Comprehensive Universe of U.S. Patents (CUSP): Data and Facts

Enrico Berkes

September 7, 2016

This Version: May 9, 2018

Abstract

Patents are commonly used as the main source of data for empirical studies related to innovation and technological change. The large amount of information about the underlying innovative process contained in each patent has certainly contributed to their popularity. Nevertheless, due to the lack of reliable data, historical analysis has focused on relatively small time frames or on specific dimensions of patents data. The goal of this paper is to fill this gap. I build and release a comprehensive time series of the universe of U.S. patents. The data set contains all the variables commonly used in the literature and, importantly, geolocates every inventor and assignee reported in each grant over the period 1836-2016.

1 Introduction

Patents have been the main source of data for empirical studies on innovation and technological change. Despite being an imperfect proxy for technological input and output,[1] the fact that patent data are easily accessible, offer a wide range of information about the invention content and the underlying innovation process, and are available for a large number of developed countries has contributed to their popularity in the literature. With some notable exceptions (e.g., Nicholas, 2010), until recently, research papers on the topic have mostly focused on the past 50 years. Similarly, historical analysis has concentrated on relatively small time frames (e.g., Moser, 2005) or on specific dimensions of patents data. The likely underlying reason is the lack of a reliable source of data for historical patents. In fact, the U.S. Patent and Trademark Office (USPTO) provides detailed data for all the patents issued from 1976 on, and studies on innovative activities prior this year often required the collection of data by hand.

[1] For example, Moser (2005) uses data from two World's Fairs at the end of the 19th century and shows that inventors from countries without patent laws focused on sectors that relied more on secrecy than patenting. This suggests that, at least in that period, patenting activity was skewed towards a certain set of industries.

More recently, thanks to the availability of increasingly reliable Optical Character Recognition (OCR) software, cheap computational power, and the publication of high-quality scan of historical patents by the USPTO various scholars started working on historical patent data. Notable examples are Akcigit et al. (2017), Sarada et al. (2017), Packalen and Bhattacharya (2015), and Petralia et al. (2016). The first two match patents data to the recently released decennial Census data and therefore mainly focus on the decades between 1880 and 1940. Packalen and Bhattacharya (2015) study the importance of physical proximity for innovation throughout history. To do so, they extract the name of the city, or county, from the text of each patent and study how the tendency of using new ideas in inventions changes with population density. Finally, Petralia et al. (2016) digitalize the images provided by the USPTO and extract information about the county of residence of inventors and assignees. The parsing of the text is supplemented with some machine learning techniques, such as neural networks, that are used to measure the plausibility of the collected data, as well as to infer the values of missing observations.

Despite the important contribution of these papers, different data sets contain different sets of variables and cover different time frames. Moreover, not all the data collected in these projects are readily available to other researchers. The aim of this paper is to fill this gap. Three years ago, I started working on a newly assembled data set of historical patent. The idea was to collect all the variables that are commonly used in the innovation literature using a consistent methodology and data sources and share the result with the rest of the community. Traditional sources, such as the USPTO and Google Patents, are complemented by newly digitalized patent documents and an extensive use of fuzzy matching is employed to extract information about the patent itself (e.g., technology classes, filing year, and backward citations), as well as about inventors and assignees. Each inventor and assignee is geolocated at the town level, the most disaggregated geographical level that is possible to identify from the patent text. The outcome is what I called the Comprehensive Universe of U.S. Patents (CUSP). It spans almost two centuries of patent data (1936-2015) and contains the richest set of variables available so far. Various sanity checks show a high degree of accuracy.

The first part of the paper describes in details the data sources and the techniques used to extract the data. I also compare CUSP with HistPat (Petralia et al., 2016), one of the most promising data sets of historical patents readily available on the Harvard Dataverse. The analysis shows a broader coverage of CUSP and a similar level of accuracy in terms of geolocation of the patents, the dimension that is most stressed in HistPat. In the second part, I report some stylized facts. Some of these are new and might point to interesting directions for future research. Some others confirm well-known patterns already

discussed in the literature (e.g., the upward trend in the average number of inventors per patent is already described by Wuchty et al., 2007). Nevertheless, this new data set offers for the first time a long-term perspective and allows us not only to observe trends but also to pin down when the trends started in history.

The rest of this paper is structured as follows. Section 2 describes in details the data contained in the CUSP and how they were assembled. Section 3 briefly compares the CUSP with some other historical patent data sets. Section 4 provides some stylized facts that might be source of inspiration for future research. Section 5 concludes.

2 Data

The data set collects a comprehensive set of variables for the entire universe of patents issued by the USPTO between 1836 and 2015. To do this, I use five distinct data sources:

1. Patent text and information reported on the USPTO website;[2]
2. State-, or in one case city-, level databases. Such databases are usually maintained either by universities or public libraries and contain all the inventions a (not always) comprehensive list of the patents whose inventor was resident in that state (or city). In many cases, these databases only cover historical patents. I was able to identify seven local inventors databases:
   (a) Cincinnati Inventors Database
   (b) Iowa Inventors Database
   (c) Nevada Inventors Database
   (d) Oklahoma Inventors Database
   (e) South Carolina Inventors
   (f) The Portal to Texas History
   (g) Wyoming Inventors Database;
3. High quality patent images digitalized with an OCR software;
4. Google Patents;

[2] http://patft.uspto.gov/netahtml/PTO/search-adv.htm

5. Patents issued after 1920 digitalized by Google and made available on the USPTO website. This was the first attempt made by Google to OCR historical patents and the result is generally of poor quality. Nevertheless, these data are used as a last resort in case it is not possible to extract the needed information from the previous sources.

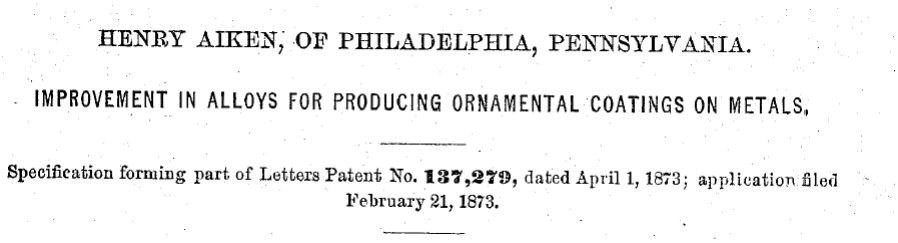
Using multiple sources reduces the probability that the information I am looking for is not available in any of them, and allows me to select the most reliable one. Given the peculiarities of each database in terms of the degree of accuracy and data availability, the database of choice is based on the year in which the patent was issued and the piece of information I am trying to collect. First, since the USPTO makes readily available all the information for the patents issued after 1976, their website is my preferred source for all the patents issued after that year. Additionally, from there it is also possible to collect information about the technology classes for all the grants going back to 1836. Second, I use the local inventor databases to extract information about inventors and assignee and their town of residence for all the patents listed there. These data have a limited coverage in terms of space and time, but the information contained in the local databases is reliable and easy to extract. Finally, I parse the patent text obtained from the three digitalization processes (mine, Google Patents, and USPTO) for all the remaining variables and patents. It was not possible to extract all the pieces of information for the universe of patents, but all the patents are listed in each table.

The rest of this section describes in more details the variables available in CUSP and the strategy employed to extract them.

