Online Appendix

"Who Profits from Patents?

Rent-Sharing at Innovative Firms"

Patrick Kline	Neviana Petkova	Heidi Williams	Owen Zidar
UC-Berkeley	US Department of Treasury	MIT	Princeton

February 20, 2019

A Appendix: Data

A.1 Description of patent data

Our patent data build draws on several sources. Three identification numbers are relevant when using these datasets. First, publication numbers are unique identifiers assigned to published patent applications. Second, application numbers are unique identifiers assigned to patent applications that in practice are quite similar to publication numbers, but sometimes one application number is associated with multiple publication numbers. Finally, patent grant numbers are unique identifiers assigned to granted patents. Note that one patent application number can be associated with more than one granted patent.

Traditionally, unsuccessful patent applications were not published by the USPTO. However, as part of the American Inventors Protection Act of 1999, the vast majority of patent applications filed in the US on or after 29 November 2000 are published eighteen months after the filing date. There are two exceptions. First, applications granted or abandoned before eighteen months do not appear in this sample unless the applicant chooses to ask for early publication. Lemley and Sampat (2008) estimate that about 17 percent of patents are granted before eighteen months, of which about half (46 percent) are published pre-patent grant. Second, applications pending more than eighteen months can "opt out" of publication if they do not have corresponding foreign applications, or if they have corresponding foreign applications but also have priority dates predating the effective date of the law requiring publication (Lemley, and Sampat 2008).[1]

1. Census of published USPTO patent applications. We observe the census of published (accepted and rejected) patent applications published by the US Patent and Trademark Office (USPTO). Our source for this data is a set of bulk XML files hosted by Google.[2] The underlying XML file formats were often inconsistent across years, so in the process of parsing these XML files to flat files we attempted to validate the data against other USPTO administrative data wherever possible. These records are at the publication number level.

2. Census of granted USPTO patents. For the published USPTO patent applications in our data, we wish to observe which of those applications were granted patents. Our source for this data is a set of bulk XML files hosted by Google.[3] As with the published USPTO patent applications data, the underlying XML file formats were often inconsistent across years, so in the process of parsing these XML files to flat files we attempted to validate the data against other USPTO administrative data wherever possible. As one specific example, even though patent numbers uniquely identify patent grants, there are twenty-one patent numbers in this data that appear in the data twice with different grant dates. Checking these patent numbers on the USPTO's online Patent Full Text (PatFT) database reveals that in each of these cases, the duplicated patent number with the earlier grant date is correct.[4] Accordingly, we drop the twenty-one observations with the later grant dates.

3. USPTO patent assignment records. Some of our published patent applications are missing assignee information. (Applicants are not required to submit assignee information to the USPTO at the time of application.) Based on informal conversations with individuals at the USPTO, we fill in missing assignee names to the extent possible using the USPTO Patent Assignment data. The USPTO Patent Assignment data records assignment transactions, which are legal transfers of all or part of the right, title, and interest in a patent or application from one or more existing owner to one or more recipient. The dataset is hosted on the USPTO website.[5] Each transaction is associated with a patent number, application number, and/or publica-

[1] For more details, see `http://www.uspto.gov/web/offices/pac/mpep/s1120.html` and the discussion in Lemley and Sampat (2010). Most applications not published eighteen months after filing are instead published sixty months after filing.

[2] See `http://www.google.com/googlebooks/uspto-patents-applications-biblio.html`.

[3] See `http://www.google.com/googlebooks/uspto-patents-grants-text.html`.

[4] PatFT can be accessed at `http://patft.uspto.gov/`.

[5] Available at: `https://www.uspto.gov/learning-and-resources/electronic-data-products/patent-assignment-dataset`.

tion number (wherever each is applicable). The patent assignment records include both initial assignments and re-assignments, but only the former is conceptually appropriate for our analysis since we want to measure invention ownership at the time of application. We isolate initial assignments by taking the assignment from this database with the earliest execution date. If a given assignment has more than one execution date (e.g., if the patent application is assigned to more than one entity), we use the latest execution date within that assignment as the transaction execution date. Using these initial assignments, we fill in assignee organization name as well as assignee address information where possible when these variables are missing from our published patent applications data.

4. USPTO patent document pre-grant authority files. A very small number (1,025 total) of published USPTO patent applications are "withdrawn," and these observations tend to be inconsistently reported across the various datasets we analyze. The USPTO patent document pre-grant authority files — an administrative data file hosted on the USPTO website — allows us to exclude all withdrawn applications for consistency.[6] Our versions of these files were downloaded on 24 March 2014 and are up to date as of February 2014. These records are at the publication number level.

5. USPTO PAIR records. We analyze several variables, such as the date of initial decisions, from the USPTO Patent Application Information Retrieval (PAIR) data, which we draw from an administrative dataset called the Patent Examination Research Dataset (PatEx).[7] With the exception of 264 published patent applications, these data are available for our full sample of published USPTO patent applications. These records are at the application number level.

6. Examiner art unit and pay scale data. Frakes and Wasserman (2017) generously provided us with examiner art unit and General Schedule (GS) pay scale data they received through FOIA requests. These data allow us to identify which examiners were active in each art unit in each year.

7. Thomson Innovation database. All of the databases listed above record information obtained directly from the USPTO. One measure of patent value that cannot be constructed based on the USPTO records alone is a measure of patent family size, as developed in Jonathan Putnam's dissertation (Putnam 1996). Generally stated, a patent "family" is defined as a set of patent applications filed with different patenting authorities (e.g., the US, Europe, Japan) that refer to the same invention. The key idea is that if there is a per-country cost of filing for a patent, firms will be more likely to file a patent application in multiple countries if they perceive the patent to have higher private value. Past work — starting with Putnam (1996) — has documented evidence that patent family size is correlated with other measures of patent value. The Thomson Reuters Innovation database collects non-US patent records, and hence allows for the construction of such a family size measure.[8] We purchased a subscription to the Thomson Innovation database, and exported data from the web interface on all available variables for all published USPTO patent applications. To construct our family size measure, we take the DWPI family variable available in the Thomson Innovation database (which lists family members), separate the country code from the beginning of each number (e.g., "US" in "US20010003111"), and then count the number of unique country codes in the family. These records are at the publication number level.

8. Hall, Jaffe, and Trajtenberg (2001) NBER data. Hall, Jaffe, and Trajtenberg (2001) constructed a match between US patents granted between January 1963 and December 1999 with the Compustat data. As part of that work, the authors constructed technology categories to describe the broad content area of different patents, based on categorizations of the patent technology class and subclass variables.[9] We match on these

[6] See `http://www.uspto.gov/patents/process/search/authority/`.

[7] See `http://www.uspto.gov/learning-and-resources/electronic-data-products/patent-examination-research-dataset-public` for the underlying PAIR data, see: `http://portal.uspto.gov/pair/PublicPair`.

[8] See `http://info.thomsoninnovation.com/`.

[9] See `http://www.nber.org/patents/`.

technology categories, and hand-fill the small number of cases in which classes or subclasses appear in our data but not in the crosswalk constructed by Hall and co-authors. These records are at the patent class level.

9. Kogan et al. (2017) patent value data. Kogan et al. (2017) provide their final estimates of patent value for their sample of granted patents at `https://iu.app.box.com/v/patents/`. In particular we downloaded the "patents.zip" file, which contains a linkage between USPTO patent grant numbers and the estimate of the patent value ξ. These data were downloaded on — and are accurate as of — 7 August 2016. To develop a measure of patent value at the application number level, we associate each application with its potentially numerous patent numbers. We then sum the values of ξ by application number to obtain a measure of the ex-post value of granted applications.

10. USPTO technology center data. Technology centers are groupings of examiner art units. The USPTO hosts a listing of all technology centers and associated examiner art units at `http://www.uspto.gov/patent/contact-patents/patent-technology-centers-management`. We use these groupings to examine heterogeneity in predicted patent value by area of invention in Appendix Table D.2.

A.2 Construction of patent application sample

We restrict the sample to USPTO patent applications filed on or after 29 November 2000 (the date when "rejected" applications started to be published), and ends with applications published on 31 December 2013. We impose a few additional sample restrictions:

- We exclude a very small number of "withdrawn" patent applications (1,025 total) given that these observations tend to be inconsistently reported across datasets. As noted above, the withdrawn applications were identified using the USPTO patent document pre-grant authority files.

- Six publication numbers are listed in the USPTO patent document pre-grant authority files but are not available in any of our other datasets;[10] we exclude these observations from our sample.

- Four publication numbers are missing from the Thomson Innovation database.[11] We include these observations in the sample, but they are missing data for all variables drawn from the Thomson Innovation data.

