Online Appendix to "Technological Innovation, Resource Allocation and Growth"

Leonid Kogan, Dimitris Papanikolaou,
Amit Seru and Noah Stoffman

A Patent Data

Our measure of innovation relies on using information on patents that a firm creates and the stock market response to news about these patents. We now discuss the data that we employ in our analysis.

Patents in the United States are granted by the United States Patent and Trademark Office (USPTO). We download the entire history of U.S. patent documents from Google Patents.[1] Each of about 7.8 million patent files was downloaded using an automation script.[2]

To construct our measure of innovation, we match all patents in the Google data to corporations whose returns are in the CRSP database. Patent regulations require that only an individual, not a corporation, can be an inventor. However, the inventor can assign the granted property rights to a corporation or another person. Therefore, when patents are granted they always have an inventor, and sometimes an "assignee", that is, one or more corporations or persons.

For most patents, Google provides a text version of the patent document, created using OCR software. We use this text version of the document to extract the names of corporations to which patents are assigned. However, OCR technology is imperfect, and many of the downloaded documents include a great deal of garbled text. We therefore make use of a number of text analysis algorithms to extract relevant information from the documents.

Our sample covers patents granted between 1926 and 2010 matched to firms with returns in CRSP database. Since we merge our patent data with data on stock returns, we are limited to the period after 1926, when the CRSP database begins.

---

[1] `http://www.google.com/patents`

[2] Google also makes available for downloading bulk patent data files from the USPTO. The bulk data does not have all of the additional "meta" information including classification codes and citation information that Google includes in the individual patent files. Moreover, the quality of the text generated from Optical Character Recognition (OCR) procedures implemented by Google is better in the individual files than in the bulk files provided by the USPTO. As explained below, this is crucial for identifying patent assignees.

Matching patents to firms

Here, we briefly discuss the steps our matching procedure followed, and provide extensive details Section B. We search the document for the words "assignee" or "assigned" and extract the text that immediately follows. This text is either a company name, or the name of an individual to whom the patent is assigned. We then count the number of times each assignee name appears across all patent documents. We compare each assignee name to more common names, and if a given name is "close", in the sense of the Levenshtein distance, to a much more common name, we substitute the common name for the uncommon name.[3] For example, one of the most common names is "General Electric Company", which is associated with over 43,000 patents. We substitute this name for the far less common, but quite similar, names "General Electbic Oohpany", "General Electbic Cqhpany", and "Genebal Electbic Compakt".

At this point, we have an assignee name for each patent. These names must be matched to a company identifier such as the CRSP permco. This is accomplished in two steps. We begin by looking only at patents that are also in the NBER database. For each assignee name identified in the steps above, we count how many different permcos are matched to patents in the NBER database. For example, all of the patents with an assignee name "General Electric Company" are matched to one permco in the NBER database. We can therefore safely assume that *all* of the patents assigned to the General Electric Company can be matched to that permco, *even for patents not included in the NBER data*. Remaining assignee names are matched to CRSP firm names using a name matching algorithm.[4] The algorithm uses a score based on the inverse word frequency to match assignee names to possible company names. For example, the word "American" is quite common in company names, and so contributes little to name matching; the word "Bausch" is quite uncommon, so it is given much more weight. Visual inspection of the matched names confirms very few mistakes in the matching.

Extracting patent citations

We extract patent citations from three sources. First, all citations for patents granted between 1976 and 2011 are contained in text files available for bulk downloading from Google. These citations are simple to extract and likely to be free of errors, as they are official USPTO data. Second, for patents granted before 1976, we extract citations from the OCR text generated from the patent files. We search the text of each patent for any 6- or 7-digit numbers, which could be patent numbers. We then check if these potential patent numbers are followed closely by the corresponding grant date for that patent; if the correct date appears, then we can be certain that we have identified a patent citation. Since we require the date to appear near any potential patent number, it is unlikely that we would incorrectly record a patent citation – it is far more likely that we would fail

[3]The Levenshtein distance is the number of edits required to make one string match another string, where an edit is inserting, deleting, or substituting one character.

[4]The algorithm is based on code written by Jim Bessen, available at `http://goo.gl/m4AdZ`.

to record a citation than record one that isn't there. Third, we complement our citation data with the hand-collected reference data of Nicholas (2008). See Section B of this Appendix for a detailed explanation of this process.

Summary statistics

We now provide some statistics that lend credence to our method for extracting patent information. Table A.1 shows the number of patents we match to companies. Of the 6.2 million patents granted in or after 1926, we find the presence of an assignee in 4.4 million. The matching procedure provides us with a database of 1.9 million matched patents, of which 523,301 (27%) are not included in the NBER data. Figure A.1 graphs the total number of patents matched by the year the patent was granted. Patents included in the NBER data, which is the most comprehensive database previously available, are shown in light shading. Patents unique to our database are presented in dark shading. Note that the two sets of data appear to fit together fairly smoothly, and that even during the period covered by the NBER data, our database adds an average of 2,187 patents to the NBER data.[5]

Table A.2 provides additional summary statistics. Overall, our data provides a matched permco for 66% of all patents with an assignee, or 31% of all granted patents. By comparison, the NBER patent project provides a match for 32% of all patents from 1976–2006, so our matching technique works quite well, even using only data extracted from OCR documents for the period before the NBER data. Another point of comparison is Nicholas (2008), who uses hand-collected patent data covering 1910 to 1939. From 1926–1929, he matches 9,707 patents, while our database includes 8,858 patents; from 1930–1939 he has 32,778 patents while our database includes 47,036 matches during this period.

[5]We use information on the patent-assignee match in the NBER data to assist with our matching, so the match during the overlapping period is mostly the same, by construction. An exception is for cases where there is apparently a mistake in the NBER match and our patent-assignee frequency-based matching system corrects an error.

B Patent Data Construction – Details

In this section we explain in detail how we constructed our new patent data set. The raw data are very large and not very well structured, and thus required a great deal of effort to clean. We used a number of techniques to extract, clean, and match assignees from patents. As with any such project there is a trade-off between type-I and type-II errors (in this case, failing to match an assignee to CRSP or incorrectly matching an assignee to CRSP). Our approach was to be as conservative as possible, attempting to minimize mismatches while at the same time extracting as many correct matches as possible.

B.1 Data sources

We use three sources of data to construct the new patent database:

1. Details of patents granted from 1976–2010 is available in high-quality text files available for bulk downloading from Google, through a special data-hosting arrangement with the United States Patent and Trademark Office (USPTO). The text files use one of two data structures that allows relatively straightforward data extraction: files for 2001–present use XML, while files for 1976–2000 use a fixed-width data structure with labeled fields.
2. Patents granted prior to 1976 are also stored on Google, but only in individual web pages (one per patent). Information during this period is drawn from Optical Character Recognition (OCR) of original patent documents, and is of highly-variable quality. There is very limited, if any, structure to these files.
3. We use the NBER patent data (Hall and Trajtenberg, 2001), which covers the period 1976–2006, to help with the matching and to validate our other data extraction methods.

Due to varying data sources and quality over time, it worth stressing that from 1976–2010 we use the *official records* of the USPTO. As we discuss below, we are able to provide some additions and corrections to the NBER data during the period of overlap with our data. Prior to 1976 the data are more difficult to work with, but we have implemented a number of sophisticated text analysis algorithms to create a very high-quality database.

Downloading individual patent files

We downloaded individual patent data from Google. The URL for each patent's summary page is of the form `http://www.google.com/patents/?id=RD0yAAAAEBAJ`, where `RD0y` is a 4-character code used by Google to identify each patent. The IDs use any of the characters {a, ..., z, A, ..., Z, 0, ..., 9, _, -}. There are $64^4 = 16.8$ million possible IDs, but only about 8 million patents. However, all 16.8 million URLs must be checked, because there is no publicly-available mapping of patent numbers to the Google ID.

A screen shot of the summary page for the patent with id RD0y, which is patent 4,345,262, is shown in Figure 1. The main page includes—when available—the title of the patent, the filing and grant dates, the abstract, inventor(s), original assignee(s), current classifications, and a record of citations (out-cites) and references (in-cites). The information reported on this page by Google was gleaned from the OCR analysis of the original patent document, and consequently less information is reported for older documents, especially patents granted before 1976.