2.1 Issue and Filing Years

A patent's issue year is readily available from the USPTO website for all the patents ever granted. The same is not true for the year in which the patent was filed. This piece of information is often missing for historical patents. However, filing years are arguably a better indicator of when the invention was completed than issuing years.[3] When not available from digital sources, it is possible to retrieve the date in which the patent was filed directly from the patent text starting from patent number 137,279 and issued on April 1, 1873. The filing date appears in the patent header preceded by "application filed on". Figure 1 shows the header of this patent. The parsing process follows two increasingly less stringent steps. First, I look for sequences of exactly four numbers preceded by the words "application", "filed", "tiled", "fiied",

[3]Figure 9 in Section 4 shows that the up until the 19th century the issue and filing years were very close, with an issuance time of less than one year for the average patent. However, the two kept diverging until the late 40s when on average a patent had to wait 4 years before being issued.



HENRY AIKEN, OF PHILADELPHIA, PENNSYLVANIA.

IMPROVEMENT IN ALLOYS FOR PRODUCING ORNAMENTAL COATINGS ON METALS.

Specification forming part of Letters Patent No. **137,279**, dated April 1, 1873; application filed February 21, 1873.

Figure 1: Header of patent number 137,279, the first patent that reports the year in which the application was filed.

or "fledi"[4] and followed by a month or its abbreviation (e.g., january or jan).[5] Second, if this procedure is not successful, I look for sequences of exactly four numbers that are on the same line as the keywords listed above.

Since the likelihood of error is different for each of the two steps, each observation in the data set is assigned a flag that will help researchers to understand how confident we should be with the value reported. The flag is set equal to 1 if the filing year comes from the USPTO (or Google Patents) website; 2 if the filing year was obtained through the first round of parsing; 3 if it was obtained by searching for sequences of four numbers appearing on the same line as the keywords "application" and "filed" (and its variations). At the end of this process, the filing year of 8,178,429 patents (or 93.2%) was obtained from an official source, 446,184 (or 5.1%) from the first round of parsing and 88,270 (or 1.01%) from the second round.[6]

Finally, issue and filing years are checked for consistency. If the first two digits are a 9 and a 1, respectively, I swap them;[7] if the issue year is outside the time frame of the dataset (1790-2015), then I replace it with a missing value; if the filing year is outside the time frame of the dataset, is larger than the issue year, or the difference between issue and filing years is bigger than 30, then I set it to missing value. In the end, issue and filing years are available for a total of 8,712,883 patents (or 99.3%).

2.2 Technological Classes

Technological classes are assigned to each patent by patent reviewers. The USPTO regularly updates class definitions and corrects the classification of patents backwards. Each patent is associated with

[4] "tiled", "fiied" and "fledi" are common mistakes made by the OCR software when reading "filed".

[5] Note that before this process, I substitute all the occurrences of "l9" ("el" followed by a nine) and "|9" with "19". Similarly, for "l8" and "|8".

[6] Note that the percentages here is calculated using 8,772,775 as denominator, that is the total number of patents minus the number of patents for which the filing year is unknown since it is not reported anywhere in the text (i.e., patents whose number is smaller than 137,279).

[7] Swapping the first two digits is a relatively common typo in patent documents.

multiple technology classes according to three different classification systems: the U.S. Patent Classification (USPC), the Cooperative Patent Classification (CPC) and International Patent Classification (IPC) schemes. Technology classes are indicative of the pieces of knowledge embedded in the invention patented. For example, a patent that describes an image processing method for TVs might be classified under the US patent categories 382 (Image Analysis) and 348 (Television). Documents classified in the USPC system are also assigned one (and only one) principal class. The principal class captures the scope of the invention as a whole or the main inventive concept using the claims as a guide (USPTO, 2012). The CPC and the IPC do not specify any main technological category. If a patent cannot be classified within the current classification system, or its principal class is unclear at that point in time, then it is assigned a 1/1 as main class. Since the USPTO reviews and updates its classification system every couple of months, the total number of 1/1 patents is quite small (15,819 patents or 0.2% of my sample).

In my data set, I collect technology classes for all the three classification schemes directly from the USPTO website.[8] Classes were collected in June 2016 and therefore represent the classes assigned at that point in time. The data set contains two tables that report the technology classes according to the USPC. The first contains the classes as they were made available by the USPTO. In the second, I assign a main class to the patents whose principal class is 1/1 based on the frequency of its secondary classes (at a three-digit level). When two or more secondary classes appear with the same frequency, the one with the smallest number is selected. For example, if a patent was assigned classes 1/1, 324/12, 324/121, 345/67, 345/87, then 1/1 is substituted with 324.[9]

2.3 Backward and Forward Citations

Backward and forward citations have been extensively used in the empirical literature to understand knowledge flows across firms and inventors, as well as as a measure of patent quality, the idea being that the more a patent is cited the more the invention it describes is valuable.[10] That is why a patent data set would not be complete without this piece of information. For each patent in the dataset, I collect the patent number of all the U.S. patents referenced in the grant. Once the backward citation matrix is populated, it is possible to obtain the list of forward citations, that is the patent numbers of all the patents that cite a certain invention, simply by "inverting" that matrix. Starting in 1947, all the patents issued by

[8] See for example http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&u=%2Fnetahtml%2FPTO%2Fsearch-adv.htm&r=1&p=1&f=G&l=50&d=PALL&S1=0137279.PN.&OS=PN/137279&RS=PN/137279 for patent number 137,279. The principal class according to the USPC is in bold.

[9] Note that even after this procedure some patents are assigned class 1/1. This is due to the fact that they do not report any additional class making the frequentistic procedure impossible.

[10] Alcácer et al. (2009) discuss the merits and demerits of using citations in empirical work.

the USPTO include a section that lists all the references cited.[11] Before that year, prior art upon which the invention was built was reported on the file history which is not publicly available. Nevertheless, some patents were directly referenced in the patent text and it is therefore possible to get a sense, albeit noisy, of knowledge flows across technology fields and regions.

The data collection strategy for backward citations crucially depends on when the patent was issued. For patents issued after 1976, citations are directly collected from the computerized patent information available on the USPTO's website. For patents issued between 1947 and 1975 I parse the text and extract the contents of the section titled "References Cited" that lists the references to other patents or, in some cases, scientific articles. From this section, I collect the number of all the U.S. patents cited. Finally, for the patents issued before 1947, I look for references to other patents directly in the patent text. In particular, I look for sentences that contain the keywords "patent" or "patents" followed by "no", "number", "numb", "num", "nos", or "numbers" and get the patent number referenced afterwards. Figure 2 shows an extract of patent no. 46,101 in which the inventor describes how his patent differs from a previously issued patent and states his claims. This strategy finds a total of 182,044 patents cited by patents issued before 1947. I apply the same strategy for all the post-1947 patents that do not include a "References Cited" section.

