # Publications</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model:</td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
</tr>
<tr>
<td>Variables</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treat×Post</td>
<td>-0.2018***</td>
<td>-0.1804***</td>
<td>-0.177***</td>
<td>-0.1814***</td>
</tr>
<tr>
<td></td>
<td>(0.0272)</td>
<td>(0.0235)</td>
<td>(0.0224)</td>
<td>(0.0216)</td>
</tr>
</tbody>
</table>

| Fixed-Effects                          |     |     |     |     |
| Scientist                              | Yes | Yes | Yes | Yes |
| Year                                  | Yes | Yes | Yes | Yes |

| Fit statistics                         |     |     |     |     |
| Observations                           | 109,029                     | 129,253                     | 139,506                     | 146,132                     |
| # Scientist                            | 8,133                       | 9,636                       | 10,399                      | 10,888                       |
| Adj-pseudo $R^2$                       | 0.16729                     | 0.17377                     | 0.18137                     | 0.1863                       |
| Log-Likelihood                         | -128,224.22                 | -156,168.70                 | -171,561.31                 | -181,641.21                 |

Clustered (Scientist) standard-errors in parenthesis. Signif Codes: ***: 0.01, **: 0.05, *: 0.1

NOTES: Fixed-effects Poisson estimations.
Source: authors’ own calculations based on publication data from Web of Science.
C Quantification estimates

The estimates used to quantify the gain in patents are reported in Table C.7.

Table C.7: Estimation of the increase in patent production for U.S. university software scientists.

<table>
<thead>
<tr>
<th>Dependent Variable: Yearly # Patents</th>
<th>(1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variables</td>
<td></td>
</tr>
<tr>
<td>Post</td>
<td>0.3731***</td>
</tr>
<tr>
<td></td>
<td>(0.1366)</td>
</tr>
<tr>
<td>Time Trend (log)</td>
<td>0.0899</td>
</tr>
<tr>
<td></td>
<td>(0.1199)</td>
</tr>
<tr>
<td>Fixed-Effects</td>
<td></td>
</tr>
<tr>
<td>Scientist</td>
<td>Yes</td>
</tr>
<tr>
<td>Fit statistics</td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>6,236</td>
</tr>
<tr>
<td># Scientist</td>
<td>460</td>
</tr>
<tr>
<td>Adj-pseudo $R^2$</td>
<td>0.10272</td>
</tr>
<tr>
<td>Log-Likelihood</td>
<td>-3,194.30</td>
</tr>
</tbody>
</table>

Clustered (Scientist) standard-errors in parenthesis.
Signif Codes: ***: 0.01, **: 0.05, *: 0.1