- Based on the kind code variable listed on the USPTO published patent applications,[12] we exclude a small number of patent applications that are corrections of previously published applications: corrections of published utility/plant patent applications (kind codes A9/P9; 3,156 total), and second or subsequent publications of the same patent application (kind codes A2/P4; 1,182 total). These kind codes more generally allow us to confirm that our sample does not include various types of documents: statutory invention registration documents (kind code H1), reexamination certificates (kind codes Bn/Cn/Fn for n=1-9), post grant review certificates (kind codes Jn for n=1-9), inter parties review certificates (kind codes Kn for n=1-9), or derivation certificates (kind codes On for n=1-9). Our final sample includes only utility patent applications (kind code A1; 3,597,787 total) and plant patent applications (kind code P1; 4,196 total).

Finally, there are two data inconsistencies that we have resolved as follows:

- Seven observations appear to be missing from Google's XML files of the published patent applications.[13] We were able to hand-code the required variables for these observations based on the published patent applications posted at `http://patft.uspto.gov` for all but three of these observations (specifically, publication

[10] Specifically, these publication numbers are: US20010003111; US20020011585; US20020054271; US20020084413; US20020084764; and US20020103782.

[11] Specifically, these publication numbers are: US20010020331; US20010020666; US20010021099; and US20010021102.

[12] For a summary of USPTO kind codes, see: `http://www.uspto.gov/patents/process/search/authority/kindcode.jsp`.

[13] Specifically, the missing publication numbers are: US20010020331; US20010020666; US20010021099; US20010021102; US20020020603; US20020022313; US20020085735.

numbers US20020020603; US20020022313; US20020085735). For those three observations, we hand-coded the required variables based on the information available at `http://portal.uspto.gov/pair/PublicPair/`; for these, we assumed that the appropriate correspondent addresses were those listed in the "Address and Attorney/Agent" field under Correspondence Address.

- The applications data contain 67 applications that were approved SIR (statutory invention registration) status but have the kind code "A1," instead of "H1" (as we would expect). We changed the kind code to "H1" for these applications, and they are therefore dropped from our sample.

A.3 Description of US Treasury tax files

All firm-level variables are constructed from annual business tax returns over the years 1997-2014: C-Corporations (Form 1120), S-Corporations (Form 1120S), and Partnerships (Form 1065). Worker-level variables are constructed from annual tax returns over the years 1999–2014: Employees (form W2) and contractors (form 1099).[14]

Variable Definitions

To define firm-level variables using the US Treasury files, we use the following line items from the 2010 business tax forms: 1120 for C-corporations, 1120S for S-corporations, and 1065 for partnerships. Note that the tax form line numbers can sometimes change slightly if, for example, a line is added for a new deduction.

- Revenue
  - Line 1c of Form 1120 for C-Corporations, Form 1120S for S-Corporations, and Form 1065 for partnerships. When 1c is not available, we use 1a, which is gross receipts. We replace negative revenue entries, which are very rare, with missing values.
- Total Income
  - For C-Corporations, line 11 on Form 1120. Note that this subtracts COGS from revenues and includes income from a variety of sources (e.g., dividends, royalties, capital gains, etc). For S-Corporations, line 6 on Form 1120S. For partnerships, line 8 on Form 1065.
- Total Deductions
  - For C-corporations, line 27 on Form 1120. For S-corporations, line 20 on Form 1120S. For partnerships, line 21 on Form 1065.
- Labor Compensation
  - For C-Corporations, sum of lines 12, 13, 24, and 25 on Form 1120.[15] For S-Corporations, sum of lines 7, 8, 17, and 18 for Form 1120S. For partnerships, sum of lines 9, 10, 18, and 19 on Form 1065. These lines are compensation to officers, salaries and wages, retirement plans, and employment benefit programs, respectively.[16]

[14] W2 data are not available in 1997–1999.

[15] Ideally, we could also add Schedule A line 3, which is the cost of labor on the COGS Form 1125-A, but these data are not available. However, the W2-based measure of compensation avoids this issue.

[16] For partnerships, the compensation to officers term is called "Guaranteed payments to partners."

- Value Added
  - Gross receipts minus the difference between cost of goods sold and cost of labor.
  - For C-Corporations, line 3 on Form 1120. For S-corporations, line 3 on Form 1120S. For partnerships, line 3 on Form 1065.[17]
- Profits
  - Yagan (2015) defines operating profits as revenues less Costs of Goods Sold and deductions where deductions are total deductions other than compensation to officers, interest expenses, depreciation, and domestic production activities deduction. We do not add back compensation to officers.
  - For C-Corporations, we define operating profits as the sum of lines 1c, 18, and 20, less the sum of 2 and 27 on Form 1120. We set profits to missing if 1c, 18, 20,2, and 27 are all equal to zero.
  - For S-Corporations, operating profits are the sum of lines 1c, 13, and 14 less the sum of 2 and 20 on Form 1120S.
  - For partnerships, operating profits are the sum of lines 1c, 15, and 16c less the sum of 2 and 21 on Form 1065.
- EBITD
  - EBITD is total income less total deductions (other than interest and depreciation).
  - For C-Corporations, it is the sum of lines 11, 18, and 20, less 27 on Form 1120.
  - For S-Corporations, it is the sum of lines 1c, 13, and 14 less 20 on Form 1120S.
  - For partnerships, it is the sum of lines 1c, 15, and 16c less 21 on Form 1065.
- Employment
  - Number of W2s associated with an Employer Identification Number (EIN).
- Wage bill per worker
  - Sum of W2 box 1 payments divided by number of W2s for a given EIN.
- Surplus
  - Sum of EBITD and Wage bill, which is the sum of W2 box 1 payments for a given EIN.
- Inventor earnings per inventor
  - Wage bill per worker for workers who are identified as inventors by Bell et al. (2019).
- Cohort earnings per worker
  - Wage bill per worker for workers who were employed at the firm in the year of application regardless of whether or not they stay at the firm.

[17] Line 3 is calculated as line 1c minus line 2.

- Stayer earnings per worker
  - Stayers are cohort earning per worker for the set of workers who are still at the firm.
- Leaver earnings per worker
  - Leavers are cohort earning per worker for the set of workers in the initial cohort who are no longer at the firm, i.e., are no longer receiving a W2 associated with the original firm that applied for a patent.
- Earnings Gap Q4-Q1
  - Average earnings within quartile four and quartile one of a firm's wage distribution.
- Separators
  - The number of workers who left the EIN in the previous year.
- Entrants
  - The number of workers who joined the EIN relative to the previous year.
- State
  - Uses the state from the business's filing address.
- Entity Type
  - Indicator based on tax-form filing type.
- Industry
  - NAICS codes are line 21 on Schedule K of Form 1120 for C-Corporations, line 2a Schedule B of Form 1120S for S-Corporations, and Box A of Form 1065 for partnerships.
- Active Firm
  - An active firms has non-zero and non-missing total income and non-missing total deductions.

A.3.1 Deflator to convert to 2014 USD

Table A.1: Deflator to convert to 2014 USD

Year	1 / 2014 CPI	Year	1 / 2014 CPI
1993	1.503062988	2004	1.219663817
1994	1.47180133	2005	1.181649182
1995	1.441698831	2006	1.146416065
1996	1.415894851	2007	1.116642696
1997	1.391942424	2008	1.095506864
1998	1.377006398	2009	1.08694
1999	1.357571973	2010	1.073775512
2000	1.327317133	2011	1.052064076
2001	1.297761328	2012	1.033016537
2002	1.278151458	2013	1.016449245
2003	1.253173459	2014	1

Notes: This table shows the deflators used to convert our dollar amounts from current dollars into 2014 USD. Deflators were calculated using price data from the US Bureau of Economic Analysis (BEA), National Income and Product Accounts (NIPA) Table 1.1.4: 'Price Indexes for Gross Domestic Product.' See US Bureau of Economic Analysis (2014).

A.4 Description of merge between patent applications data and US Treasury tax files

Our analysis relies on a new merge between published patent applications submitted to the US Patent and Trademark Office (USPTO) and US Treasury tax files. Below we describe the details of this merge, which relies on a fuzzy matching algorithm to link USPTO assignee names with US Treasury firm names.

A.4.1 Creating standardized names within the patent data

Published patent applications list an assignee name, which reflects ownership of the patent application. Due to, e.g., spelling differences, multiple assignee names in the USPTO published patent applications data can correspond to a single firm. For example, "ALCATEL-LUCENT U.S.A., INC.", "ALCATEL-LUCENT USA, INCORPORATED," and "ALCATEL-LUCENT USA INC" are all assigned the standardized name "alcatel lucent usa corp".