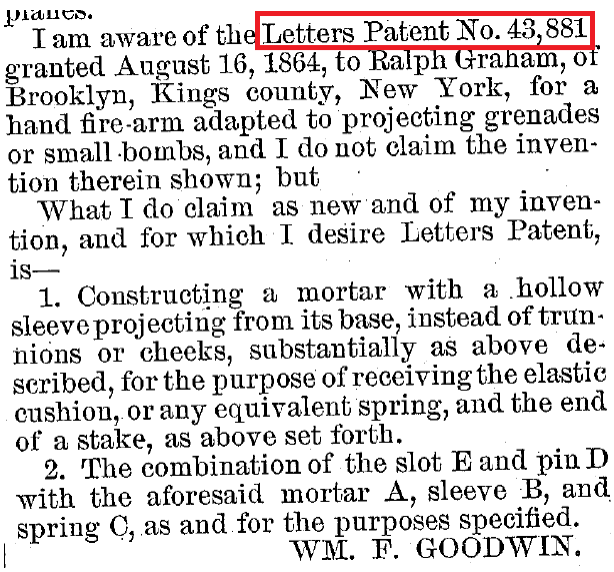
The two tables that contain backward and forward citations are structured with a long form. The first column contains the number of the citing patent, whereas the second column the number of the patent cited. Each line correspond to a single citation. A patent that cites multiple grants will appear on multiple lines. The table containing backward citations has two additional columns. The first is a binary variable that takes value 1 if the citations was added by the examiner, as reported by Google Patents. The second column contains a flag that take value 1 if the citations are collected from a digital source (i.e., either Google Patents or the USPTO website); value 10 and 11 if the citations are obtained from the "References Cited" section of patents OCR'ed by the USPTO and by myself, respectively; value 5 if the citations come from the main text of the patent as in Figure 2.[12]

2.4 Inventors Name and Location

The collection of the inventors names and locations is the most challenging and sensitive task. For this reason, particular attention was devoted to this phase of the data collection. Fuzzy matching techniques are employed to overcome some of the problems that occur due to the fact that the performance of OCR

[11] The first patent to include a "References Cited" section is Patent Nr. 2,415,068.

[12] This step only uses the patents OCR'ed by myself since the quality of the text is generally superior.



planes.

I am aware of the Letters Patent No. 43,881, granted August 16, 1864, to Ralph Graham, of Brooklyn, Kings county, New York, for a hand fire-arm adapted to projecting grenades or small bombs, and I do not claim the invention therein shown; but

What I do claim as new and of my invention, and for which I desire Letters Patent, is—

1. Constructing a mortar with a hollow sleeve projecting from its base, instead of trunnions or cheeks, substantially as above described, for the purpose of receiving the elastic cushion, or any equivalent spring, and the end of a stake, as above set forth.

2. The combination of the slot E and pin D with the aforesaid mortar A, sleeve B, and spring C, as and for the purposes specified.

WM. F. GOODWIN.

Figure 2: The figure shows an extract from patent number 46,101. The patent references another patent in the text. This piece of information is used to build a data set of citations prior to 1947.

programs heavily relies on the original image quality and sometimes the digitalized text displays various typos. As for backward citations, use the information available on the USPTO website to extract the name of the inventors and their residence for all the patents issued after 1976.[13] The maximum number of inventors in a single patent in the sample is 76. This is the number of inventors of grant number 7,581,231, a software patent filed by Microsoft.

For patents whose patent number is smaller than 1,583,767, I use a three step approach to collect the relevant information. First, I parse the end of the patent and identify the inventors' signatures (in print). Figure 3 shows the very end of patent number 580, a bee hive. The name of the inventor is reported in capital letters together with the name of two witnesses. The fact that signatures are printed in capital letters and with a larger font minimizes the amount of typos during the digitalization process. Second, I parse the patent header (Figure 1) looking for the residence of the inventors identified at the end of the grant. The patent header is characterized by the keywords "United States Patent Office" or "assignor" (and some variations that take into account frequent typos), whereas the inventors' location is extracted by looking for keywords like "of" or checking weather the name of a state is contained in the header string. Third, if the code is unable to extract the information from the header (for example because none of the keywords listed above is present), then I parse the beginning of the patent text (Figure 4). All the patents prior to 1,583,767 start with a formula similar to the one in the Figure: "*To all whom it may*

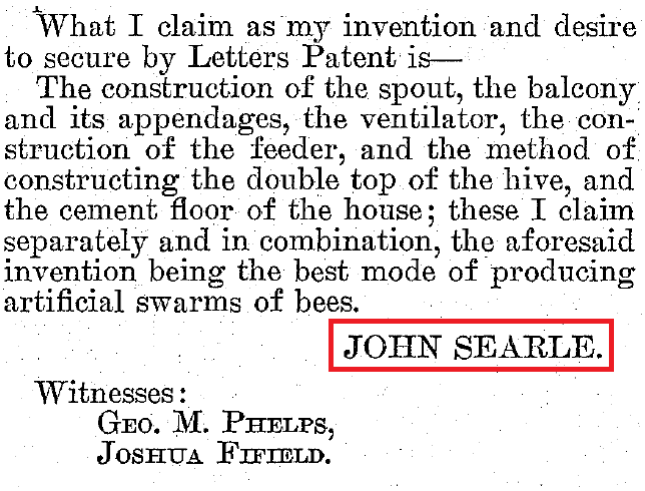
[13]Note that the USPTO reports the details of some patents even before 1976. In this case, all the details are collected from there.

concern: Be it known that I, <inventor name>, residing at <name of city>, in the county of <name of county> and State of <name of state>". By searching for this pattern, it is possible to extract the city and state of residence of a certain inventor. This technique is however left as last resort, since parsing the patents text is prone to more typos than the header (which is written in a bigger capitalized font), and the formula changes from time to time, making the pattern matching task more difficult.

The strategy used to extract inventors names and locations for the patents whose number is between 1,583,767 (included) and 1,920,165 (excluded) is similar to the one above, with the exception that the third step had to be dropped, since the patent text does not contain information about the inventors anymore. This is the same strategy employed to extract information for all the patents issued after patent 1,920,165. However, for these the parser needed to be modified to take into account the new structure of the header (e.g., the keywords used to identify the header are different). Note that for the majority of patents issued starting from the end of the 19th century, the name of the inventors is readily available from Google Patents. In that case, the name of the inventors is taken from there and the steps described above are used for the sole purpose of getting information about their residence.

Possible typos in the location names are then fixed by using a frequentistic approach. First, I count how often a city/state pair appears in the data set. Second, I iterate over all the inventors in the data set and compare the reported location with those in the previously built dictionary. If the dictionary contains a city in the same state with a Levenshtein distance of 2 or less that appears more frequently in the original data set, then I assign that city to that inventor. Similarly, if the data set contains a city/state pair with a Levenshtein distance of 1 or less for both the city and state (e.g., Chicago, IL and Chicag, HI) and a higher frequency then I assign to that inventor the more recurring pair. Finally, when no location is reported for an inventor, I check in the previous and if there is an inventor whose location is not missing with the same name and who filed a patent one year before or after. If that is the case I assign the location of the latter to the former.

I assign a latitude and longitude to each inventor using a two step approach. First, I use an offline database of geolocated U.S. cities. If I find a perfect match in the database, then the latitude and longitude reported there is assigned to the inventor. Second, if the city and state do not match any entry, I send a query to Google Maps. The advantage of Google Maps is that it handles typos and errors very well. For example, looking for "Ls Angeles, WI" correctly returns the latitude and longitude of Los Angeles, CA. Once each inventor is geolocated, I assign to the them a county and correct back states and countries of residence through ArcGIS. In other words, I assign to them the country, state, and county



What I claim as my invention and desire to secure by Letters Patent is—

The construction of the spout, the balcony and its appendages, the ventilator, the construction of the feeder, and the method of constructing the double top of the hive, and the cement floor of the house; these I claim separately and in combination, the aforesaid invention being the best mode of producing artificial swarms of bees.

JOHN SEARLE.

Witnesses:
GEO. M. PHELPS,
JOSHUA FIFIELD.