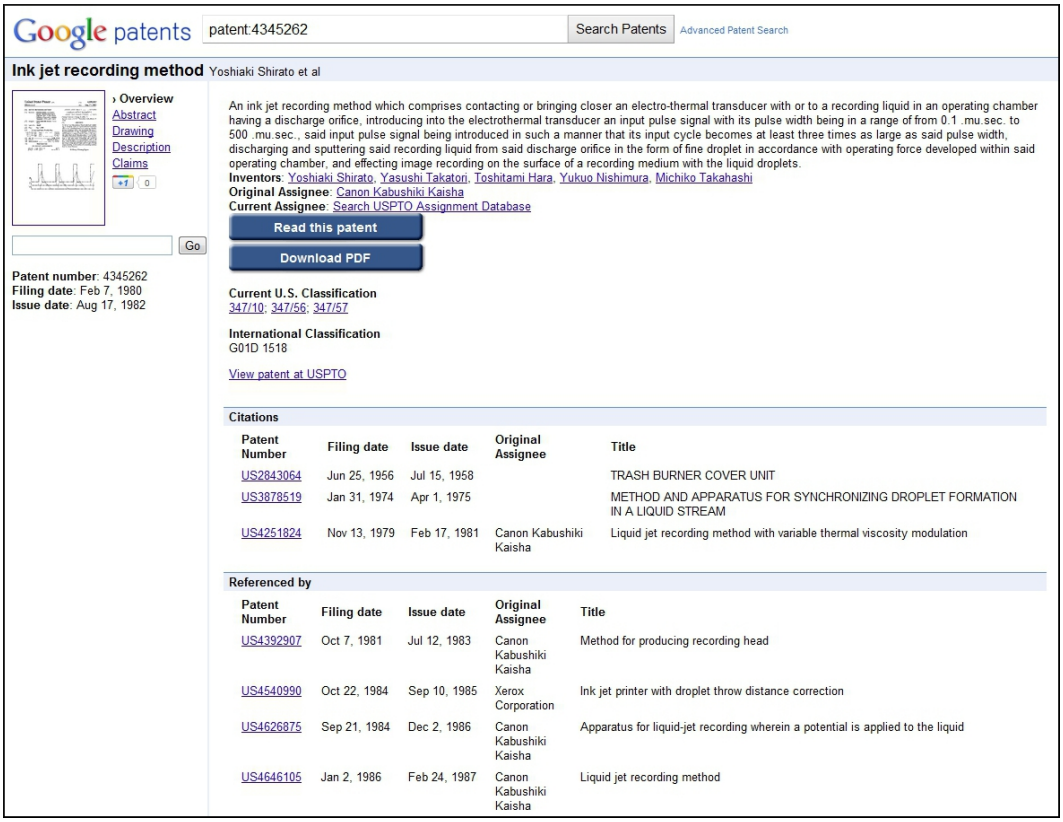


Figure 1: Google summary page for U.S. patent 4,345,262

Using a Perl automation script, we sequentially navigated to each of the 16.8 million patent summary pages.[6] From this page, we stored all available information. The script then loaded the "Read this patent" link, which loads a PDF version of the patent document. From here, we loaded the "plain text" version of the document, which is simply the text derived from OCR of the PDF document. Examples of these pages are shown in Figures 2 and 3. We saved the complete text of the plain text version of each patent. After compression, the complete archive of text requires approximately 56 gigabytes of disk space.

[6]Google generally blocks users from downloading so many web pages. We are grateful to Hal Varian for his assistance with arranging permission to access these pages.

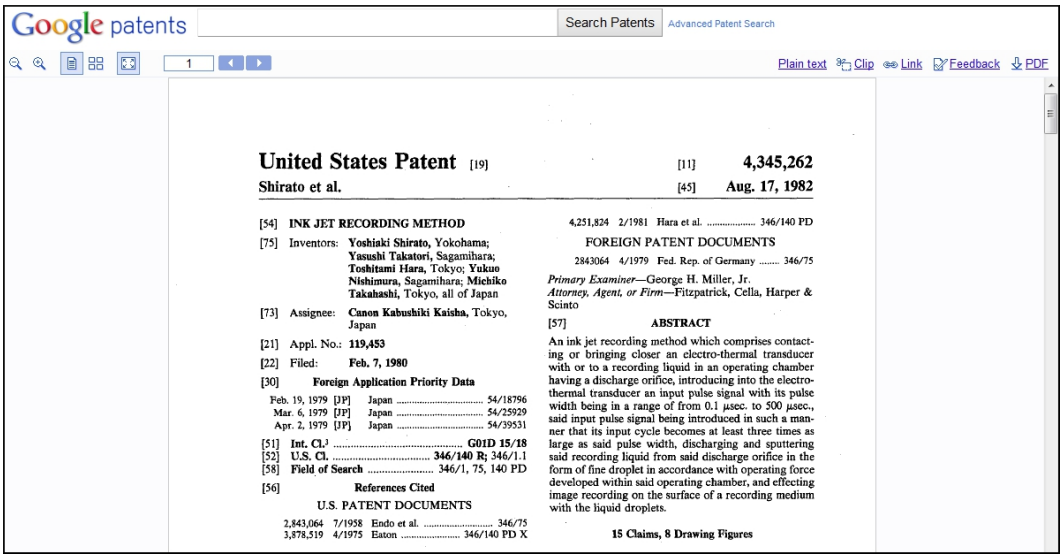


Figure 2: PDF view

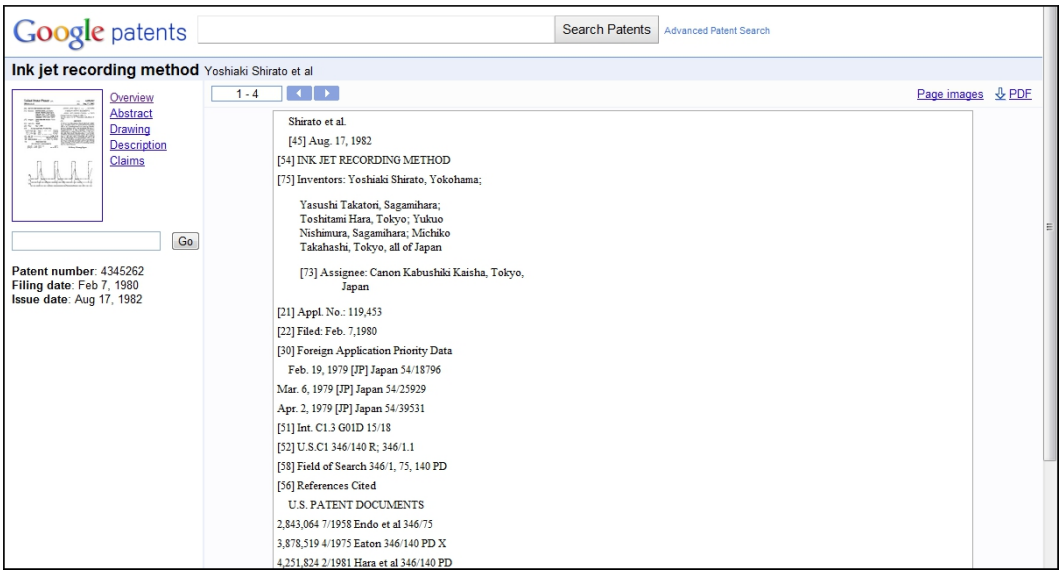


Figure 3: Plain text view

Download bulk patent files

As part of a special arrangement with the USPTO, Google also makes available for downloading bulk patent data files. The bulk data does not have all of the additional "meta" information including classification codes and citation information that Google includes in the individual patent files. Moreover, the quality of the text generated from OCR procedures implemented by Google is better in the individual files than in the bulk files provided by the USPTO. We therefore do not use the bulk download files for data in the pre-NBER period.

For the post-NBER period, however, the bulk data files are of extremely high quality because

they are based on digital patent records as opposed to OCR data drawn from images of patent documents. These data files are provided either in XML format or in a fixed-width record format. In both cases, all fields (inventor name, grant date, etc.) are clearly identified. We rely on these files to construct the database during the post-NBER period (2006–2009) and to make additions and corrections to the NBER data.

B.2 Identifying assignees

Extracting assignee names

For data during the post-1976 period, we can use the XML files available for bulk download to identify the assignee with virtually no errors.

During the pre-1976 period, we cannot rely solely on Google's extraction of the filing and grant dates or the assignee name because the OCR for patents frequently has errors. As an example, consider patent 1,131,249, shown in Figure 4.

UNITED STATES PATENT OFFICE.

EARLE R. KNIGHT, OF NORWOOD, OHIO, ASSIGNOR, BY MESNE ASSIGNMENTS, TO ALLIS-CHALMERS MANUFACTURING COMPANY, A CORPORATION OF DELAWARE.

RETAINING-RING FOR DYNAMO-ELECTRIC MACHINES.

1,131,249. Specification of Letters Patent. Patented Mar. 9, 1915.

Application filed May 1, 1909. Serial No. 492,399.