We employed a name standardization routine as follows. Starting with names in unicode format, we transform the text into Roman alphabet analogs using the "unidecode" library to map any foreign characters into their applicable English phonemes, and then shift all characters to lowercase.[18] We then standardize common terms that take multiple forms, such as "corp." and "corporation"; these recodings were built on the name standardization routine used by the National Bureau of Economic Research (NBER)'s Patent Data Project, with modifications as we saw opportunities to improve that routine.[19] We additionally eliminate any English articles (such as "a" or "an"), since these appeared to be uninformative in our attempts to uniquely identify entities. We then tokenize standardized names by splitting on natural delimiters (e.g., spaces and commas), after which we remove any non-alphanumeric

[18]The unidecode library is available at `https://github/com/iki/unidecode`, and is a direct Python port of the Text::Unidecode Perl module by Sean M. Burke.

[19]The NBER Patent Data Project standardization code is available at `https://sites.google.com/site/patentdataproject/Home/posts/namestandardizationroutinesuploaded`.

punctuation. Finally, sequences of single-character tokens are merged into a combined token (e.g., "3 m corp" would become "3m corp"). The resultant ordered list of tokens constitutes our standardized entity name. We refer to the USPTO standardized firm name as $SNAME_{USPTO}$.

A.4.2 Creating standardized names within the US Treasury tax files

In the US Treasury tax files, firms are indexed by their Employer Identification Number (EIN). Each EIN is required to file a tax return for each year that it is in operation. Specifically, we restrict our analysis to firms with valid 1120, 1120S, or 1065 filings over the years 1997–2014. We apply the same name standardization algorithm to the Treasury firm names that was applied to the USPTO names. We refer to the Treasury standardized firm name as $SNAME_{Treasury}$.

A.4.3 Merging standardized names across the USPTO data and the US Treasury tax files

We then conduct a fuzzy merge of $SNAME_{USPTO}$ to $SNAME_{Treasury}$ using the SoftTFIDF algorithm, which is described below. We use this algorithm to allocate each $SNAME_{USPTO}$ to a single $SNAME_{Treasury}$, provided that match quality lies within a specified tolerance. To choose the tolerance we used a hand coded match of applications to Compustat firms as a validation dataset (see Section A.4.4). The tolerance (and other tuning parameters) were chosen to minimize the sum of Type I and II error rates associated with matches to Compustat firms. The resulting firm-level dataset has one observation per $SNAME_{Treasury}$ in each year. However, there are some cases in which multiple EINs are associated with a given $SNAME_{Treasury}$. In those cases, we chose the EIN with the largest total income in the year of application in order to select the most economically active entity associated with that standardized name.[20]

SoftTFIDF algorithm Our firm name matching procedure of name $a \in SNAME_{USPTO}$ to name $b \in SNAME_{Treasury}$ works as follows. Among all the words in all the firm names in $SNAME_{USPTO}$ that are close to a given word in b, we pick the word with the highest SoftTFIDF index value, which is a word-frequency weighted measure of similarity among words. We do this for each word in the firm's name. For instance, American Airlines Inc would have three words. We then take a weighted-sum of the index value for each word in the firm name where the weights are smaller for frequent words like "Inc." This weighted sum is the SoftTFIDF value at the level of firm-names (as opposed to words in firm names). We assign a to the firm name b with the highest SoftTFIDF value above a threshold; otherwise, the name a is unmatched. Because of computational limitations, we limit comparisons to cases in which both a and b start with the same letter. Therefore we will miss any matches that do not share the same first letter. This subsection provides details on this procedure and example matches.

SoftTFIDF of firm names A score between groups of words X, Y is given by

$$\text{SoftTFIDF}(X,Y) := \sum_{w \in X} \text{weight}(w,X) \cdot \alpha(w,Y)$$

where $\text{weight}(w,X)$ is a word frequency-based importance weight and $\alpha(w,Y)$ is a word match score that uses a word similarity index. Specifically, the importance weight for the word w in the set of words Z is: $\text{weight}(w,Z) := \frac{\text{tfidf}(w,Z)}{\sqrt{\sum_{w' \in Z} \text{tfidf}(w',Z)^2}}$, where

- $\text{tfidf}(w,Z) := \text{tf}(w,Z) \times \text{idf}(w, \mathscr{Z})$,
- $\text{tf}(w,Z) := \frac{n(w,Z)}{\sum_{w' \in Z} n(w',Z)}$,

[20] For example, if two EINs shared the same standardized name $SNAME_{Treasury}$ but one EIN made 50 million in total income and the other showed three million in total income, we chose the EIN that earns 50 million.

- $\text{idf}(w, \mathscr{Z}) := \log\left(\frac{|\mathscr{Z}|}{|\{Z \in \mathscr{Z} \mid w \in Z\}|}\right)$,
- $n(w,Z)$ is the number of occurrences of word w in a set of words Z,
- $\mathscr{Z}$ is the set of all words in either $SNAME_{USPTO}$ or $SNAME_{IRS}$.

We compute the word match score $\alpha(w,Y)$ for words that are close to those in $SNAME_{USPTO}$. To determine which names are close, we use a Jaro-Winkler distance metric to measure the distance between two strings.

Jaro-Winkler metric of distance between strings We use this metric since it has been shown to perform better at name-matching tasks (Cohen, Ravikumar, and Fienburg 2003) than other metrics such as Levenshtein distance, which assigns unit cost to every edit operation (insertion, deletion, or substitution). A key component of the Jaro-Winkler metric is the Jaro metric. The Jaro metric depends on the length of $SNAME_{USPTO}$, the length of $SNAME_{Treasury}$, the number of shared letters, and the number of needed transpositions of shared letters.

Specifically, consider strings $s = s_1 \dots s_K$ and $t = t_1 \dots t_L$ and define $H = \frac{\min\{|s|,|t|\}}{2}$, which is half the smaller of K and L. We say a character s_i is **in common with** t if $\exists j \in [i-H, i+H]$ s.t. $s_i = t_j$. Let s', t' be the ordered sets of **in-common** characters (hence we will re-index). Then define $T_{s',t'} := \frac{1}{2}|\{i \mid s'_i \neq t'_i\}|$. The similarity metric is given by

$$\text{Jaro}(s,t) := \frac{1}{3} \cdot \left(\frac{|s'|}{|s|} + \frac{|t'|}{|t|} + \frac{|s'| - T_{s',t'}}{|s'|}\right).$$

The Jaro-Winkler metric is given by

$$\text{Jaro-Winkler}(s,t) := \text{Jaro}(s,t) + \frac{P'}{10} \cdot (1 - \text{Jaro}(s,t)),$$

where P as the longest common prefix of t and s and then $P' = \max\{P, 4\}$, which is the normalization used in Cohen, Ravikumar, and Fienburg (2003).

Word match score $\alpha(w,Z)$ We define the word match score as follows:

$$\alpha(w,Z) = \begin{cases} 0 & \text{if } \text{closest}(\theta, w, Z) = \emptyset \\ \max\limits_{w' \in \text{closest}(\theta,w,Z)} \text{weight}(w', Z) \cdot \text{Jaro-Winkler}(w, w') & \text{otherwise} \end{cases}$$

where

$$\text{closest}(\theta, w, Z) := \left\{v \in Z \mid \forall v' \in Z,\ (\text{Jaro-Winkler}(w,v) \geq \text{Jaro-Winkler}(w,v')) \ \wedge\ \text{Jaro-Winkler}(w,v) > \theta\right\}.$$

In words, we select the word w that is the closest importance-weighted match among words that are close to the word w in Z given closeness threshold θ. The accuracy of this matching procedure, which has also recently been used by Feigenbaum (2016), will likely become clearer after reviewing the following examples and discussing how we selected the tuning parameters (such as the closeness threshold θ).

Example

USPTO Assignee Name	Compustat Firm Name (best match)	Match Score
angiotech pharmaceuticals corp	angiotech pharmaceuticals	.9982
assg brooks justin	brooks resources corp	.5857
hewlett packard development corp	hewlett packard corp	.8482
huawei device corp	huatue electronics corp	.0013
matsushita electric works corp	matson corp	.0012
olympus corp	olympus capital corp	.9109
safety crafted solutions corp	safety first corp	.3862
sc johnson home storage corp	sc holdings corp	.5144

This table provides a small sample of candidate matches from a USPTO to Compustat match.[21]

A.4.4 Validation: Compustat-USPTO match

This section describes the hand matching process we used to determine the true mapping of USPTO names to Compustat names for a random sample of USPTO names. We describe the hand coding task and how we use the hand coded linkages to select the tuning parameters.