Figure 3: The figure shows an extract from the end patent number 580. There the name of the inventor is listed in capital letters together with the name of two witnesses.

their geo-coordinates fall into based on a map of year 2000.[14]

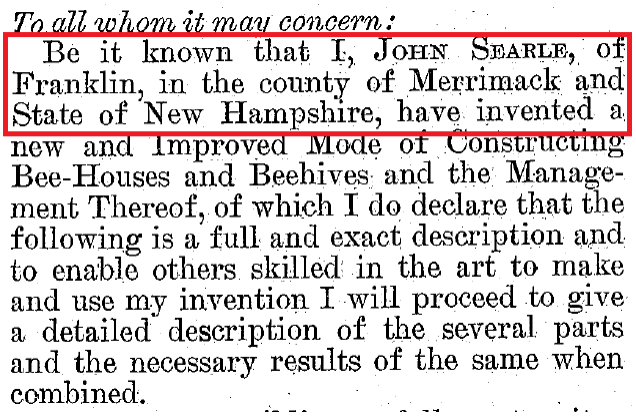
2.5 Assignees Name and Location

Extracting the location and name of assignees from the patent documents is a more straightforward task compared to extracting information about inventors, but also one that is prone to more mistakes. In fact, there is no redundancy in the documents: details about the assignees appear one and only one place: the header of the patent. Using the same procedure developed for inventors, I identify the header of each patent and check for the presence or absence of the string "assign". If this sequence of characters is not contained in the header, then I conclude that the patent has no assignees, otherwise I parse the rest of the line searching for the name and location of the assignees. Unfortunately, their location is not always available: sometimes only the assignee name is reported, while other times the assignee name is followed by "a corporation of <name of state>" without any further detail. From a careful review of a number of patents, it seems that the state reported there represents where the firm is registered, and does not necessarily indicate the location of the branch where the inventor works.[15] For this reason, when either the location is missing or the assignee name is followed by "a corporation of <name of state>" without any reference to the city where the assignee is actually located, I assign the company to the same location of the first inventors, when they are all reported to live in the same location.[16] This approach biases the

[14]Note that counties might therefore differ from other datasets of historical patents because they usually collect the county where the town was located when the patent was filed.

[15]For example patent number 1,898,054 is assigned to the National Lead Company of New York, N.Y., a corporation of New Jersey.

[16]In the next iteration of the data set, I will flag these instances to allow researchers to drop them for robustness checks.



To all whom it may concern:

Be it known that I, JOHN SEARLE, of Franklin, in the county of Merrimack and State of New Hampshire, have invented a new and Improved Mode of Constructing Bee-Houses and Beehives and the Management Thereof, of which I do declare that the following is a full and exact description and to enable others skilled in the art to make and use my invention I will proceed to give a detailed description of the several parts and the necessary results of the same when combined.

Figure 4: The figure shows an extract from the end patent number 580. There the name of the inventor is listed in capital letters together with the name of two witnesses.

distance between inventors and assignees towards zero. Some of the facts reported in Section 4 should therefore be interpreted as a lower bound. Similarly to what I did for the inventors, I fix possible typos using a frequentistic approach and missing values looking at assignees with the same name one year before and after the filing year of the patent. Geographical coordinates are also assigned in the same way.

3 Validation and Comparison

An in-depth comparison with other data sets of historical patents is beyond the scope of this paper. The interested reader is referred to Andrews (2017) who presents a newly assembled data set of geo-referenced historical patents and, in doing so, he compares it with other four existing data sets.[17] However, to give credibility and motivate the data collection effort, it might be useful to contrast CUPS with HistPat, a very renowned publicly available data set of historical patents described in Petralia et al. (2016) and available on the Harvard's Dataverse. Table 1 schematically shows the variables available in the two data sets.[18] Figure 5 compares the number of patents contained in the two data sets and the actual number of patents reported by the USPTO by issue year. The dashed yellow line shows the official number of patents issued by the USPTO in each year. The total number of patents in CUSP almost perfectly match this series.[19] Since HistPat seems to only include patents for which all the inventors and assignees are located in the U.S., the red line shows the number of patent that satisfy this requirement in CUSP, whereas the

[17] Unfortunately, it was not possible to gain access to the data (or their aggregate statistics) underlying the work of Akcigit et al. (2017) and Packalen and Bhattacharya (2015). He therefore excludes them from the analysis in the present version of the paper.

[18] The rows and first column of the table are taken from Andrews (2017) and reported here upon the generous agreement of the author.

[19] Some minor differences are due to the fact that some patents were withdrawn after being issued. Those patents are discarded from my data set.

	HistPat	CUSP
Years Covered	1836-1976	1836-2015
Inventor First Name	N	Y
Inventor Last Name	N	Y
Inventor Town	N	Y
Inventor County	Y	Y
Inventor State	Y	Y
Full Patent Text	N	Available upon request
Patent Number	Y	Y
Application Date	N	Y
Grant Date	Y	Y
Names of Multiple Inventors	N	Y
Names of Assignees	N	Y
Assignee Town	N	Y
Assignee County & State	Y	Y
Patent Class	N	Y
Backward and Forward Citations	N	Y

Table 1: The table shows a schematic comparison of the variables available in HistPat and CUSP.

green line represents the number of patents in HistPat. The difference between the two series is always relatively small except for the period between the two World Wars when HistPat systematically covers less patents. Although CUSP contains a larger amount of variables and patents, Petralia et al. (2016) put a lot of intellectual and computational effort in identifying the county of residence of each inventor and assignee listed in the patent. The real test is therefore to compare the two data sets along a geographical dimension.

When multiple inventors and assignees are reported on a patent, HistPat assigns to that grant multiple locations without giving any information about whose residence is the one reported. For comparison purposes, I therefore extract all the counties assigned to the assignees and inventors of a patent and compare them with the ones listed in HistPat. If CUSP reports all the counties reported by HistPat for a given patent,[20] I count that as a success, otherwise the patent is categorized as a non-match. The resulting matching rate is about 80%. The exercise includes all the patents issued in the period 1836-1976 and available in both data sets. Figure 6 reports the share of non-matched patents by issue year. The share remains quite stable around 20% over the whole period with a peak of about 35% in 1919-1920. Analyzing by hand a random sample of the patents not matched shows mixed results. Sometimes CUSP contains the right location of the assignee but the wrong location of the inventor (or viceversa), whereas HistPat contains the wrong location of the assignee but the right county of the inventor (or viceversa);

[20]Note that in some cases CUSP actually contains more entries than HistPat.

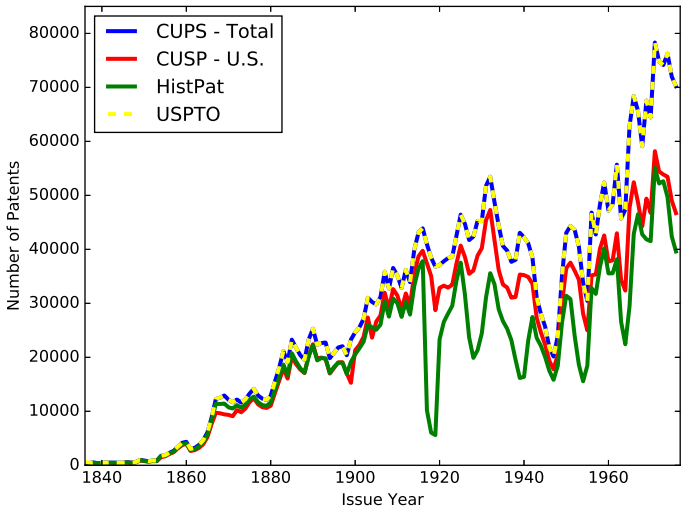


Figure 5: The figure compares the number of patents reported in HistPat and CUSP. For CUSP I only selected the patents for which at least one inventor or one assignee are U.S. residents.

sometimes CUSP is off track and other times HistPat is off track.[21] From time to time the mismatch is due to the fact that while HistPat reports the county stated in the text, CUSP reports the county that contained the town in 2000. For example, when patent number 48 was granted Portsmouth, VA was part of Norfolk County, whereas nowadays is an independent city. I include in the data set a list of the patent numbers that, according to the procedure described above, do not match in the two data sets. The list will provide guidance on where to concentrate my efforts for the next iteration of the data set.