To all whom it may concern:

Be it known that I, EARLE R. KNIGHT, a citizen of the United States, residing at Norwood, in the county of Hamilton and State of Ohio, have invented certain new and useful Improvements in Retaining-Rings for ring is shown in perspective in Fig. 3. It comprises two segmental duplicate interchangeable parts each part of which has two male portions 16 and two female portions 17. The male and female portions of one part of the split retaining rings are respectively

Figure 4: Title page of patent 1,131,249

It is clear to a human reader that this patent was assigned to the Allis-Chalmers Manufacturing Company, but the OCR for this patent reads

```
EASLS B. KNIGHT, OF NORWOOD, OHIO, ASSIGNOR,, BY MESH'S ASSIGN1IBNTS, TO ALUSCHALME&S
MANOTAC/rURING- COMPANY, A COBPOBAT'lOH OF DELAY/ABE.
```

Consequently, Google records the assignee as "BY MESH S ASSIGNIIBNTS", which is clearly not accurate.

We therefore rely on a number of textual analysis algorithms to extract the assignee name from the full text files we saved for each patent. In general, our approach to performing a "fuzzy" match on a text string is to use the maximum likelihood n-gram approach described by Norvig (2009).

We begin by identifying the text where the assignee, if there is one, will be named. We do this by searching for words that appear similar to "assign", "assignor", or "assignee". When found near the beginning of the patent document, this word is typically followed closely by the name of the assignee, so we extract a text string of 200 characters for further processing. The assignee may be a person, or a corporation, in which case the name will include a word like "company", "corporation" or "incorporated". If the word "assign" and its variants are not found, we assume the inventor did not assign the patent to another entity.

Cleaning assignee names

After extracting the string that is likely to contain the assignee name, additional cleaning is necessary. Because of OCR errors, company names may be garbled. For example, the General Electric Company, which has more than 43,000 patents in our data, appears as "General Electbic Oohpany", "General Electbic Cqhpany", and "Genebal Electbic Compakt", among hundreds of other misspellings. To fix these, we first count how many patents have been granted to each assignee name, regardless of how the assignee name is spelled. In this example, General Electric Company appears in 42,693 patents, while each of the misspelled variants appears fewer than 5 times.

We then calculate the Levenshtein edit distance[7] between each assignee name and all other names that have more patents. If any assignee name is close to another assignee name that is associated with many more patents, then the more common assignee name is substituted for the less common name. This algorithm correctly identifies all of the misspellings noted above as being General Electric.

After cleaning assignee names, we manually checked which misspelled names were matched to the 500 assignees with the most patents to confirm that no significant errors were introduced in this step.

Matching to CRSP

Having extract a list of assignee names, the next step is to match company names to the CRSP permco identifier. This is accomplished in three steps.

We begin by looking only at those patents that are included in the NBER patent database.

[7]The Levenshtein distance is the number of edits required to transform one string into another string, where allowed edits are inserting, deleting, or substituting one character. For example, the Levenshtein distance between "patent" and "parent" is 1, while the distance between "patent" and "apparent" is 3.

For each assignee name identified in the steps above, we count how many *different* permcos are matched to patents in the NBER database. For example, all of the patents with an assignee name "General Electric Company" are matched to one permco in the NBER database. We can therefore safely assume that *all* of the patents assigned to the General Electric Company can be matched to that permco, *even for patents not included in the NBER data.* This step allows us to draw on the extensive data cleaning and matching project undertaken by Hall and Trajtenberg (2001) while at the same time identifying some errors in the NBER database. For example, patent 4,994,660 was assigned to General Electric but is identified in the NBER data as being assigned to Hitachi, Ltd. Because our algorithm relies on name matching, and the assignee name in that patent is General Electric, the patent is correctly identified in our data.

The first step only helps us match assignees with patenting activity during the period covered by the NBER database. We therefore proceed with a second step to match remaining assignee names. We do this with a name matching algorithm based on code written by Jim Bessen, available at `http://goo.gl/m4AdZ`. The algorithm uses a score based on the inverse word frequency to match assignee names to possible company names. For example, the word "American" is quite common in company names, and so contributes little to name matching; the word "Bausch" is quite uncommon, so it is given much more weight. Visual inspection of the matched names confirms very few mistakes in the matching.

Finally, we identify the top 250 assignees (by patents) with no CRSP matches. We manually matched these to CRSP whenever possible. Examples of firms requiring manual matching include research subsidiaries such as 3M Innovative Properties Company, which was not successfully matched to CRSP because its name differs substantially from its parent. Although we only checked 250 assignees, this manual check allowed us to match an additional 64,000 patents. Firms with high patenting activity but not matched to CRSP are either private companies or foreign firms that are not listed on U.S. exchanges, an example of which is Hoffmann-La Roche, the large Swiss drug company.

B.3 Correcting grant dates

The filing and grant dates of the patents are subject to the same sort of OCR errors as the assignee information. The grant dates are particularly important for our purposes because we use them to calculate the return around the grant date. Since patent numbers are sequential by grant dates, it

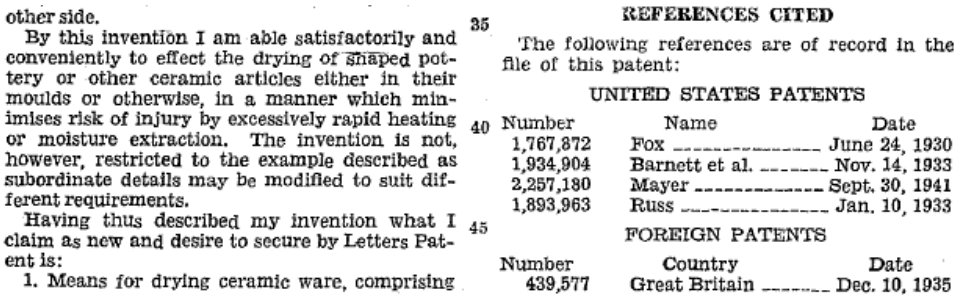
is easy to infer missing or incorrect grant dates by comparing patent dates to the grant dates of adjacent patents. The same is not true of filing dates, but do not use filing dates in our current work.

To populate missing patent dates and correct mistakes we identify the 3 non-missing grant dates immediately preceding and following each patent. For example, if patent k's grant date is missing but patents ($k-3$,...,$k-1$,$k+1$,...,$k+3$) have grant date D, then we set patent k's grant date to D. By applying this procedure iteratively we are able to correct most grant dates, with the exception of patents whose grant dates are missing and lie at a boundary between two grant dates. We fill in these missing boundary dates by manually checking their grant dates on the USPTO's web site.

While we don't rely on filing dates in the paper, it is possible to correct large errors in filing dates by identifying cases where filing dates occur after the grant date, or much earlier than the filing dates of adjacent patents. These errors often occur only in the year, so we can keep the recorded month and day the same while setting the year of the patent filing to the median filing year of a 20-patent window centered on a patent with an apparent error.

B.4 Extracting citations

Extracting patent citations from the patent text documents presents another challenge. The format of a patent document has changed several times, as has the location and formatting of citations within the document. For example, Figure 5 shows the references section of patent 2,423,030, granted in 1947. The format seen here is the first format used after patent citation began in February, 1947.



other side.

By this invention I am able satisfactorily and conveniently to effect the drying of shaped pottery or other ceramic articles either in their moulds or otherwise, in a manner which minimises risk of injury by excessively rapid heating or moisture extraction. The invention is not, however, restricted to the example described as subordinate details may be modified to suit different requirements.

Having thus described my invention what I claim as new and desire to secure by Letters Patent is:

1. Means for drying ceramic ware, comprising

REFERENCES CITED

The following references are of record in the file of this patent:

UNITED STATES PATENTS

Number	Name	Date
1,767,872	Fox	June 24, 1930
1,934,904	Barnett et al.	Nov. 14, 1933
2,257,180	Mayer	Sept. 30, 1941
1,893,963	Russ	Jan. 10, 1933

FOREIGN PATENTS

Number	Country	Date
439,577	Great Britain	Dec. 10, 1935

Figure 5: A patent citation section

A human reader has no problem identifying the citations in this patent. But to understand the considerable challenge faced in automating this identification, consider the OCR for this part of the patent:

```
other side. 35 REFERENCES CITED
By this invention I am able satisfactorily and The following references are of record in the
conveniently to effect the drying of'Shaped pot- jjle of tllis patent:
tery or other ceramic articles either in their
moulds or otherwise, in a manner which min- UNITED STATES PATENTS
imises risk of injury by excessively rapid heating 40 Number Name Date
or moisture extraction. The invention is not, 1,767,872 Pox June 24, 1930
however, restricted to the example described as 1^934,904 Barnett et al Nov. 14', 1933
subordinate details may be modified to suit dif- 2,257,180 Mayer Sept. 30, 1941
ferent requirements. 1,893,963 Russ Jan. 10,1933
Having thus described my invention what I 45
claim as new and desire to secure by Letters Pat- * ("uu-^ f A 1 Jun 11>
entis: Number Country Date
1. Means for drying ceramic ware, comprising 439,577 Great Britain Dec. 10,1935
```

Our approach is to identify any text that could be a patent number (a 6- or 7-digit number, perhaps separated by commas, spaces, or other "noise" characters) and is closely followed by the correct grant date for the cited patent. In particular, for every potential patent number we identify, we determine its grant date and then search near the possible citation for that date. If the date appears, we can be very confident that we have correctly identified a citation. For example, for the patent shown in Figure 5 we extract the patent number 1,767,872 and then confirm that its grant date—June 24, 1930—appears somewhere nearby in the text. By using this two-step process to identify citations, our citation extraction is of very high quality—the probability that some random 7-digit number will be followed closely by the correct date is clearly extremely small.