Hand coding tasks We hired several workers on Upwork (formerly Odesk) as well as University of Chicago undergraduates to hand match two lists of names. The goal for these workers was to match every name in a source file (a list of 100 randomly selected USPTO names) to a target file of Compustat names or to conclude that there is no matching name in the target file. To increase accuracy, we informed these workers that (1) we had hand-coded several of these names ourselves, (2) every name in the source file would be assigned to multiple workers, and (3) we would only accept reasonably accurate work. We also instructed them to use Google to confirm that matches were true matches. For example, "infinity bio ltd" may seem like a match with "infinity pharmaceuticals inc," but Googling the first reveals that the former is a small Brazilian energy company while the latter is a pharmaceutical company headquartered in the US. If one worker found a match but another did not, we considered the non-empty match to be correct. Overall, we ended up assigning 2,196 assignee names to workers, of which 286 (13%) had matches in the Compustat data.[22]

Using hand coding tasks to select tuning parameters We use these hand-coded linkages to establish the "true mapping" from USPTO names to Compustat names, which enables us to select tuning parameters that minimize the sum of type I and type II errors (relative to these "true linkages").

We constructed a grid and for each set of parameters on the grid executed a match. We then compared these fuzzy matchings to the "true mapping." Type I errors occur when SoftTFIDF returns a match but either (a) the match is inconsistent with the hand-coded match or (b)the hand coded linkage shows no match at all. Type II errors occur when SoftTFIDF does not return a match but the hand-code process had a match.

The parameters that minimize the sum of these false positive and false negatives are: $\theta = .95$, token type of standardized names (instead of raw names), $P = 0$, and a threshold match score of .91. We remind the reader that the parameter θ governs the threshold similarity for two words to be considered "close." Only "close" words contribute to a match score, hence $\theta = .95$ sets a relatively high cutoff below which two similar words do not increase the match score between two fill names. The prefix $P = 0$ suggests that not boosting scores by a common

[21] We present results using Compustat names instead of Treasury names for disclosure reasons.

[22] This match rate is sensible: the number of Compustat names is roughly 20% the number of assignee names, so this rate is consistent with a reasonable proportion of Compustat firms applying for a patent.

prefix doesn't improve performance, which makes sense given that we block by the first letter already.[23] Finally, the threshold match score of .91 shows that we should only consider names a match if they are very close by our similarity metric. With these parameters, Type I and II errors are each below six percent.

A.4.5 Validation: Individual-inventor match

The Bell et al. (2019) inventor-level merge between patent applications and W2 reports in theory can — via the EINs provided on W2 reports — provide a linkage between patent applications and firms, but ex-ante we expect this inventor-based match to measure something conceptually different from a firm-based match. For example, many inventors work at firms that are not the assignee of their patents, in which case we would not expect our assignee-based merge to match to the same EIN as the Bell et al. (2019) inventor-based merge. However, the Bell et al. (2019) merge nonetheless provides a very valuable benchmark for assessing the quality of our assignee-based merge. Bell et al. generously agreed to share their inventor-based merge with us, and our preliminary results comparing the two linkages provide a second set of evidence supporting the quality of our assignee-based linkage. In the simplest comparison, around 70% of patent applications are associated with the same EIN in the two linkages. The characteristics of this match also look sensible, e.g., the match rates are higher if we limit the sample to patent applications that Bell et al. (2019) match to inventors who all work at the same firm. Given that we do not expect a match rate of 100% for the reasons detailed above, we view the results of this second validation exercise as quite promising.

[23] It is computationally infeasible to compare every single entity against every other, so we utilized first-letter blocking in order to reduce the sizes of sets being compared against one another. In particular, the target names (either Compustat or Treasury names) are chunked by the first letter of their standardized names and grouped with the source (USPTO) names with the same first letter. Hence, we will miss any matches that differ on the first letter.

B Appendix: Further Details on Sample Restrictions

This section describes the way we implement the sample restrictions in more detail.

The last row of Table 1 Panel B shows a decline from 35,643 firms (EINs) to 9,732. As mentioned in Section 5, we estimate the Poisson model of patent value described before making the restriction to active firms. Specifically, the set of firms in the Poisson analysis are the 35,643 firms that (1) successfully matched an application in the USPTO-tax merge, (2) corresponded to the first application by that EIN, (3) did not have a prior grant, and (4) were the EIN with the largest revenue in the application year. As noted in Section 5, the 596 of these 35,643 firms with a valid KPSS value are used to estimate the Poisson model for patent values ξ_j reported in Table 4. Due to missing covariates, we were unable to form predicted patent values $\hat{\xi}_j$ from the Poisson model for 805 firms, reducing the sample to 34,838 firms.

After forming the Poisson predictions, we make the final restriction to focus on active firms, which are EINs with non-zero/non-missing total income or total deductions in the application year and in the three previous years, a positive number of employees in the application year, and revenue less than 100 million in 2014 USD. When we restrict to EINs with non-zero/non-missing total income or total deductions in the application year and in the three previous years and a positive number of employees in the application year, we drop of 24,633 of 34,838 EINS. When we further restrict that sample to EINs with revenue less than 100 million in 2014 USD in the application year, we drop another 473 firms and end up with our main estimation sample of 9,732 firms. As noted in Section 5, imposing these restrictions leaves only 159 firms with valid raw KPSS values ξ_j, however all 9,732 firms have valid predicted values $\hat{\xi}_j$ from the Poisson model.

C Appendix: Poisson model of patent value

Recall that the probability mass function for a Poisson distributed outcome Y with mean λ can be written:

$$p(Y|\lambda) = \exp(Y\lambda - \exp(\lambda))/Y!$$

Let $Y_a = (Y_1, ..., Y_{m_a})$ and $X_a = (X_1, ..., X_{m_a})$ denote the vectors of outcomes and covariates respectively in an art unit a. Supposing $Y_j|X_a \overset{iid}{\sim} Poisson\left(X_j'\delta + \nu_a\right)$ where ν_a is a scalar art unit effect, we can write:

$$\begin{aligned}\ln p(Y_a|X_a, \nu_a) &= \sum_{j=1}^{m_a} \ln p\left(Y_j|X_j'\delta + \nu_a\right) \\ &= \sum_{j=1}^{m_a} Y_j\left(X_j'\delta + \nu_a\right) - \exp\left(X_j'\delta + \nu_a\right) - \ln(Y_j!)\end{aligned}$$

The random effects Poisson likelihood of an art unit a can be written:

$$L(Y_a|X_a) = \frac{1}{\sqrt{2\pi}\sigma_\eta}\int \exp\left\{\ln p(Y_a|X_a) - \frac{1}{2}\frac{\nu^2}{\sigma_\nu^2}\right\} d\nu$$

By independence across art units, the full log likelihood can be written $\sum_a \ln L(Y_a|X_a)$.

The first order condition for the coefficient vector δ is:

$$\begin{aligned}\frac{d}{d\delta}\sum_a \ln L(Y_a|X_a) &= \sum_a \frac{\int\left(\sum_{j=1}^{m_a}\left[Y_j - \exp\left(X_j'\delta + \nu\right)\right]X_j\right)\exp\left\{\ln p(Y_a|X_a,\nu) - \frac{1}{2}\frac{\nu^2}{\sigma_\nu^2}\right\}d\nu}{\int \exp\left\{\ln p(Y_a|X_a,\nu) - \frac{1}{2}\frac{\nu^2}{\sigma_\nu^2}\right\}d\nu} \\ &= \sum_a\sum_{j=1}^{m_a}\left[Y_j - \int \exp\left(X_j'\delta + \nu\right)\omega_a(\nu)\,d\nu\right]X_j = 0\end{aligned}$$

where the weighting function $\omega_a(z) = \frac{\exp\left\{\ln p(Y_a|X_a,z) - \frac{1}{2}\frac{z^2}{\sigma_\nu^2}\right\}}{\int \exp\left\{\ln p(Y_a|X_a,\nu) - \frac{1}{2}\frac{\nu^2}{\sigma_\nu^2}\right\}d\nu}$ is the posterior density of ν given the observables in art unit a. Note that this is a shrunken version of the usual Poisson orthogonality condition that is robust to misspecification of features of the conditional distribution other than the mean (Wooldridge 2010). The weights, however, rely on the exponential nature of the Poisson density function which, if misspecified, will yield inconsistency in small art units. In large art units, however, the posterior will spike around the "fixed effect" estimate of ν, which is again robust to misspecification of higher moments of the conditional distribution.