4 Stylized Facts

4.1 Numbers

Fact 1.1: Patenting activity in the U.S. has steadily increased over time; the growth started accelerating in the 80s.

The number of patents filed at the USPTO has experienced an important acceleration starting in the 80s. This trend seems to be mainly driven by two factors. First, the number of U.S. patents that had been decreasing since the 60s shows a dramatic reversal of the trend in that decade. The change might be due to the growing importance of software patents. Second, the number of foreign patents also accelerated

[21] Note that sometimes patents contain contradictory information. For example patent number 9 states in the header that the inventor is from New York, which is what is reported in CUSP, whereas in the text the inventor is from Springfield, MA, which is what is reported in HistPat.

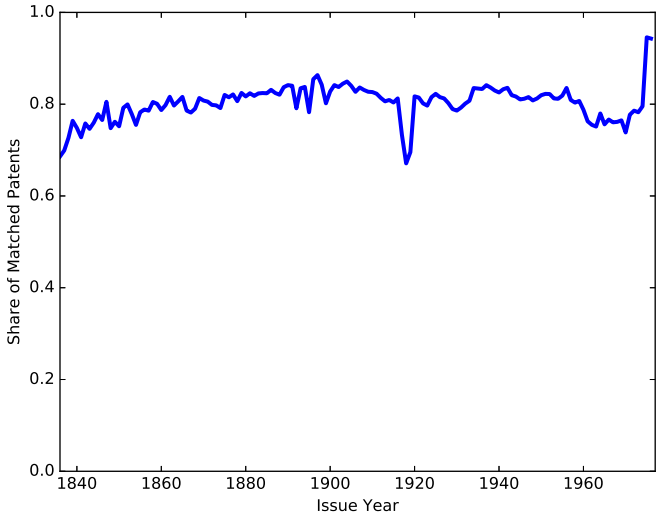


Figure 6: The figure shows the share of patents for which all the locations that appear in HistPat are also contained in CUSP. The denominator is given by the number of patents available in HistPat in a given year.

in those years, although the upward trend started already in the 50s. Figure 7 plots the total number of patents issued by the USPTO according to their filing year and country of residence of their inventors. The blue line represents the total number of patents by filing year, whereas the red and green lines show the patents whose inventors are foreign or a U.S. residents, respectively.[22] The graph highlights two additional interesting facts. First, the share of foreign patents before the 60s was almost negligible. Second, in 2010 the number of patents whose inventors are foreign residents passed the number of patents filed by inventors whose residence is in the United States.

Fact 1.2: The share of patents resulting from international collaborations started increasing in the 50s.

An international collaboration is defined as a patent for which at least one inventor is a U.S. resident and at least one other has her residence outside the United States. The number of international collaborations has importantly increased over the years with a steady growth that started in the 80s. Despite this, international collaborations still remain a small fraction of the total number patents filed at the USPTO. Figure 8 shows this pattern graphically. In 2010, less than 5% of the patents filed were the result of international collaborations.

[22]Note that the graph excludes what in Fact 1.2 I define international collaborations, so the green and red lines do not necessarily sum up to the blue line. However, as it is shown in Fact 1.2 the share of foreign collaborations is small. The main patterns of the graph do not change if I used the country of residence of the first inventor to classify patents into U.S. and foreign patents, instead.

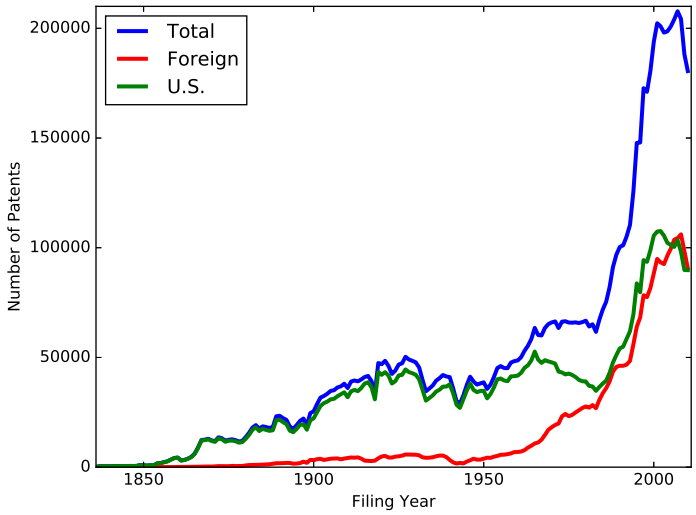


Figure 7: The graph plots the number of patents granted by the USPTO by filing year and country of residence of their inventors. The blue line represents the total number of patents issued by the USPTO. The green line shows the number of patents whose inventors are U.S. residents. The red line shows the number of patents whose inventors are foreign residents.

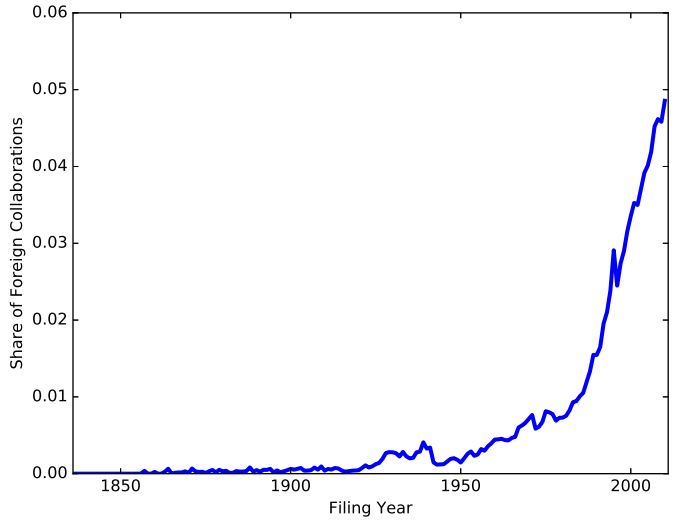


Figure 8: The figure shows the share of patents that were the result of an international collaboration by filing year. A grant is considered an international collaboration if at least one inventor is a U.S. resident and at least another one has her residence outside the United States.