Our citation extraction method provides more citations than what is available on the Google summary page. For example, the Google summary page for the patent shown in the previous example provides no citations at all, while our algorithm correctly extracted all four citations. (We exclude citations to foreign patents, as these patents are not in our database.) In general, Google does not currently report out-cites from patents granted before 1976, so we use this extraction method on all patents granted between 1926 and 1975.

B.5 Data validation

As previously mentioned, any data extraction project such as this can lead to two types of errors: matching a patent to a firm that is not the assignee, or failing to match a patent to a any firm when it does have an assignee. Our strategy makes the first error very unlikely, as a match occurs only

when a name closely resembling a CRSP company name appears around the word "assignee" at the beginning of patent document. We cannot be sure how many errors of the second type we made, but we have taken care to ensure that our algorithms allow as flexible matching as possible.

We also did two final checks to check the quality of our matching strategy. First, we visually inspected a random sample of 500 patents granted between 1926 and 1975 and confirmed that assignees had been correctly extracted, and correctly matched if the assignee appeared in CRSP. This is obviously a very small sample of patents, but this careful check confirmed that no serious errors existed.

Second, we applied the extraction and matching algorithms we used in the pre-1976 period to a random sample of 25,000 patents granted between 1976 and 1999. We then compared our matches to the matches in the NBER data. None of our matches was incorrect, and only 3 patents were incorrectly not matched to an assignee. In other words, applying the techniques we used on pre-1976 data to data from the NBER period yields results that are virtually identical to those in the NBER database.

C Additional Results and Descriptive Statistics

C.1 Alternative Distributional Assumptions

Here, we briefly describe how we allowed for different distributional assumptions to construct the filtered value of a patent.

Allowing for a non-zero mean

We now assume that patent value is drawn from a non-zero mean distribution, i.e., $v_j \sim N(\mu, {\sigma_{vj}}^2)$, truncated at zero. In this case, the filtered value of v_j as a function of the stock return is equal to

$$E[v_j|r_j^l] = (1-\delta_j)\mu_j + \delta_j\, r_j^l + \sqrt{\delta_j}\,\sigma_{\xi j}\,\frac{\phi(R_j)}{1-\Phi(R_j)},$$

where ϕ and Φ are the standard normal pdf and cdf, respectively, and R and δ are the normalized return and the signal-to-noise ratio respectively,

$$R_j = -\frac{(1-\delta_j)\mu_j + \delta_j\, r_j^l}{\sqrt{\delta_j}\sigma_{\xi j}},$$

and δ_j is the signal-to-noise ratio as defined in the paper.

Relative to the paper, however, the additional issue we need to worry about is how to estimate μ_j. Following the same procedure as the paper to allow the mean μ to vary by firm and year is not possible, since at higher frequencies the standard deviation term dominates the mean. Hence, we will need additional assumptions to recover μ_j. We will do so in two ways. In both methods it is useful to note that the average stock return on patent announcement dates equals

$$E[r_j] = E[v_j] = \mu_j + 0.7979\,\sigma_{vj}$$

1. The first way assumes that μ is constant across firm-years. In this case, we estimate μ_j by first subtracting $0.7979\,\sigma_{vft}$ from firm returns on patent announcements, and then estimate μ_j as the difference in average returns on announcement versus non-announcement days using the full sample. Our point estimates imply a $\mu_j = -0.6\%$. We use this estimate when forming the conditional value of a patent as outlined above.
2. Alternatively, we assume that $\mu + 0.7979\,\sigma_v$ is constant across firm-years. That is, we allow

μ to vary exactly with σ_v. In this case, we estimate the difference in average returns on announcement versus non-announcement days using the full sample. The point estimate is 0.02%. We then recover our estimate $\mu_j = 0.0002 - 0.7979\,\sigma_{vj}$.

We find that in both of the cases above, our results are very similar to the benchmark case where we assume $\mu = 0$. Specifically, the correlation between the filtered value of the patent $E[v_j|r_j^l]$ constructed using different assumptions on μ in (a) and (b) above with our benchmark values are in excess of 99%. Indeed, allowing the pre-truncation mean to vary has mostly a scaling effect on our estimates. Since the results are essentially extremely similar to the paper, we do not reproduce them here.

Exponential

As before, we assume that the stock return of the firm on the patent grant date is given by

$$r = v + \varepsilon.$$

We now assume that v is exponentially distributed with parameter $1/\sigma_v$. As in the paper, we still assume the noise term is normally distributed, $\varepsilon \sim N(0, \sigma_\varepsilon^2)$. Under these assumptions, we can solve for the conditional expectation – equivalent of equation (4) in the paper – in closed form,

$$E[v|R] = R + \sigma_\varepsilon \left(\sqrt{\frac{2}{\pi}} \frac{\exp(-\tilde{R}^2/2)}{G^c(\tilde{R}/\sqrt{2})} - \frac{\sigma_\varepsilon}{\sigma_v} \right)$$

where G^c is the complementary error function and

$$\tilde{R} = \frac{\sigma_\varepsilon}{\sigma_v} - \frac{r}{\sigma_\varepsilon}.$$

As in the paper, we assume that the ratio $\sigma_v^2/\sigma_\varepsilon^2$ is constant across firms. We use our estimates from section 1.4, which imply $\sigma_v^2/\sigma_\varepsilon^2 = 0.0142$, so we use that as our baseline case.

Cauchy

As before, we assume that the stock return of the firm on the patent grant date is given by

$$r = v + \varepsilon.$$

where now v is distributed according to the positive part of a Cauchy distribution centered at zero with scale parameter γ_v and ε follows a Cauchy distribution centered at zero with scale parameter γ_ε. In this case, r on announcement date is also Cauchy with scale $\gamma_v + \gamma_\varepsilon$. Under the assumption that both ε and v are Cauchy distributed, the conditional value of the patent is now given by:

$$E(v|r) = \frac{\left(\gamma_\varepsilon \left(c(r) - {\gamma_v}^2\right) \ln\left(c(r)\right) + 2\, r \left(c(r) + {\gamma_v}^2\right) \arctan\left(\frac{r}{\gamma_\varepsilon}\right) - 2\, \gamma_\varepsilon \left(c(r) - {\gamma_v}^2\right) \ln\left(\gamma_v\right) + r\pi \left(r^2 + (\gamma_\varepsilon - \gamma_v)^2\right)\right) \gamma_v}{\left(2\, \gamma_\varepsilon\, \gamma_v\, r \ln\left(c(r)\right) + 2\, \gamma_v \left({\gamma_v}^2 - \gamma_\varepsilon^2 + r^2\right) \arctan\left(\frac{r}{\gamma_\varepsilon}\right) - 4\, \gamma_\varepsilon\, \gamma_v\, r \ln\left(\gamma_v\right) + \pi \left(\gamma_v + \gamma_\varepsilon\right) \left(r^2 + (\gamma_\varepsilon - \gamma_v)^2\right)\right)}$$

where

$$c(r) = r^2 + \gamma_\varepsilon^2.$$

Since the second moments of the Cauchy distribution do not exist, we need alternative ways of estimating its parameters than what we used in the text. We estimate the scale of the noise term, γ_ε, using one-half the interquartile range of firm-year idiosyncratic returns, with an adjustment similar to the equation in footnote 14 in the paper. Regarding the estimation of the noise-to-signal ratio $\tilde{\delta} = \gamma_v/(\gamma_v + \gamma_\varepsilon)$, we can no longer estimate it using equation (17) in the paper. Absent a different alternative, we use the same estimates as in the paper implying a $\tilde{\delta} = 0.014$.