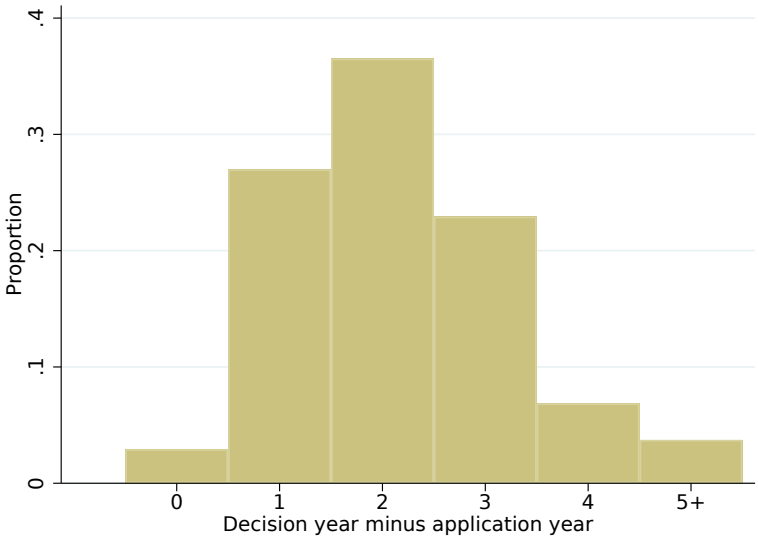
The first order condition for the variance σ_ν is:

$$\begin{aligned}\sum_a \frac{d}{d\sigma_\nu}\ln L(Y_a|X_a) &= \sum_a \left\{-\frac{1}{\sigma_\eta} + \frac{\int \frac{\eta^2}{\sigma_\eta^3}\exp\left\{\ln p(Y_a|X_a,\nu) - \frac{1}{2}\frac{\nu^2}{\sigma_\nu^2}\right\}d\nu}{\int \exp\left\{\ln p(Y_a|X_a,\nu) - \frac{1}{2}\frac{\nu^2}{\sigma_\nu^2}\right\}d\nu}\right\} \\ &= \frac{1}{\sigma_\eta^3}\sum_a\left[\int \omega_a(\nu)\,\nu^2 d\nu - \sigma_\nu^2\right] = 0.\end{aligned}$$

This simply says that the posterior variance of ν in each art unit should average across art units to σ_ν^2.

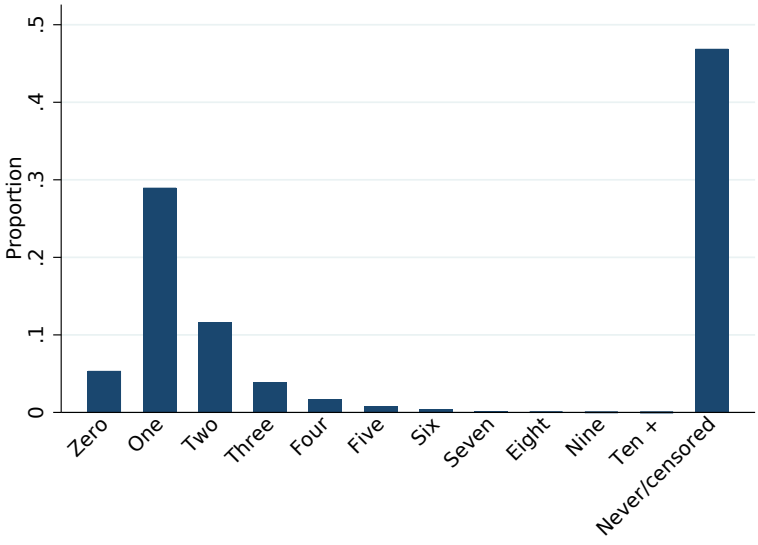
D Appendix: Additional figures and tables

Figure D.1: Years Until Initial Decision



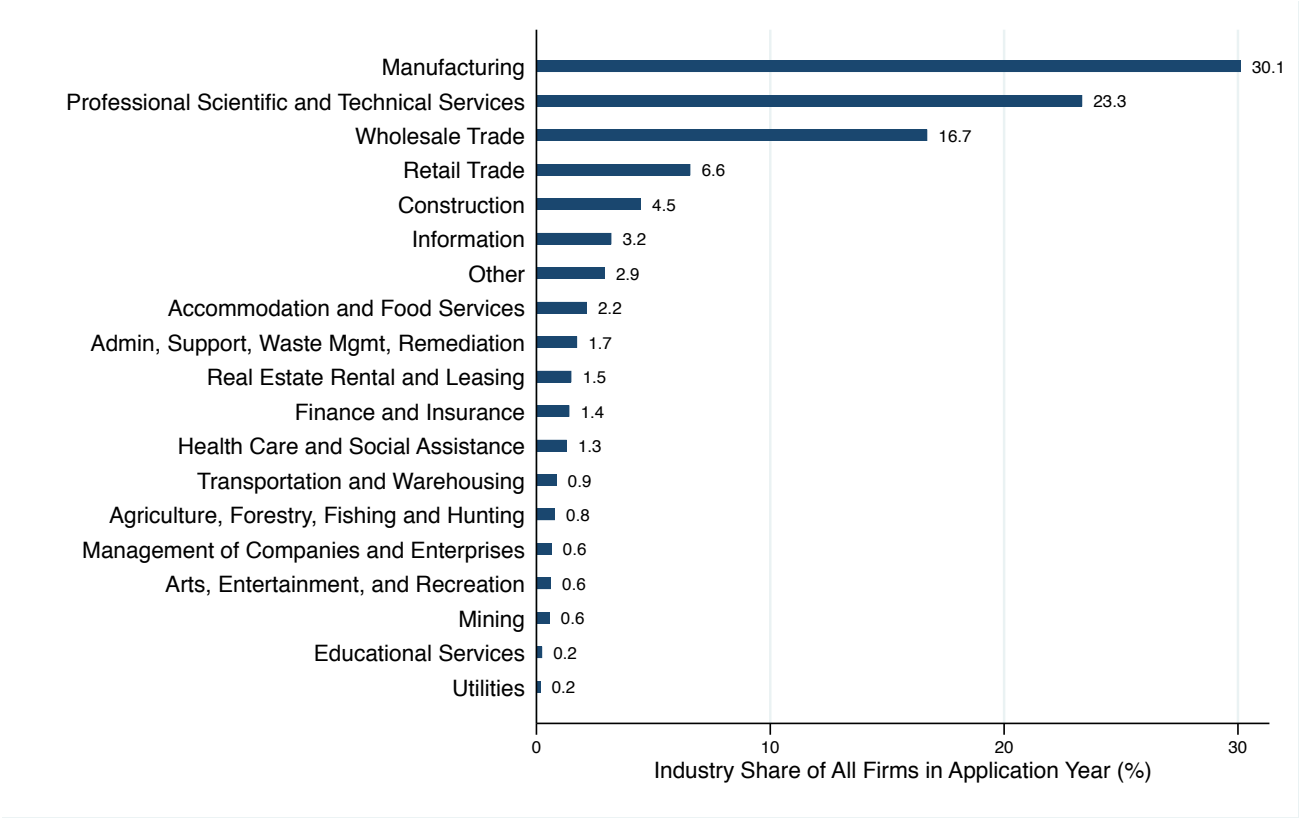
Notes: This figure plots a histogram of the years until the initial patent application decision for the sample of patent assignees by application pairs in the bottom row of Panel A of Table 1 (N=99,871).

Figure D.2: Years Until Patent Grant for Initially Rejected Patent Applications



Notes: This figure plots a histogram of the years until a patent grant for the subsample of patent assignee by application pairs in the bottom row of Panel A of Table 1 (N=99,871) which receive an initial rejection (N=88,298).

Figure D.3: Industry Composition of Firms



Notes: This figure plots the distribution of firms in our sample by industry. The distribution of firms whose patent application is initially granted is similar.

Table D.1: Testing for Spatial Correlation in Initial Allowance Decisions

	Initially allowed							
	State		Zip		4-D NAICS		4-D NAICS × State	
	(1a)	(1b)	(2a)	(2b)	(3a)	(3b)	(4a)	(4b)
Panel A: analysis sample								
Intra-class correlation (ρ)	0.000	0.000	0.068	0.058	0.000	0.000	0.000	0.000
p-value	1.000	1.000	0.297	0.465	1.000	1.000	1.000	1.000
Observations	9,732	8,647	9,732	8,647	9,732	8,647	9,732	8,647
Number of categories	51	51	4,501	4,231	355	347	3,376	3,185
AU-AY FEs		✓		✓		✓		✓
Panel B: top quintile								
Intra-class correlation (ρ)	0.113	0.132	0.000	0.000	0.000	0.000	0.000	0.046
p-value	0.165	0.184	1.000	1.000	1.000	1.000	1.000	0.416
Observations	1,946	1,666	1,946	1,666	1,946	1,666	1,946	1,666
Number of categories	49	49	1,252	1,113	250	242	1,119	1,013
AU-AY FEs		✓		✓		✓		✓

Notes: This table reports the results of tests for whether initial patent allowances are geographically clustered, separately for the analysis sample and for the top-quintile predicted patent value sample. "Intra-class correlation (ρ)" reports the ratio of a random effects estimate of the geographic variance component to the sum of the geographic and idiosyncratic variance components. The p-value reports a Breusch-Pagan Lagrange multiplier test of the null hypothesis that ρ=0. AU-AY FEs denotes the inclusion of Art Unit (AU) by application year (AY) fixed effects.