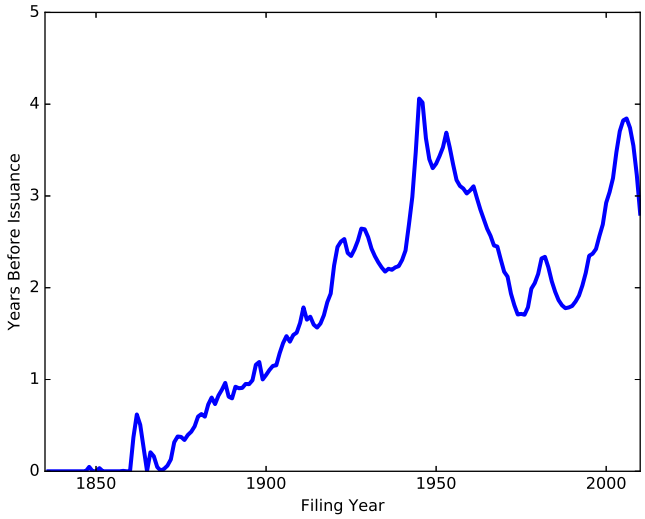
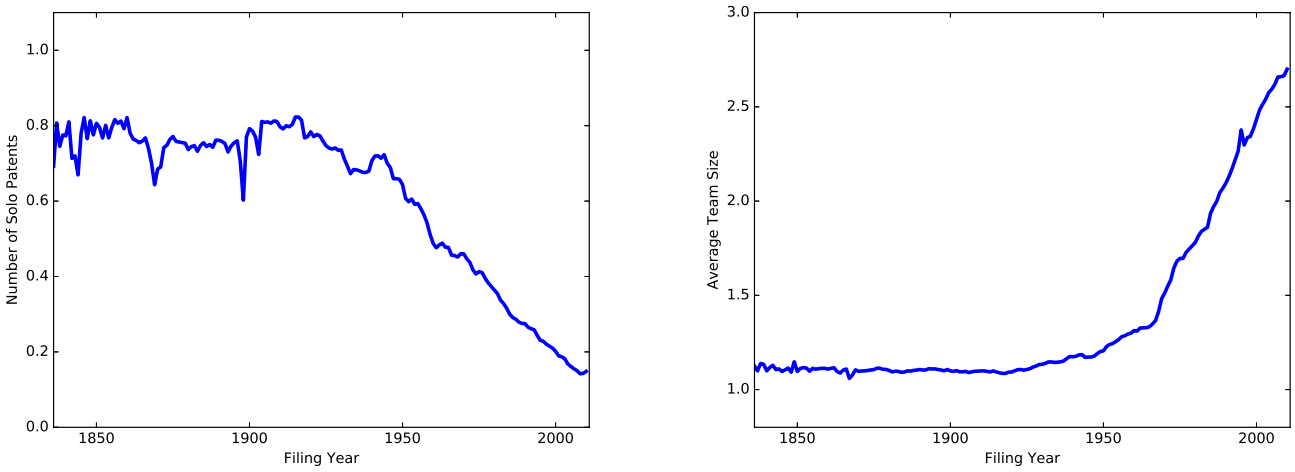


Figure 9: The figure shows the average time (in years) that a patent application filed in a certain year had to wait before being granted.

Fact 1.3: The time needed to issue a patent was negligible in the 19th century; it was on average 2-3 years in the 20th century.

Since information about the year in which the grant was filed is often absent in data sets of historical patents data, it is common practice in the literature to proxy the filing year with the year in which the patent was granted. Authors often argue that in the past the time necessary to examine a patent was shorter due to the smaller amount of applications and their relative simplicity (see for example, Akcigit et al., 2017). Figure 9 tests this hypothesis. The average issuance time for patents filed before 1900 was indeed below one year, but it was already almost 2 years by 1915 and more than 2.5 years in the 1920s. The average issuance time experienced an important increase during WWII reaching 4 years in 1947 and gradually went back to about 2 years in the period between the 1970s and the 2000s, when it started rising again to reach another peak in 2005 when the average patent had to wait about 4 years before being issued. The decrease at the end of the sample might be due to the decrease in the number of applications received by the USPTO during and after the Great Recession, or might be simply due to truncation problems (patents filed in 2010 with an issuance time larger than 5 years do not appear in the sample).



(a) (b)

Figure 10: The figure shows the decline of single-authored patents over time. Panel a reports the share of patents filed by a single inventor by filing year. Panel b the average number of inventor for each filing year.

4.2 Inventors

Fact 2.1: The share of single-authored patents was around 80% up until 1920 when it started declining.

Single-authored patents have become increasingly rare in the past century. Figure 10 provides evidence for this fact from two different angles. Panel a shows the share of patents filed by a single inventor, whereas panel b the average team size by filing year. The share of patents filed by a single inventor has steadily decreased over time since the end of the 1920s. In the 19th century between 70% and 80% of the inventions patented were single-authored. By 2010, this share had decreased below 20%. Similarly, average team size remains surprisingly stable, around 1.2 inventors per patent, up until the late 40s when it starts a rapid increase. In 2010, the number of inventors for the average patents is about 2.7, more than double compared to 60 years before. Wuchty et al. (2007) document this pattern for the period (1975-2000). Thanks to the larger time frame covered by CUSP, it is possible to put this finding into a historical context and pin down the moment when the shift happens. Interestingly, de Solla Price (1963) documents that the cost of research as a share of GDP did not increase before WWII when it started and exponential growth. It would be interesting to understand what factors have driven the decline of single-authored patents which started in the 1920s and accelerated in the late 1940s, and if technology fields contributed differential to this trend. This is left for future research.

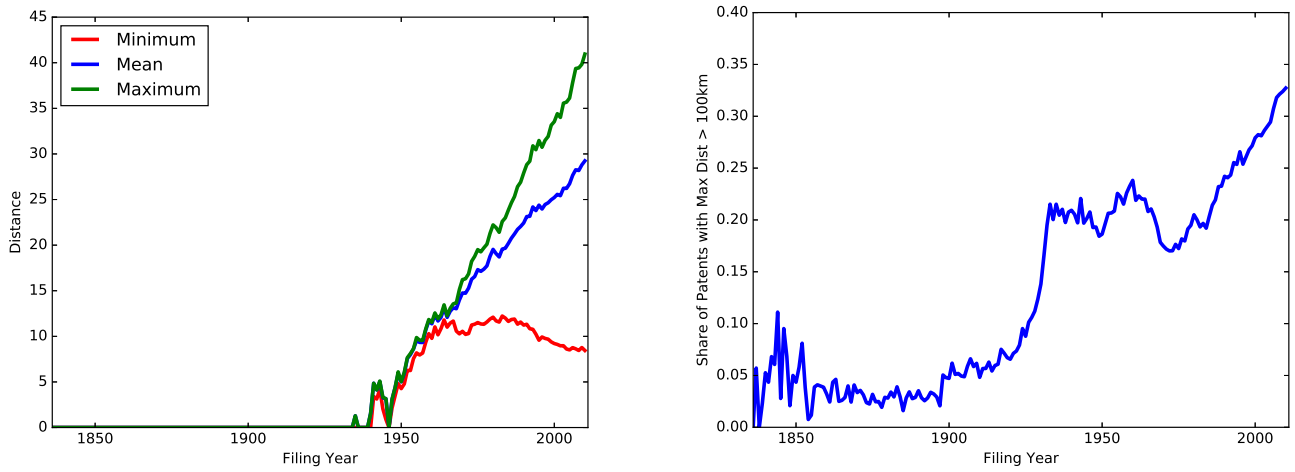
Fact 2.2: Average and maximum distance among the inventors of the median patent started an upward trend in the 50s; minimum distance increased at first and then plateaued.

An important idea in the innovation literature is that the decline in communication costs have made the collaboration with people living in other cities or countries less costly and hence proximity less important. Consistently with this intuition, Packalen and Bhattacharya (2015) find that inventors in more dense cities were adopting ideas faster throughout the 20th century, but the advantage of living in a large city has disappeared more recently. However, this insight appears to be in contrast with other observations, such as the existence of large innovation hubs, or of seminars and conferences that allow scholars to personally discuss with their peers.