Table A.1: Number of patents

Data step	Number of patents
Total downloaded patents	7,797,506
Granted in 1926 or later	6,272,428
Identified as having an assignee	4,374,524
Matched to CRSP	1,928,123
Of which:	
Present in NBER data	1,404,822
New to this paper	523,301

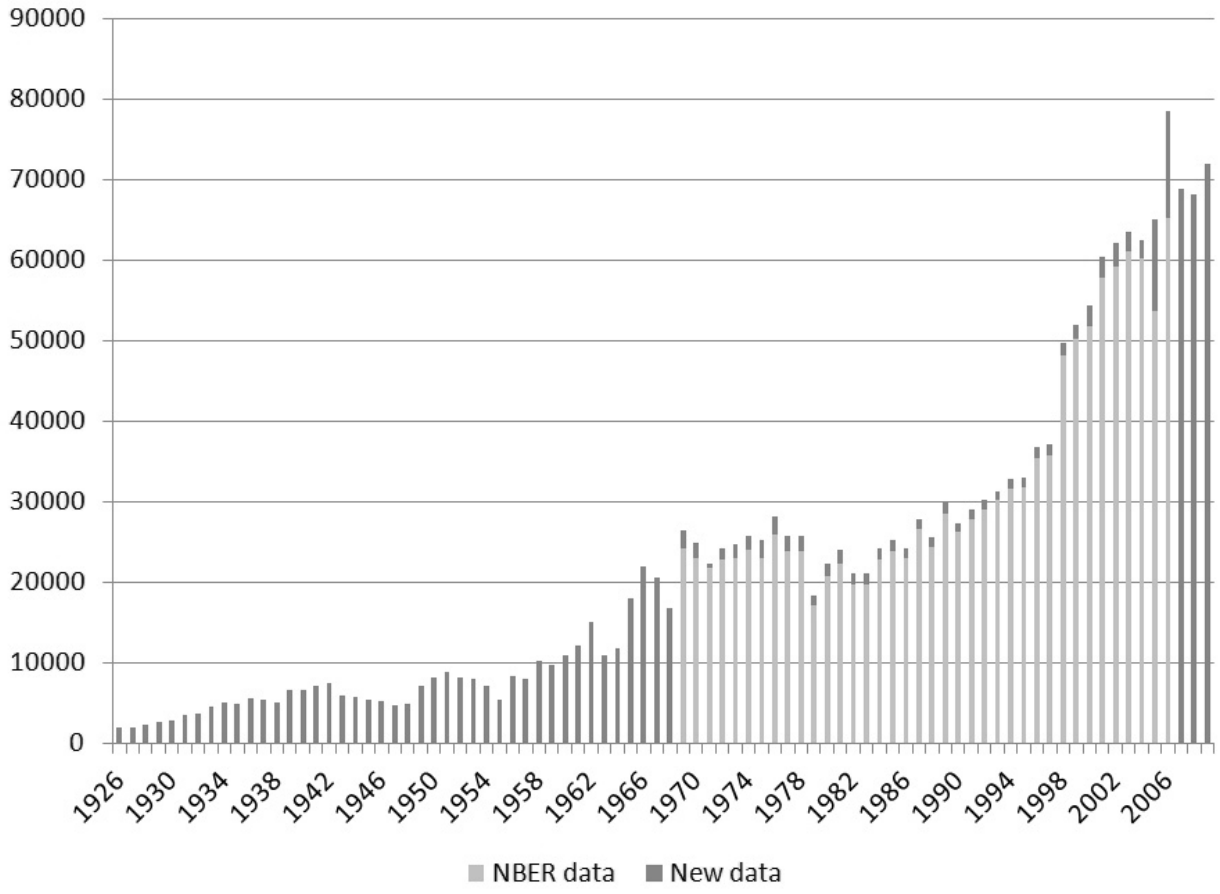
The table provides details on patents in our sample discussed in Section II.A of the paper. We begin with all patents downloaded from Google Patents, and restrict the sample to post-1926. Not all patents have assignees, and among those that do, not all are companies in CRSP. We are able to match 1,928,123 patents to CRSP firms, of which 523,301 (27%) are new to this study. Further details are reported in Table 2 and Figure 1. In the paper, we restrict attention to the patents that have a unique assignee, patents for which we have non-missing data on three day announcement return, market capitalization and return volatilities needed to compute our $\hat{\Theta}$ measure. The sample contains 1,801,879 patents.

Table A.2: Assignee matching by Decade

	Number of patents			Number of unique	
Years	Total	With assignee	Matched to CRSP	Matched firms	CRSP firms
1926–1929	174,022	48,433	8,858	182	786
1930–1939	442,700	172,925	47,029	355	951
1940–1949	307,499	141,345	60,616	451	1,042
1950–1959	425,953	171,157	82,255	587	1,246
1960–1969	567,599	265,524	165,409	1,175	3,177
1970–1979	690,459	393,661	247,102	2,086	7,204
1980–1989	708,735	579,518	235,525	2,756	11,715
1990–1999	1,109,398	933,705	352,005	3,664	14,882
2000–2010	1,846,063	1,668,256	729,324	4,415	11,900
All years	6,272,428	4,374,524	1,928,123	7,864	26,660

The table shows summary statistics for patents in our sample by decade discussed in Section II.A of the paper. Column 2 shows the total number of patents, and column 3 shows how many patents are identified as having an assignee. Column 4 shows how many of those patents with assignees are matched to a company in CRSP. (The remaining assignees are either individuals, private companies, or the matching process was unable to identify the correct company.) Columns 5 and 6 show how many unique firms there are matched to patents or in CRSP.

Figure A.1: Number of Patents with Matched Assignees



The figure shows the number of patents matched to CRSP firms by year of patent grant. Light shading denotes patents included in the NBER patent data set, while dark shading denotes patents that are new in our paper.

Table A.3: Innovation and Firm Size

Size (book assets)	**1**	**2**	**3**	**4**	**5**
Patents, citation-weighted (Θ^{cw})	1.2	2.3	3.9	8.2	90.4
Citations to Patents	2.6	2.6	2.5	2.3	2.2
Patents, citation-weighted, scaled by assets (%)	6.6	4.4	3.0	2.5	2.4
Patents, citation-weighted, scaled by mkcap (%)	7.9	6.5	5.8	5.4	22.5
Patents, SM weighted (Θ^{sm})	0.3	1.2	3.5	15.5	603.8
Total Value to Number of Patents	0.6	1.1	2.0	4.3	18.1
Patents, SM weighted, scaled by assets (%)	3.5	3.3	3.5	4.7	10.6
Patents, SM weighted, scaled by mkcap (%)	1.8	2.4	2.8	3.9	12.3
Size (Market cap of equity)	**1**	**2**	**3**	**4**	**5**
Patents, citation-weighted (Θ^{cw})	1.3	3.4	6.0	14.9	81.4
Citations to Patents	2.2	2.4	2.5	2.5	2.3
Patents, citation-weighted, scaled by assets (%)	4.0	4.4	4.0	3.4	3.1
Patents, citation-weighted, scaled by mkcap (%)	21.9	9.8	7.2	6.0	3.4
Patents, SM weighted (Θ^{sm})	0.1	0.8	2.2	9.3	618.4
Total Value to Number of Patents	0.3	0.6	1.2	2.7	19.1
Patents, SM weighted, scaled by assets (%)	1.2	2.4	3.5	4.7	13.9
Patents, SM weighted, scaled by mkcap (%)	2.4	2.9	3.1	4.3	10.6

Table reports mean value within each quintile. SM values are deflated by CPI (units are USDm in 1982). Quintiles are computed using annual breakpoints. Citation-weighted patent counts are computed as $\sum_j 1 + C_j/\bar{C}_j$ where C_j is number of cites to patent j and $\bar{C}_j$ is the mean number of cites to patents granted in the same year as patent j.

Table A.4: Firm-level innovation measure: changes in distribution across decades

Decade	Mean	Sd	p25	p50	p75	p90	p95	p99
1950	3.1	6.3	0.0	0.4	3.1	9.4	16.2	32.7
1960	4.7	10.0	0.0	0.0	4.7	14.4	23.8	51.6
1970	1.7	5.3	0.0	0.0	0.7	4.3	9.3	30.4
1980	1.2	4.0	0.0	0.0	0.0	3.1	8.1	21.7
1990	3.0	11.1	0.0	0.0	0.0	6.7	17.9	56.0
2000	5.6	19.2	0.0	0.0	1.5	14.6	32.6	86.5

Table reports the distribution of our baseline measure θ_f^{sm} across decades. Units are in percentage terms.

Table A.5: Mean innovation across industries

Ind Code	Industry Name	θ^{cw}	θ^{sm}
1	Food Products	0.76	0.98
2	Beer & Liquor	0.16	2.10
3	Tobacco Products	0.32	1.59
4	Recreation	2.19	1.39
5	Printing and Publishing	0.66	0.18
6	Consumer Goods	4.02	3.48
7	Apparel	0.51	0.22
8	Healthcare, Medical Equipment, Pharmaceutical Products	9.09	9.13
9	Chemicals	6.97	5.93
10	Textiles	1.05	0.33
11	Construction and Construction Materials	2.50	1.20
12	Steel Works Etc	1.78	1.38
13	Fabricated Products and Machinery	6.67	3.66
14	Electrical Equipment	8.09	4.58
15	Automobiles and Trucks	4.67	2.72
16	Aircraft, ships, and railroad equipment	6.22	3.85
17	Precious Metals, Non-Metallic, and Industrial Metal Mining	0.52	0.32
18	Coal	0.23	0.09
19	Petroleum and Natural Gas	0.72	1.43
21	Communication	0.41	0.67
22	Personal and Business Services	2.15	2.25
23	Business Equipment	7.45	7.19
24	Business Supplies and Shipping Containers	2.78	2.39
25	Transportation	0.05	0.05
26	Wholesale	0.42	0.24
27	Retail	0.13	0.12
28	Restaurants, Hotels, Motels	0.05	0.03

Table reports mean value of normalized firm-level innovation θ (multiplied by 100) within each Fama-French industry (using their 30 industry classification). We exclude financial firms and utilities.