Table D.2: Mean $\hat{\xi}$ by Technology Center

Technology center	$\overline{\hat{\xi}}$	N	Technology center	$\overline{\hat{\xi}}$	N
Business Methods - Finance	15.079	152	Telecomms: Analog Radio	3.080	43
Electronic Commerce	10.237	365	Mining, Roads, & Petroleum	2.991	518
Databases & File Mgmt	9.726	261	Microbiology	2.983	83
Tires, Adhesives, Glass, & Plastics	8.035	134	Semiconductors, Circuits, & Optics	2.903	237
2180: Computer Architecture	8.029	68	Molec Bio & Bioinformatics	2.891	68
Combust & Fluid Power Systems	7.803	111	Amusement & Education Devices	2.780	236
Aero, Agriculture, & Weaponry	7.129	224	Static Structures & Furniture	2.627	560
Selective Visual Display Systems	6.387	200	Fuel Cells & Batteries	2.437	177
Computer Graphic Processing	6.012	299	Business Methods	2.416	193
Optics	5.650	341	2110: Computer Architecture	2.163	50
Organic Chemistry & Polymers	5.622	116	Software Development	2.141	76
Organic Compounds	5.316	115	Medical Instruments	2.110	132
Organic Chemistry	4.913	62	Multiplex & VoIP	1.990	72
Manufact Devices & Processes	4.870	443	Metallurgy & Inorganic Chemistry	1.980	102
Memory Access & Control	4.865	49	Chemical Apparatus	1.947	171
Selective Communication	4.674	294	Semiconductors & Memory	1.936	185
Surface Transportation	4.428	294	Cryptography & Security	1.898	76
Electrical Circuits & Systems	4.090	266	Medical & Surgical Instruments	1.823	118
Coating, Etching, & Cleaning	3.615	83	Computer Networks	1.733	145
Misc. Computer Applications	3.459	129	Radio, Robotics, & Nucl Systems	1.597	85
Material & Article Handling	3.248	255	Receptacles, Shoes, & Apparel	1.444	470
Graphical User Interface	3.217	152	Kinestherapy & Exercising	1.330	138
Refrigeration & Combustion	3.084	265	Fluid Handling	0.706	188

Notes: This table reports the mean predictions of ex-ante value $\hat{\xi}$ by USPTO technology center of the application; technology centers are administrative groupings of art units designated by the USPTO. The sample is observations from our analysis sample whose application belongs to a technology center with more than 20 observations in the analysis sample (N=6,402). $\hat{\xi}$ is measured in millions of 1982 USD.

Table D.3: Impacts on Closely Held Firm Aggregates

	(1) Positive employment	(2) Log firm size	(3) Revenue per worker	(4) Value added per worker	(5) EBITD per worker	(6) Wage bill per worker	(7) Surplus per worker	(8) Labor comp per worker	(9) W2 + 1099 per worker	(10) Income tax per worker
High value ($Q5$)	-0.04	0.33	79.75	32.72	21.84	8.49	31.11	7.13	5.17	2.21
	(0.08)	(0.14)	(24.60)	(8.24)	(6.08)	(3.42)	(7.44)	(1.38)	(1.98)	(1.64)
Mean of outcome ($Q5$)	0.71	3.00	305.20	119.30	20.11	49.40	70.60	46.82	43.62	18.21
% Impact ($Q5$)	-5.2		26.1	27.4	108.6	17.2	44.1	15.2	11.9	12.2
Lower value ($< Q5$)	-0.02	0.02	-3.29	-0.70	0.71	0.30	0.53	0.34	-0.11	1.23
	(0.02)	(0.05)	(13.95)	(6.19)	(3.19)	(1.51)	(4.03)	(2.55)	(1.41)	(1.07)
Observations	75,132	49,943	49,943	49,943	49,943	49,943	49,943	49,943	51,998	49,808

Notes: This table reports difference-in-differences estimates of the effect of initial patent allowances on firm and worker outcomes, separately for high and low ex-ante valuable patent applications. It restricts the analysis in Table 5 to S-corporations and partnerships ("closely held" firms). Estimates correspond to coefficients on interactions of the designated value category with a post-decision indicator and an indicator for the application being initially allowed. Controls include main effect of value category interacted with a post-decision indicator, firm fixed effects, and art unit by application year by calendar year fixed effects, as in equation (8). Standard errors (reported in parentheses) are two-way clustered by (1) art unit, and (2) application year by decision year. EBITD is earnings before interest, taxes, and deductions. Surplus is EBITD + wage bill. Labor compensation measures total deductions for labor expenses claimed by the firm. "W2 + 1099" measures to the sum of W2 and 1099 earnings divided by the sum of the number of W2's and 1099's filed. "Income tax per worker" is the average worker's individual income tax liability. Revenue, value added, EBITD, wage bill, surplus, labor compensation, and W2 + 1099 pay are reported in thousands of 2014 USD.

Table D.4: Heterogeneity across larger and smaller firms

	(1) Positive employ-ment	(2) Log firm size	(3) Revenue per worker	(4) Value added per worker	(5) EBITD per worker	(6) Wage bill per worker	(7) Surplus per worker	(8) Labor comp per worker	(9) W2 + 1099 per worker	(10) Average stayer earnings
Panel A: large firms										
High value ($Q5$)	0.05	0.20	39.36	14.04	8.89	2.16	10.67	1.69	1.98	6.81
	(0.03)	(0.07)	(17.51)	(6.95)	(5.38)	(2.73)	(6.26)	(3.74)	(2.10)	(3.13)
Mean of outcome ($Q5$)	0.72	4.02	278.10	103.90	7.10	57.91	65.28	52.21	52.61	73.47
% Impact ($Q5$)	7.5		14.2	13.5	125.2	3.7	16.3	3.2	3.8	9.3
Lower value ($< Q5$)	0.01	-0.07	1.71	4.78	-0.03	-0.18	0.25	2.10	-0.46	-1.48
	(0.02)	(0.05)	(11.48)	(5.13)	(2.64)	(1.15)	(3.08)	(1.79)	(1.32)	(2.32)
Panel B: small firms										
High value ($Q5$)	-0.10	0.28	19.59	16.56	8.82	5.15	13.28	7.37	2.99	7.52
	(0.08)	(0.15)	(25.06)	(7.68)	(4.03)	(3.08)	(5.18)	(4.41)	(3.69)	(6.68)
Mean of outcome ($Q5$)	0.66	1.77	335.00	135.10	12.12	55.58	69.66	59.99	45.74	71.08
% Impact ($Q5$)	-15.4		5.9	12.3	72.8	9.3	19.1	12.3	6.6	10.6
Lower value ($< Q5$)	-0.02	0.13	-22.21	-3.57	-2.93	1.81	-0.89	0.49	1.48	6.88
	(0.02)	(0.06)	(14.07)	(5.57)	(1.95)	(1.26)	(2.36)	(2.35)	(0.98)	(2.23)
Observations	155,646	103,437	103,437	103,437	103,437	103,437	103,437	103,437	107,789	99,558

Notes: This table reports difference-in-differences estimates of the effect of initial patent allowances interacted with large and small firm size on firm and worker outcomes, separately for high and low ex-ante valuable patent applications, in our analysis sample. Large firms are those above median baseline employment, and small firms are those below median baseline employment. Estimates correspond to coefficients on interactions of the designated value category with a post-decision indicator, an indicator for the application being initially allowed, and whether the employment of the firm was above or below median baseline employment. Controls include main effect of value category interacted with a post-decision indicator, firm fixed effects, and art unit by application year by calendar year fixed effects. Standard errors (reported in parentheses) are two-way clustered by (1) art unit, and (2) application year by decision year. Panel A reports the estimates for the large firm indicator interacted with a post-decision indicator and an initial allowance indicator. Panel B reports the estimates for the small firm interacted with a post-decision indicator and an initial allowance indicator. EBITD is earnings before interest, taxes, and deductions. Surplus is EBITD + wage bill. Stayers are defined as those who were employed by the same firm in the year of application. Labor compensation measures total deductions for labor expenses claimed by the firm. "W2 + 1099" measures to the sum of W2 and 1099 earnings divided by the sum of the number of W2's and 1099's filed. "% Impact" reports the percent change in the outcome at the mean for winning an initial allowance. Revenue, value added, EBITD, wage bill, surplus, earnings, labor compensation, and W2 + 1099 pay are reported in thousands of 2014 USD.