A possible explanation to these two seemingly contradictory facts is that proximity still matters at the very beginning of a project and for certain specific tasks. For example, informal exchanges of ideas might play a crucial role in first stages of a project and proximity might be important to, say, analyze and brainstorm about the outcomes of lab experiments. If this was the case, I would expect to observe an increase in the geographical dispersion of teams of inventors over time. The inventors who need to work on tasks that require proximity should be clustered in space, but could potentially be geographically disconnected from the other members of the team. Figure 11 tries to shed some light on this by plotting the minimum, mean, and maximum distance among the inventors of the median patent. More precisely, I calculate the minimum, average, and maximum distance among the inventors of each U.S. patent filed in a given year by two or more inventors.[23] The left panel of Figure 11 reports the median of these distribution. The graph shows a clear increase in the three series between 1950 and 1970, when they started diverging. After 1970, the minimum distance of the median patent stabilized around 10 kilometers, whereas mean and maximum distances kept their growth and reached 30 and 40 kilometers, respectively, in 2010. The right panel shows the share of patents for which at least one inventor is reported to live at least 100 kilometers away from any other inventor in the patent. This series shows two breaks. One between 1930 and 1940 that brought the share of these patents from about 7% to about 22%, and one in the 1970s when a still ongoing upward trend started. In 2010, about 33% of the filed patents had at least one inventor more than 100 kilometers apart. Figure 12 shows the share of patents for which at least two inventors live in the same city.[24] This share has also experienced an important decline between 1930 and

[23]Note that these three statistics coincide when there are only two inventors and that solo patents are discarded for this analysis.

[24]Note that this is not necessarily the specular image of Figure 11, panel b, since a patent with three inventors, two living in Chicago and one in Columbus, would contribute to both graphs.



(a) (b)

Figure 11: This figure analyzes the distance patterns among the inventors of a same patent. The left panel shows the minimum, mean, and maximum distance across the inventors of the median patent. The right panel reports the share of patents with at least two inventors that live more than 100 km apart.

1950, but it then stabilized just below 40%.In future research, it could be interesting to study whether some technology fields have contributed differentially to this trends and also if they are confirmed when keeping the team size constant. Since team size has also been increasing over the same period of time, a null model in which inventors are added in a pseudo-random way could be consistent with the pattern described in Figures 11 and 12.

4.3 Assignees

Fact 3.1: The share of patents with an assignee has steadily increased over time.

Figure 13 shows the share of patents whose rights were assigned, in full or in part, to a third-party. A third-party could be an individual or a company that commissioned or sponsored the development of the invention described in the grant. The share of patents without assignee has been shrinking over time and has been less than 20% for the past 40 years. The increasingly capital intensive nature of R&D activities or a trend towards market concentration might be at the root of this trend.

Figure 12: The figure shows the share of patents for which at least two inventors live in the same city.

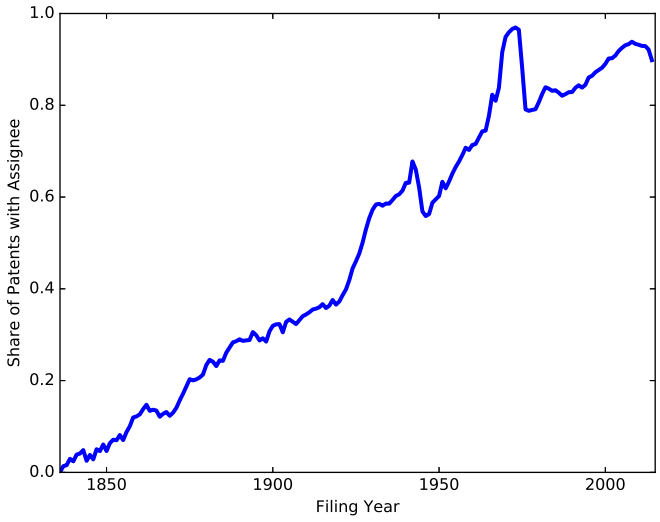
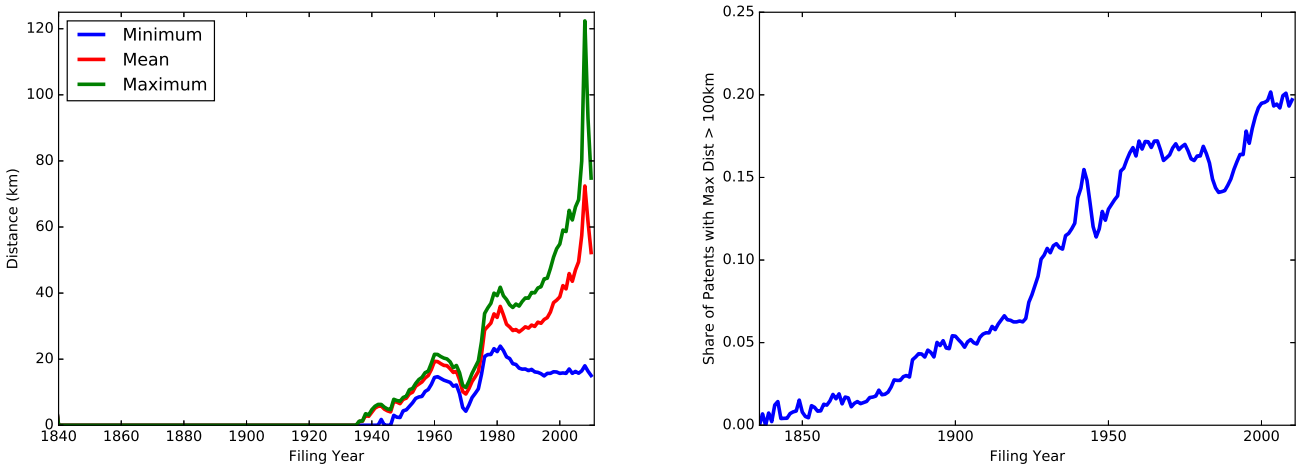


Figure 13: The figure shows the share of patents by filing year that were assigned, in full or in part, to at least one person (or company) different from the inventors.



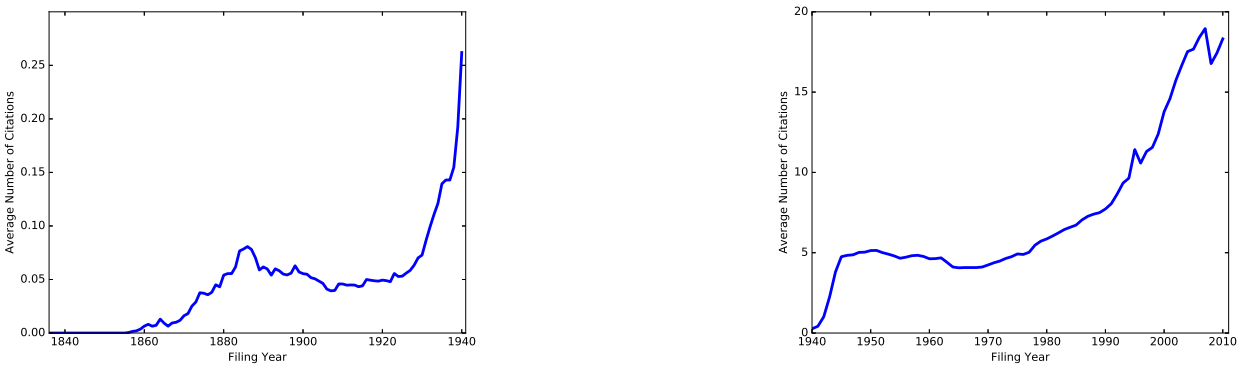
(a) (b)

Figure 14: The figure analyzes the distance patterns between assignees and inventors. The left panel shows the minimum, mean, and maximum distance between the inventors and assignee for the median patent. The right panel reports the share of patents for which the distance between the assignee and at least one of the inventors is larger than 100 kilometers.

Fact 3.2: Average and maximum distance between inventors and assignees of the median patent started an upward trend in the 50s; minimum distance increased at first and then plateaued.