Table A.6: Estimates of Patent Value: Descriptive Statistics

Moment	C	$C/\bar{C}$	R_f	Baseline		Exponential		Cauchy	
				$E[v\|R_f]$	ξ	$E[v\|R_f]$	ξ	$E[v\|R_f]$	ξ
			(%)	(%)	USDm	(%)	USDm	(%)	USDm
Mean	10.26	1.18	0.07	0.32	10.36	0.40	12.79	0.15	5.13
Std. Dev	20.13	1.98	3.92	0.20	32.04	0.30	39.75	0.11	16.13
Percentiles									
p1	0	0.00	-9.93	0.11	0.01	0.13	0.01	0.05	0.00
p5	0	0.00	-5.15	0.14	0.04	0.16	0.05	0.06	0.02
p10	0	0.00	-3.55	0.16	0.11	0.19	0.13	0.07	0.05
p25	1	0.20	-1.67	0.20	0.73	0.24	0.89	0.09	0.33
p50	5	0.62	-0.09	0.27	3.22	0.33	3.95	0.13	1.52
p75	11	1.38	1.62	0.37	9.09	0.46	11.23	0.18	4.45
p90	24	2.78	3.82	0.53	22.09	0.66	27.28	0.27	10.95
p95	38	4.06	5.73	0.68	38.20	0.85	47.27	0.34	19.13
p99	90	8.84	11.49	1.07	121.39	1.35	150.46	0.55	60.04

The table reports the distribution of the following variables across the patents in our sample: the number of future citations till the end of our sample period C; the number of citations scaled by the mean number of cites to patents issued in the same year $\bar{C}$; the market-adjusted firm returns R_f on the 3-day window around patent grant dates; the filtered component of returns $E[v|R_f]$ related to the value of innovation – using equation (4); and the filtered dollar value of innovation ξ using equation (3) deflated to 1982 (million) dollars using the CPI. In addition to the baseline case, we also report results using two alternative distributional assumptions. First, we assume that the component of firm return due to the patent, v, is exponentially distributed with scale parameter $1/\sigma_v$. As before, we assume the signal-to-noise ratio is constant across firms; using our estimates from equation (6) in the paper, we obtain $\sigma_v/\sigma_\varepsilon \approx 0.014$, so we use that. As before, we allow σ_ε to vary by firm-year and follow the same exact procedure as in the baseline case. Second, we assume that v is distributed according to a Cauchy truncated at zero with scale γ_v, while ε is distributed according to a Cauchy with parameter γ_ε. We estimate the scale of the noise term, γ_ε, using one-half the interquartile range of firm-year idiosyncratic returns, with an adjustment similar to the equation in footnote 13 in the paper. Regarding the estimation of the noise-to-signal ratio $\delta = \gamma_v/(\gamma_v + \gamma_\varepsilon)$, we can no longer estimate it using equation (6) in the paper because the variance of the Cauchy distribution does not exist. Absent a different alternative, we use the same estimate as in the paper. We restrict attention to the patents for which we have non-missing data on three day announcement return, market capitalization and return volatilities needed to compute our $\hat{\Theta}$ measure. The sample contains 1,801,879 patents.

Table A.7: Forward Citations and Patent Market Values – Alternative Distributions

	(1)	(2)	(3)	(4)	(5)
			A. Exponential		
$\log(1+C_j)$	0.174	0.099	0.055	0.013	0.004
	(9.99)	(9.44)	(10.28)	(13.84)	(5.12)
			B. Cauchy		
$\log(1+C_j)$	0.173	0.096	0.059	0.016	0.004
	(10.25)	(9.49)	(10.35)	(12.86)	(4.84)
Controls					
Firm Market Capitalization	-	Y	Y	Y	Y
Volatility	-	-	Y	Y	-
Fixed Effects	TxC	TxC	TxC	TxC	TxC
				F	FxT

Table reports the equivalent of Table 2 in the paper under two alternative distributional assumptions. Panel A presents results under the assumption that the component of firm return due to the patent, V, is exponentially distributed with scale parameter $1/\sigma_v$. Panel B presents results under the assumption that v is distributed according to a Cauchy truncated at zero with scale γ_v, while ε is distributed according to a Cauchy with parameter γ_ε. See notes to Table A.6 for more details.

Table A.8: Innovation and Firm Profit Growth – Alternative Distributions

A. Exponential									
Firm					Competitors				
1	2	3	4	5	1	2	3	4	5
0.018	0.029	0.036	0.042	0.046	-0.015	-0.029	-0.032	-0.035	-0.038
[3.61]	[4.49]	[3.74]	[3.81]	[3.60]	[-2.96]	[-5.05]	[-7.25]	[-5.98]	[-5.84]
B. Cauchy									
Firm					Competitors				
1	2	3	4	5	1	2	3	4	5
0.018	0.027	0.035	0.040	0.045	-0.012	-0.024	-0.027	-0.031	-0.034
[5.74]	[5.95]	[5.18]	[4.98]	[4.97]	[-2.19]	[-3.40]	[-4.95]	[-5.05]	[-4.81]

Table reports the equivalent of Table 4, Panel A in the paper under two alternative distributional assumptions. Panel A presents results under the assumption that the component of firm return due to the patent, V, is exponentially distributed with scale parameter $1/\sigma_v$. Panel B presents results under the assumption that v is distributed according to a Cauchy truncated at zero with scale γ_v, while ε is distributed according to a Cauchy with parameter γ_ε. See notes to Table A.6 for more details.

Table A.9: Innovation and Firm Growth: Results using Alternative Scaling (Market Capitalization)

a. Profits										
Firm					Competitors					
1	2	3	4	5	1	2	3	4	5
0.006	0.017	0.023	0.028	0.034	-0.027	-0.034	-0.038	-0.043	-0.043
[2.45]	[5.53]	[4.34]	[4.21]	[4.07]	[-5.64]	[-3.29]	[-4.14]	[-4.33]	[-4.37]
b. Output										
Firm					Competitors					
1	2	3	4	5	1	2	3	4	5
-0.003	0.001	0.003	0.012	0.021	-0.041	-0.051	-0.058	-0.056	-0.064
[-1.34]	[0.27]	[0.54]	[1.60]	[2.23]	[-6.23]	[-3.52]	[-3.65]	[-3.47]	[-3.84]
c. Capital										
Firm					Competitors					
1	2	3	4	5	1	2	3	4	5
0.003	0.007	0.011	0.015	0.021	-0.013	-0.027	-0.039	-0.050	-0.062
[1.74]	[2.19]	[2.39]	[2.58]	[2.74]	[-3.33]	[-4.94]	[-5.63]	[-6.22]	[-6.71]
d. Labor										
Firm					Competitors					
1	2	3	4	5	1	2	3	4	5
-0.001	0.002	0.006	0.011	0.014	-0.019	-0.026	-0.032	-0.033	-0.032
[-0.61]	[0.79]	[1.53]	[1.92]	[1.99]	[-6.19]	[-4.61]	[-5.29]	[-4.40]	[-4.23]
e. TFPR										
Firm					Competitors					
1	2	3	4	5	1	2	3	4	5
0.003	0.010	0.012	0.016	0.017	-0.005	-0.009	-0.013	-0.013	-0.013
[1.09]	[3.17]	[2.74]	[4.22]	[4.74]	[-2.62]	[-3.59]	[-4.53]	[-3.09]	[-2.58]

Table repeats the analysis in Table 4 in the paper. Rather than book assets, we now scale the firm's dollar value of innovation by its end of year market capitalization. Similarly, innovation by competing firms is constructed as the dollar value of innovation divided by their total market capitalization, in a manner analogous to equation (11) in the paper. See notes to Table 4 in the paper for more details.