Table D.5: Within-Firm Inequality

	(1) Avg male earnings	(2) Avg female earnings	(3) Gender earnings gap	(4) Avg inventor earnings	(5) Avg earnings of non-inventors	(6) Inventor earnings gap	(7) Wage bill per worker (Q_1)	(8) Wage bill per worker (Q_4)	(9) Wage bill per worker ($Q_4 - Q_1$)
High value ($Q5$)	5.88	0.15	6.90	16.87	2.24	14.92	-0.04	8.06	8.12
	(1.97)	(1.51)	(1.97)	(8.47)	(1.43)	(7.75)	(1.01)	(2.85)	(2.56)
Mean of outcome ($Q5$)	66.43	39.53	27.68	139.00	51.72	85.98	18.22	120.50	102.30
% Impact ($Q5$)	8.9	0.4	24.9	12.1	4.3	17.4	-0.2	6.7	7.9
Lower value ($< Q5$)	0.35	-0.49	-0.06	-1.24	0.47	-1.72	0.05	2.82	2.77
	(1.16)	(0.51)	(1.06)	(4.60)	(0.81)	(4.70)	(0.32)	(2.42)	(2.34)
Observations	95,004	84,562	80,222	52,471	100,901	50,045	82,750	81,536	81,536

Notes: This table reports difference-in-differences estimates of the effect of initial patent allowances on within-firm inequality measures, separately for high and low ex-ante valuable patent applications, in our analysis sample. Estimates correspond to coefficients on interactions of the designated value category with a post-decision indicator and an indicator for the application being initially allowed. Controls include main effect of value category interacted with a post-decision indicator, firm fixed effects, and art unit by application year by calendar year fixed effects, as in equation (8). Standard errors (reported in parentheses) are two-way clustered by (1) art unit, and (2) application year by decision year. "Gender earnings gap" measures the difference between average male and female earnings at firms where both genders are present. "Inventor earnings gap" measures the difference between average inventor and non-inventor earnings at firms where both inventors and non-inventors are present. "Wage bill per worker (Q_1)" measures the average wage bill in within-firm wage quartile one. "Wage bill per worker (Q_4)" measures the average wage bill in within-firm wage quartile four. "Wage bill per worker $(Q_4 - Q_1)$" measures the difference between average Q_4 and Q_1 earnings. Stayers are defined as those who were employed by the same firm in the year of application. Entrants are defined as those employees who were not employed at the firm in the previous year. Earnings and wage bill are measured in thousands of 2014 USD.

Table D.6: Within-Firm Inequality (Using a Balanced Sample)

	(1) Avg male earnings	(2) Avg female earnings	(3) Gender earnings gap	(4) Avg inventor earnings	(5) Avg earnings of non-inventors	(6) Inventor earnings gap	(7) Wage bill per worker (Q_1)	(8) Wage bill per worker (Q_4)	(9) Wage bill per worker ($Q_4 - Q_1$)
High value ($Q5$)	7.11	0.21	6.90	14.44	-0.48	14.92	-0.06	8.06	8.12
	(1.79)	(1.54)	(1.97)	(7.59)	(1.86)	(7.75)	(1.01)	(2.85)	(2.56)
Mean of outcome ($Q5$)	67.53	39.85	27.68	140.40	54.38	85.98	18.16	120.50	102.30
% Impact ($Q5$)	10.5	0.5	24.9	10.3	-0.9	17.4	-0.3	6.7	7.9
Lower value ($< Q5$)	-0.23	-0.18	-0.06	-1.35	0.37	-1.72	0.05	2.82	2.77
	(1.12)	(0.45)	(1.06)	(4.91)	(1.18)	(4.70)	(0.32)	(2.42)	(2.34)
Observations	80,222	80,222	80,222	50,045	50,045	50,045	81,536	81,536	81,536

Notes: This table replicates Table D.5 using a common sample for each comparison group. Each column reports difference-in-differences estimates of the effect of initial patent allowances on within-firm inequality measures, separately for high and low ex-ante valuable patent applications, in a subsample of our analysis sample where the composition of firms is held constant across the related outcome columns. Estimates correspond to coefficients on interactions of the designated value category with a post-decision indicator and an indicator for the application being initially allowed. Controls include main effect of value category interacted with a post-decision indicator, firm fixed effects, and art unit by application year by calendar year fixed effects. Standard errors (reported in parentheses) are two-way clustered by (1) art unit, and (2) application year by decision year. "Gender earnings gap" measures the difference between average male and female earnings at firms where both genders are present. "Inventor earnings gap" measures the difference between average inventor and non-inventor earnings at firms where both inventors and non-inventors are present. "Wage bill per worker (Q_1)" measures the average wage bill in within-firm wage quartile one. "Wage bill per worker (Q_4)" measures the average wage bill in within-firm wage quartile four. "Wage bill per worker $(Q_4 - Q_1)$" measures the difference between average Q_4 and Q_1 earnings. "% Impact" reports the percent change in the outcome at the mean for winning an initial allowance. Earnings and wage bill are measured in thousands of 2014 USD.

Table D.7: Earnings Impacts by Stayer Subgroups

	(1) Avg male stayer earnings	(2) Avg female stayer earnings	(3) Stayer gender earnings gap	(4) Avg inventor stayer earnings	(5) Avg non-inventor stayer earnings	(6) Stayer inventor earnings gap
High value ($Q5$)	13.83	2.73	8.89	17.41	5.75	9.27
	(2.74)	(1.95)	(3.51)	(11.21)	(1.72)	(8.35)
Mean of outcome ($Q5$)	85.03	48.35	37.77	156.00	64.37	91.85
% Impact ($Q5$)	16.3	5.7	23.5	11.2	8.9	10.1
Lower value ($< Q5$)	2.17	0.70	-0.83	0.58	2.03	-2.78
	(1.96)	(0.83)	(1.80)	(6.31)	(1.26)	(6.91)
Observations	88,100	71,591	66,270	47,063	94,909	42,640

Notes: This table reports difference-in-differences estimates of the effect of initial patent allowances on within-firm inequality measures, separately for high and low ex-ante valuable patent applications, in a subsample of our analysis sample where the composition of firms is held constant across the related outcome columns. Estimates correspond to coefficients on interactions of the designated value category with a post-decision indicator and an indicator for the application being initially allowed. Controls include main effect of value category interacted with a post-decision indicator, firm fixed effects, and art unit by application year by calendar year fixed effects. Standard errors (reported in parentheses) are two-way clustered by (1) art unit, and (2) application year by decision year. "Stayer gender earnings gap" measures the difference between average male stayer and female stayer earnings at firms where both genders are present. "Stayer inventor earnings gap" measures the difference between average inventor stayer and non-inventor stayer earnings at firms where both inventors and non-inventors are present. "% Impact" reports the percent change in the outcome at the mean for winning an initial allowance. Stayers are defined as those who were employed by the same firm in the year of application. Earnings are measured in thousands of 2014 USD.

Table D.8: Earnings of Officers / Owners

	All firms			Pass-through entities					
	(1) Officer earnings per W2	(2) Wages and salaries per W2	(3) Non-officer comp per W2	(4) Officer earnings per W2	(5) Wages and salaries per W2	(6) Non-officer comp per W2	(7) Avg earnings of owner-employees	(8) Avg pay of owner-employees	(9) Avg W2 earnings of non-owner-employees
High value ($Q5$)	3.81	0.13	-0.05	7.17	-1.43	-1.67	42.87	84.08	5.99
	(1.31)	(1.89)	(2.10)	(3.27)	(1.66)	(1.56)	(17.01)	(41.36)	(2.41)
Mean of outcome ($Q5$)	15.89	35.07	38.91	14.86	27.85	31.25	149.70	246.60	41.76
% Impact ($Q5$)	24.0	0.4	-0.1	48.3	-5.1	-5.4	28.6	34.1	14.4
Lower value ($< Q5$)	-0.83	1.46	1.97	-1.28	0.46	1.22	-0.28	-3.71	0.66
	(0.94)	(1.01)	(1.02)	(1.38)	(1.31)	(1.49)	(5.45)	(18.72)	(0.84)
Observations	103,437	103,437	103,437	49,943	49,943	49,943	31,962	45,318	43,531