Table A.10: Innovation and Firm Profit Growth – Patent citations measured within a fixed window

A. Citations within 3-years of patent grant									
Firm					Competitors				
1	2	3	4	5	1	2	3	4	5
0.007	0.012	0.017	0.021	0.025	-0.005	-0.006	-0.007	-0.006	-0.006
[4.41]	[4.92]	[5.08]	[4.99]	[5.31]	[-1.85]	[-1.56]	[-1.76]	[-1.13]	[-1.06]
B. Citations within 5-years of patent grant									
Firm					Competitors				
1	2	3	4	5	1	2	3	4	5
0.008	0.014	0.019	0.024	0.027	-0.005	-0.007	-0.009	-0.007	-0.007
[4.73]	[5.58]	[5.35]	[5.35]	[5.71]	[-1.69]	[-1.78]	[-2.11]	[-1.43]	[-1.24]
C. Citations within 10-years of patent grant									
Firm					Competitors				
1	2	3	4	5	1	2	3	4	5
0.008	0.015	0.022	0.027	0.032	-0.003	-0.005	-0.009	-0.007	-0.007
[4.86]	[5.77]	[5.62]	[5.69]	[6.59]	[-0.89]	[-1.29]	[-2.11]	[-1.28]	[-1.23]

Table reports the equivalent of Table 5, Panel A in the paper under different ways of adjusting patent citations for truncation lags. In each of the panels A, B, and C, we measure forward citations over the first N years after the patent is issued, where $N = 3, 5, 10$. We then repeat the analysis in Table 5 by also excluding the last N years from the sample. See notes to Table 5 in the paper for more details.

Table A.11: Innovation and Firm Growth: Controlling for R&D spending of Firm and Competitors

Profits									
Firm					a. Competitors				
1	2	3	4	5	1	2	3	4	5
0.017	0.027	0.034	0.039	0.043	-0.018	-0.034	-0.037	-0.040	-0.045
[3.33]	[4.12]	[3.39]	[3.48]	[3.31]	[-3.24]	[-5.42]	[-8.23]	[-6.36]	[-6.38]
b. Output									
Firm					Competitors				
1	2	3	4	5	1	2	3	4	5
0.008	0.013	0.018	0.022	0.028	-0.016	-0.034	-0.044	-0.048	-0.056
[2.85]	[3.07]	[2.85]	[2.62]	[3.15]	[-3.64]	[-6.33]	[-8.43]	[-7.91]	[-7.98]
c. Capital									
Firm					Competitors				
1	2	3	4	5	1	2	3	4	5
0.010	0.021	0.028	0.034	0.039	0.002	-0.005	-0.012	-0.020	-0.029
[8.40]	[6.69]	[5.76]	[4.40]	[4.13]	[0.38]	[-0.82]	[-1.59]	[-2.35]	[-3.15]
d. Labor									
Firm					Competitors				
1	2	3	4	5	1	2	3	4	5
0.007	0.014	0.019	0.023	0.025	-0.007	-0.017	-0.021	-0.021	-0.019
[5.65]	[4.39]	[4.11]	[3.76]	[3.30]	[-1.64]	[-3.86]	[-4.34]	[-3.67]	[-3.00]
e. TFPR									
Firm					Competitors				
1	2	3	4	5	1	2	3	4	5
0.012	0.015	0.017	0.021	0.022	-0.001	-0.006	-0.010	-0.015	-0.017
[2.18]	[2.07]	[2.56]	[3.28]	[3.98]	[-0.52]	[-2.20]	[-3.20]	[-4.59]	[-4.09]

Table repeats the analysis of Table 4 in the paper including the firm's R&D spending as an additional control. We control for the firm's ratio of R&D spending to sales, as well as the ratio of total R&D spending to total sales of competing firms. See notes to Table 4 in the paper for additional details.

Table A.12: Innovation and Firm Growth: Controlling for Measures of Investor Attention

Horizon	1	2	3	4	5
	A. Control for number of WSJ articles				
Firm	0.017	0.030	0.036	0.043	0.045
	[2.60]	[4.31]	[2.89]	[3.23]	[3.12]
Competitor	-0.033	-0.053	-0.047	-0.057	-0.074
	[-13.84]	[-5.67]	[-5.53]	[-6.45]	[-6.78]
N	28966	26293	24077	19541	15529
	B. Control for number of analysts				
Firm	0.011	0.018	0.022	0.026	0.028
	[2.66]	[4.45]	[3.50]	[3.37]	[3.07]
Competitor	-0.020	-0.040	-0.042	-0.042	-0.046
	[-2.47]	[-5.36]	[-7.12]	[-5.38]	[-4.99]
N	77483	69322	62209	55734	49858
	control for institutional ownership				
Firm	0.019	0.029	0.036	0.040	0.042
	[3.18]	[4.51]	[3.70]	[3.23]	[3.00]
Competitor	-0.015	-0.032	-0.034	-0.037	-0.040
	[-2.26]	[-4.33]	[-5.68]	[-4.53]	[-4.61]
N	70874	61737	53941	47165	41338

Table repeats the analysis in Table 4 in the paper using additional controls for the degree of investor attention. In panel A we control for the (log one plus the) number of articles that mention the firm in the Wall Street Journal. The data is available over the 2000-2007 period. The data is from matching news articles in Factiva to firms following the procedure in Butler and Gurun, 2012. In Panel B, we control for the (log of one plus the) number of analysts covering the stock. The data is from I/B/E/S and covers the 1975-2010 period. In Panel C, we control for the fraction of institutional ownership. The data is from Thomson Reuters Institutional (13f) Holdings - Stock Ownership Summary and covers the 1980-2010 period. See the notes to Table 4 in the paper for additional details.

Table A.13: Innovation and Firm Growth: IV using tax price of R&D

Profits										
Firm					a. Competitors					
1	2	3	4	5	1	2	3	4	5
0.116	0.221	0.273	0.352	0.409	-0.116	-0.207	-0.266	-0.327	-0.373
[1.92]	[2.09]	[1.94]	[1.93]	[1.75]	[-2.54]	[-2.50]	[-2.37]	[-2.30]	[-2.08]
b. Output										
Firm					Competitors					
1	2	3	4	5	1	2	3	4	5
0.069	0.099	0.148	0.208	0.262	-0.070	-0.134	-0.187	-0.233	-0.273
[1.91]	[1.60]	[1.67]	[1.76]	[1.69]	[-2.29]	[-2.49]	[-2.44]	[-2.35]	[-2.16]
c. Capital										
Firm					Competitors					
1	2	3	4	5	1	2	3	4	5
0.076	0.128	0.176	0.233	0.296	-0.092	-0.171	-0.242	-0.319	-0.386
[2.12]	[2.02]	[2.00]	[2.04]	[1.97]	[-2.99]	[-3.06]	[-3.08]	[-3.12]	[-2.94]
d. Labor										
Firm					Competitors					
1	2	3	4	5	1	2	3	4	5
0.052	0.083	0.113	0.148	0.195	-0.073	-0.139	-0.181	-0.213	-0.233
[1.64]	[1.43]	[1.39]	[1.39]	[1.39]	[-2.55]	[-2.76]	[-2.55]	[-2.28]	[-1.95]
e. TFPR										
Firm					Competitors					
1	2	3	4	5	1	2	3	4	5
0.092	0.149	0.170	0.212	0.229	-0.063	-0.095	-0.117	-0.130	-0.142
[2.26]	[2.44]	[2.31]	[2.38]	[2.16]	[-1.81]	[-1.77]	[-1.78]	[-1.65]	[-1.55]

Table repeats the analysis of Table 4 in the paper using instrumental variables. We use the R&D tax credit variation as an instrument for our innovation measure, following Bloom, Schankerman, and Van Reenen (2013). This R&D price is constructed at an annual level for each firm using state-level R&D tax credits. See Bloom et al. (2013) for more details on the construction of this variable. We instrument for the firm's own innovation θ_{ft} using the firm-level tax price; we instrument for the innovation by competing firms $\theta_{I\setminus f,t}$ using the average R&D price of competing firms. The first-stage F statistics vary from 17.1 to 61 across specifications and horizons. We cluster standard errors by firm. See notes to Table 4 in the paper for additional details.

Table A.14: Industry Output and Innovation – Comparison with Citation-Weighed Patents

Horizon (years)	1	2	3	4	5	6	7	8
	A. Industry output (quantity)							
SM	0.005	0.009	0.013	0.018	0.026	0.042	0.055	0.065
	[2.73]	[2.49]	[2.83]	[3.02]	[3.29]	[3.49]	[3.32]	[3.23]
R-sq	0.054	0.085	0.119	0.150	0.183	0.217	0.241	0.261
CW	0.007	0.016	0.024	0.031	0.038	0.044	0.049	0.053
	[3.07]	[3.61]	[4.05]	[4.11]	[3.96]	[3.81]	[3.61]	[3.29]
R-sq	0.051	0.086	0.119	0.148	0.176	0.197	0.215	0.229
	B. Industry output (value added)							
SM	0.001	0.002	0.004	0.007	0.012	0.027	0.038	0.047
	[0.61]	[0.44]	[0.64]	[0.79]	[1.15]	[2.35]	[2.78]	[2.83]
R-sq	0.014	0.027	0.044	0.059	0.077	0.098	0.121	0.140
CW	0.004	0.009	0.013	0.017	0.021	0.025	0.028	0.027
	[1.57]	[1.75]	[1.77]	[1.76]	[1.78]	[1.75]	[1.68]	[1.45]
R-sq	0.015	0.029	0.045	0.060	0.077	0.092	0.109	0.122
N	1395	1364	1333	1302	1271	1240	1209	1178

Table reports the relation between innovation and output growth at the industry level. We construct industry-level innovation indices as

$$\theta^i_{I,t} = \frac{\sum_{f \in I} \Theta^i_{f,t}}{\sum_f B_{ft}},$$

using both the market based measure ($i = sm$) as well as for cohort-adjusted, citation-weighted patent counts ($i = cw$). We report the estimated coefficients a_τ from a specification similar to equation (12) in the paper,

$$x_{t+\tau} - x_t = a_0 + a_\tau \, \theta^i_{I,t} + \rho x_t + Z_{It} + u_{t+\tau}.$$

where x is log industry output (quantity in Panel A, value added in Panel B) and Z is a vector of controls that includes log capital, log employment, mean industry id. volatility and time effects. We compute standard errors using Newey-West. To compare across the two measures, we scale $\theta^i_{I,t}$ to unit standard deviation.

Table A.15: Innovation and Aggregate Growth – Comparison with Citation-Weighted Patents

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
				A. Aggregate Output				
χ^{cw}	0.005	0.009	0.013	0.015	0.017	0.022	0.023	0.022
	[1.08]	[1.10]	[1.36]	[1.40]	[1.70]	[2.31]	[2.24]	[2.13]
R-sq	0.078	0.113	0.19	0.254	0.289	0.347	0.373	0.423
				B. Aggregate TFP				
χ^{cw}	0.004	0.007	0.011	0.016	0.02	0.022	0.023	0.026
	[2.30]	[2.20]	[2.27]	[2.94]	[3.12]	[3.14]	[3.33]	[3.89]
R-sq	0.201	0.305	0.397	0.493	0.535	0.559	0.592	0.652

Table repeats the analysis of Figure 5 in the paper using an alternative index of innovation that is constructed using patent citations. Specifically, in a direct analogy to equation (18) in the paper, the value of the index in year t is given by

$$\chi_t^{cw} = \frac{\sum_{j \in J_t} \hat{C}_j}{Y_t},$$

where $\hat{C}$ is the number of citations to patent j in the first 10 years since its grant date, J_t is the set of patents issued in year t (including both private and public firms) and Y is aggregate output. Due to truncation, the sample ends in 2000. We report the estimated coefficients a_τ from the following specification

$$x_{t+\tau} - x_t = a_0 + a_\tau \log \hat{\chi}^{cw} + \sum_{l=0}^{L} c_l \, x_{t-l} + u_{t+\tau}.$$

Here, x is log aggregate output (panel A) or log TFP (panel B). We scale $\log \hat{\chi}^{cw}$ to unit standard deviation. We examine horizons of one to five years. We select the number of lags L using the BIC criterion, which advocates a lag length of one to two years depending on the specification. We compute standard errors using Newey-West.

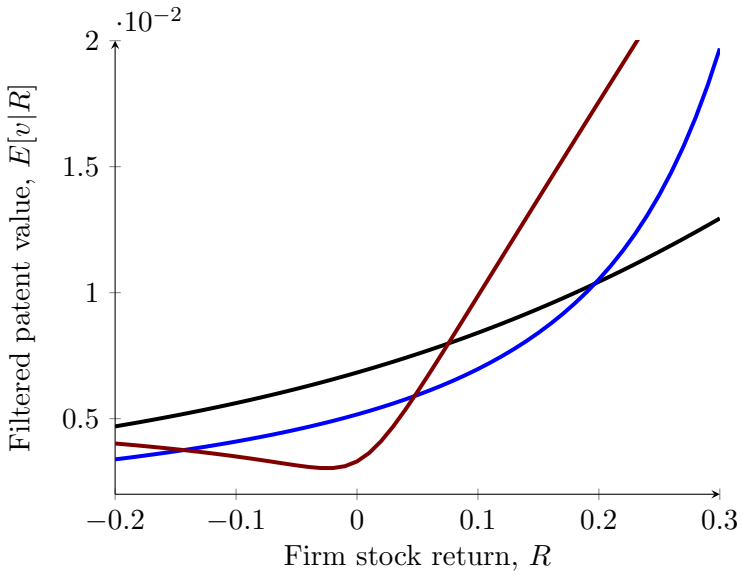


Figure A.2: Plot compares the filtered values, $E[v|R]$ across three different assumptions: our baseline case (black); the assumption that v is exponentially distributed (blue); the assumption that ε is Cauchy distributed and v follows a truncated Cauchy (at zero). See notes to Table A.7 for more details on the estimation of the parameters. We use the sample mean for the variance (or scale) of the error term to draw these graphs. We use the implied estimates from equation (6) in the main text to calibrate δ across the three cases.

Figure A.3: Innovation and Aggregate Growth – VAR results

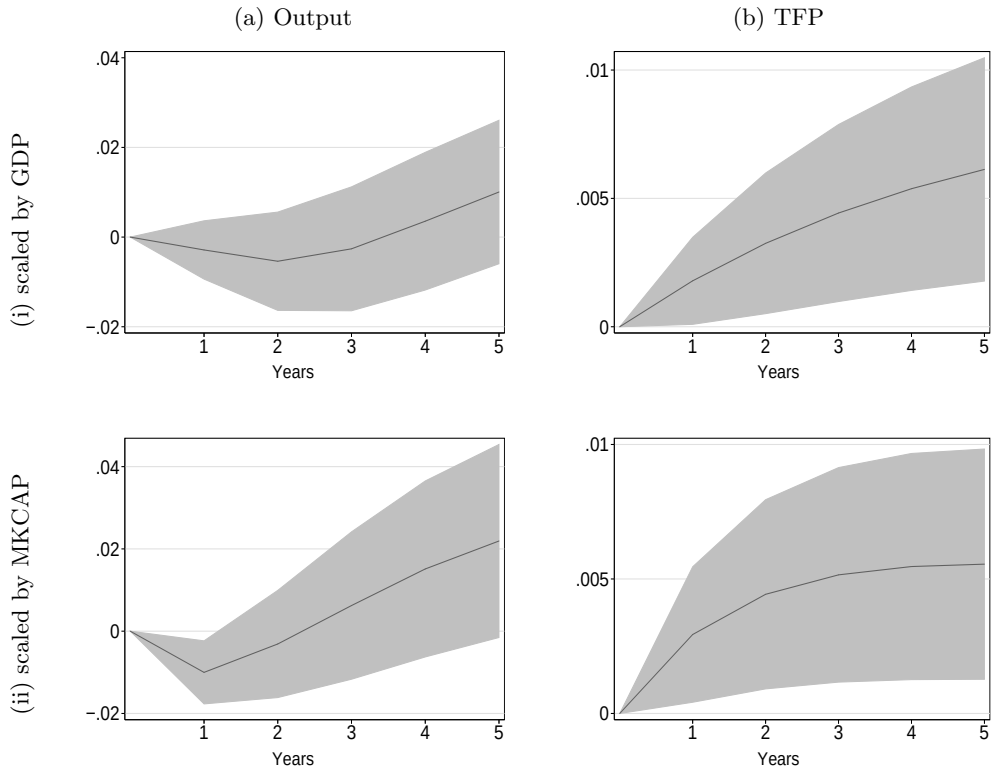


Figure shows impulse response of output per capita and productivity to innovation using bi-variate VARs. We obtain impulse responses by ordering our innovation measure last. We select lag length based on the BIC criterion. Dotted lines represent 90% confidence intervals using standard errors are computed using 500 bootstrap simulations. Productivity is utilization-adjusted TFP from Basu, Fernald, and Kimball (2006). Output is gross domestic product (NIPA Table 1.1.5) divided by the consumption price index (St Louis Fed, CPIAUCNS). Output per capita is computed using population from the U.S. Census Bureau.

References

Basu, S., J. G. Fernald, and M. S. Kimball (2006). Are technology improvements contractionary? *American Economic Review 96*(5), 1418–1448.

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

Hall, B., A. J. and M. Trajtenberg (2001). The NBER patent citation data file: Lessons, insights and methodological tools. Technical report, NBER Working Paper 8498.

Nicholas, T. (2008). Does innovation cause stock market runups? Evidence from the great crash. *American Economic Review 98*(4), 1370–96.

Norvig, P. (2009). Natural langugage corpus data. In T. Segaran and J. Hammerbacher (Eds.), *Beautiful Data*. O'Reilly Media.