Notes: This table reports difference-in-differences estimates of the effect of initial patent allowances on officer and non-officer earnings measures for all firms and pass-through entities. Estimates correspond to coefficients on interactions of the designated value category with a post-decision indicator and an indicator for the application initially allowed. Controls include main effect of value category interacted with a post-decision indicator, firm fixed effects, and art unit by application year by calendar year fixed effects. Standard errors (reported in parentheses) are two-way clustered by (1) art unit, and (2) application year by decision year. "% Impact" reports the percent change in the outcome at the mean for winning an initial allowance. Earnings are measured in thousands of 2014 USD. The outcome variables are defined as follows. (1) Officer earnings per W2 is the officer compensation component of labor compensation that we define in A.3. Specifically, the numerator of (1) is Line 12 on C-corporation form 1120, line 7 on S-corporation form 1120S, and line 10 on Partnership form 1065. Each of these line numbers corresponds to the 2010 IRS forms. Similarly, the numerator of Column (2) is the salaries and wages component of labor compensation, which is line 13, 8, and 9 on the C-corporation, S-corporation, and partnership forms, respectively. The numerator of Column (3) is labor compensation less officer compensation, so it includes the other components of labor compensation (i.e., it includes salaries and wages, pension profit-sharing plans, and employee benefit programs). On form 1120, for example, these are lines 13, 23, and 24. Columns (4)-(6) restrict the sample to only S-corporations and Partnerships, but have the same definitions, respectively, as Columns (1)-(3). Column (7) is the average W2 wage earnings that go to firm owners, which was constructed using matched firm-owner data from Smith et al. (2019). Specifically, for each S-corporation and partnership, we calculate the sum of W2 wages that accrue to its owners and divide this sum by the number of owners who also get positive W2 income from the firm. Column (8) is average owner pay, which is the sum of W2 wage earnings that accrue to owners plus their business income divided by the number of owners. Each owners' business income comes from Smith et al. (2019)'s linkage between firms and owners, which was constructed using Schedule K1 of form 1120S for S-corporations and form 1065 for Partnerships. Being able to identify business owners directly is only possible for S-corporations and partnerships, so only aggregate firm-level business income is available for C-corporations (on line 28 of form 1120). Finally, Column (9) is the total W2 earnings for non-owners divided by the number of non-owner W2s.

Table D.9: Pass-Through Estimates: Three-Year Average of Surplus per Worker

	Wage bill per worker		Avg male earnings		Avg non-inventor earnings		Avg stayer earnings		Avg earnings of stayers minus earnings in app yr		Avg non-inventor stayer earnings	
	(1a) OLS	(1b) IV	(2a) OLS	(2b) IV	(3a) OLS	(3b) IV	(4a) OLS	(4b) IV	(5a) OLS	(5b) IV	(6a) OLS	(6b) IV
Surplus / worker	0.20	0.35	0.23	0.70	0.16	0.31	0.26	0.66	0.25	0.61	0.21	0.60
	(0.01)	(0.15)	(0.01)	(0.25)	(0.01)	(0.14)	(0.01)	(0.34)	(0.01)	(0.32)	(0.01)	(0.28)
Elasticity	0.239	0.416	0.238	0.717	0.211	0.405	0.246	0.625	.	.	0.218	0.634
Observations	83,212	83,212	77,066	77,066	81,632	81,632	81,075	81,075	81,075	81,075	77,939	77,939
1^{st} stage F		58.05		58.05		67.35		67.35		52.67		52.67
Exogeneity		73.190		73.190		2.888		2.888		65.000		65.000

Notes: This table reports OLS and IV estimates of the effect of increases in surplus per worker on selected earnings outcomes using three-year averages of surplus per worker. The excluded instrument is the interaction of top quintile of ex-ante value $\hat{\xi}$ category with a post-decision indicator and an indicator for the application being initially allowed. Controls include main effect of value category interacted with post-decision indicator and interaction of lower quintile value category with a post-decision indicator times an indicator for initially allowed, firm fixed effects, and art unit by application year by calendar year fixed effects. Standard errors (reported in parentheses) are two-way clustered by (1) art unit, and (2) application year by decision year. "Exogeneity" reports p-value for test of null hypothesis that IV and OLS estimators have same probability limit. Stayers are defined as those who were employed by the same firm in the year of application. Surplus is EBITD (earnings before interest, tax, and depreciation) + wage bill. Earnings, wage bill, and surplus are measured in thousands of 2014 USD.

Table D.10: Sensitivity analysis of retention estimates and model based interpretation

	(1) Baseline	(2) Low incumbent premium	(3) High retention elasticity	(4) Low retention elasticity	(5) High pass-through rate	(6) Low pass-through rate
Panel A. Calibrated wage premium						
Calibrated inputs						
w_j^I/w_j^m	1.8	1.2	1.8	1.8	1.8	1.8
$d\ln G\left(w_j^I\right)/d\ln w_j^I$	1.2	1.2	1.8	0.6	1.2	1.2
π	0.61	0.61	0.61	0.61	0.91	0.31
Model-based outputs						
η	2.7	7.3	4.0	1.4	2.7	2.7
$c'(N_j/I_j)/w_j^m$	1.1	0.2	1.0	1.4	1.1	1.1
θ	0.73	0.88	0.80	0.59	0.73	0.73
ε	6.0	3.3	4.2	-	-	1.7
Panel B. Calibrated elasticity of product demand						
Calibrated inputs						
ε			6.0	6.0	6.0	6.0
$d\ln G\left(w_j^I\right)/d\ln w_j^I$			1.8	0.6	1.2	1.2
π			0.61	0.61	0.91	0.31
Model-based outputs						
w_j^I/w_j^m			2.9	1.3	-	-
η			2.7	2.7	-	0.6
$c'(N_j/I_j)/w_j^m$			2.6	0.4	-	-
θ			0.73	0.73	-	0.37

Notes: This table reports estimates of model-based parameters under different calibrations of the retention elasticity and pass-through rate. Panel A allows for different calibrations of the incumbent wage premium w_j^I/w_j^m. Column (1) reports our baseline calibration, where the wage premium is the ratio of the average earnings of incumbent workers to average earnings of recent entrants in the year of application. The baseline retention elasticity $d\ln G\left(w_j^I\right)/d\ln w_j^I$ is estimated in Table D.1and the pass-through rate is estimated in Table 8. Column (2) recalibrates all model-based parameters assuming a lower wage premium of 1.2. Columns (3) and (4) recalibrate assuming the retention elasticity is one standard deviation above and below the baseline, respectively. Columns (5) and (6) re-estimate the model-based parameters assuming the pass-through rate is one standard deviation above and below the baseline, respectively. Panel B holds the baseline elasticity of product demand ε constant and allows the wage premium to vary. Column (1) in Panel B would replicate the baseline results, so it was excluded from this table. Column (2) is not reported in Panel B as it reports the results of calibrating the incumbent wage premium. Dashes denote cases in which the parameter value lies outside a feasible range. See Section 2 for how the parameters are calculated.

References

Bell, Alex, Raj Chetty, Xavier Jaravel, Neviana Petkova, and John Van Reenen, "Who Becomes an Inventor in America? The Importance of Exposure to Innovation," *Quarterly Journal of Economics*, 2019.

Cohen, William, Pradeep Ravikumar, and Stephen Fienburg, "A Comparison of String Distance Metrics for Name-Matching Tasks," in "IIWeb" 2003, pp. 73–78.

Feigenbaum, James, "Automated Census Record Linking: A Machine Learning Approach," 2016. Working paper.

Frakes, Michael and Melissa Wasserman, "Is the Time Allocated to Review Patent Applications Inducing Examiners to Grant Invalid Patents? Evidence from Micro-Level Application Data," *Review of Economics and Statistics*, 2017, *99* (3), 550–563.

Hall, Bronwyn, Adam Jaffe, and Manuel Trajtenberg, "The NBER U.S. Patent Citations Data File: Lessons, Insights, and Methodological Tools," 2001. NBER working paper no. 8498.

Kogan, Leonid, Dimitris Papanikolaou, Amit Seru, and Noah Stoffman, "Technological Innovation, Resource Allocation, and Growth," *Quarterly Journal of Economics*, 2017, *132* (2), 665–712.

Lemley, Mark and Bhaven Sampat, "Is the Patent Office a Rubber Stamp?," *Emory Law Journal*, 2008, *58*, 181–203.

_ and _ , "Examining Patent Examination," *Stanford Technology Law Review*, 2010, (4), 1–11.

Putnam, Jonathan, "The Value of International Patent Rights," 1996. Yale PhD dissertation.

Smith, Matthew, Danny Yagan, Owen Zidar, and Eric Zwick, "Capitalists in the Twenty-First Century," 2019. NBER Working Paper No. 25442.

US Bureau of Economic Analysis, "National Income and Product Accounts Table 1.1.4: Price Indexes for Gross Domestic Product," 2014.

Wooldridge, Jeffrey, *Econometric Analysis of Cross Section and Panel Data*, 2nd ed., Cambridge, MA: MIT Press, 2010.

Yagan, Danny, "Capital Tax Reform and the Real Economy: The Effects of the 2003 Dividend Tax Cut," *American Economic Review*, 2015, *105* (12), 3531–3563.