Similarly to what we did for inventors, it is possible to analyze the distance between inventors and their assigness. A priori it is not obvious what to expect. On the one hand, the advantage in terms of resources of big firms and a tendency towards concentration (see for example Grullon, 2017) should reduce the average distance between inventors and assignees. On the other hand, more outsourcing and the decline in communication and transportation prices should work as a centrifugal force. Figure 14 suggests that centrifugal forces dominate centripetal ones. The left reports the evolution of the minimum, average, and maximum distance between the inventors their assignees for the median patent in the sample.[25] The right panel shows the share of patents for which the maximum distance between the inventors and their assignee is at least 100 kilometers. The graphs show a clear tendency towards decentralization, although the minimum distance has remained constant since the 1980s. Similarly to what was argued for Fact 2.1, it might the case that R&D operations are directed by researchers working for the assignee and some specific tasks are outsourced to other labs.

[25]Note that these three statistics coincide for solo authored patents with an assignee.



(a) Pre-1940

(b) Post-1940

Figure 15: The two figures report the average number of citation by filing year. The left panel shows the series for the years between 1836 and 1940, whereas the right panel for the years after 1940.

4.4 Citations

Fact 4.1: The average number of backward citations per patent has steadily increased over time.

The average number of patents cited by each patent has been steadily increasing over time. Figure 15 shows this trend over time. The left panel shows the series for the years between 1836 and 1940, whereas the right panel for the years after 1940. The data are split into two figures to take into account the introduction of a mandatory section containing the list of references cited in 1947 that has importantly increased the number of citations successfully extracted from the data. As it is possible to see in the left panel of Figure 15, the average number of citations prior to the mandatory disclosure of the references is order of magnitudes smaller, but not negligible. Nevertheless, as the right panel of Figure 15 highlights, the amount of references obtained in this way is probably only a small fraction of the full list of prior art considered when examining the patent.[26]

Citations might have steadily increased over time for two main reasons. First, digitalization have made it easier for inventors and reviewers to find inventions related to the one described in the patent. This would explain the acceleration in the average number of citations after 1980. Second, the number of inventions upon which newer inventions are built on has also increased over time. Inventions have become increasingly complex and if in the past a new idea relied on basic knowledge, nowadays it builds on a large number of previous discoveries (e.g., Jones, 2009). Such an increase in complexity would translate in an increase in the number of references.

[26]Future research should investigate the informativeness of pre-1947 citations.

Figure 16: The figure shows the share of patents that have not received any citation at any point in time after being filed.

Fact 4.2: The share of patents without forward citations was around 90% until 1910; it has then declined to 10% and remained mostly stable.

The share of patents without forward citations has dramatically decreased between 1910 and 1940, but was stable in the years before and after this period. Before 1910 about 90% of patents did not receive any citation since they were filed, whereas after 1940 this share was around 10%. Figure 16 reports this pattern over time. The low share before the 20s might be related to the introduction of the mandatory reference sector in 1947. More interesting is the extremely low share of patents without forward citations in the second half of the 20th century. Three facts might explain this trend. First, higher patenting costs might have contributed to attract more meaningful patents. Second, the second part of the 20th century witnessed a significant increase in the number of foreign patents filed in the United States. Because of the costs involved in the patenting process, it is usually believed that grants filed to multiple patent offices are particularly valuable. Finally, there might have been an increase in the amount of self-citations. As suggested by Jones et al. (2007), an increase in the number of inventors per patent is likely to increase the number of self-citations. It could be interesting to explore these explanations in future research.

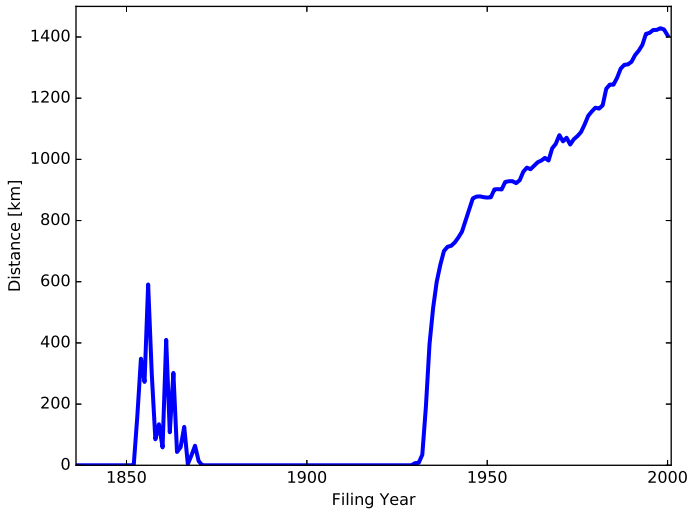


Figure 17: The figure shows the average distance of citations received by the median patent in the 10 years following its filing.

Fact 4.3: The average distance of citations received by the median patent in the first 10 years after filing was 0 up to the 40s; it has been increasing ever since.

Figure 17 analyzes the average distance of the citations received by the median patent in the 10 years after being filed. More precisely, I calculate the average distance between first inventors for each patent filed in each year. The figure reports the median of this distribution. With the exception at the beginning of the sample which is manly due to the small number of citations in the 19th century, the series shows a clear upward trend that started in the mid-40s and is still ongoing. This trend seems to support the idea that the decreasing cost of communication facilitate the diffusion of knowledge across space.

4.5 Classes

Fact 5.1: In the past 200 years only 9 classes made it to the top 1% terms of citations received per decade.

Given the long time span provided by these data, we can ask what technologies were the most valuable in each decade. To do so, I exploit a standard measure of patent relevance used in the literature, namely the number of citations received by each grant. More precisely, I rank all the patents filed in each decade by the number of citations received and I select those in the top percentile. I define the most frequent principal class among the patents selected as the leading technology for that decade.[27] Table 2 reports

[27]Using the top 5%, instead of the top percentile, leads to similar results. Berkes et al. (2018) explore more refined definitions of leading technology exploiting the network structure of patent citations.

Decade	Leading Class	Description
1836-1845	E02	Hydraulic Engineering; Foundations
1846-1865	F16	Engineering Elements or Units
1866-1875	A01	Agriculture
1876-1885	D05	Sewing
1886-1895	B41	Printing
1896-1905	D03	Weaving
1906-1945	F16	Engineering Elements or Units
1946-1995	A61	Medical or Veterinary Science; Hygiene
1996-2015	G06	Computing; Calculating; Counting

Table 2: The table reports the leading technology for each decade from 1836 to 2015. A leading technology is defined as the most frequent technological class in the top percentile of the distribution of citations received.

the results of this procedure. The table highlights two interesting facts. First, despite its simplicity, this methodology is able to capture the well-known technological waves in the United States over the past two centuries. The industrial revolution at the beginning of the twentieth century, the rise of medical science after the second world war with the development of vaccines and antibiotics, and finally the digital revolution in the second part of the 90s. Second, the length of the technological waves seems to have increased over time. Although this might be due to the nature of the data that are more noisy at the beginning of the sample, this fact might be explained by two other observations. On the one hand, it might be that since innovation becomes more complex in every field over time, it is more rare to have a breakthrough that moves the center of gravity towards another technology. On the other hand, it might be that the more recent waves enjoy more ideas to build upon and it takes longer to exhaust their creative momentum.

5 Conclusions

Since Hall et al. (2001), patents have been the preferred measure of innovation in the literature. The more than 3000 citations received by that paper alone in less than 20 years testify the high-demand for high-quality data on the topic. Because of the new opportunities offered by newly released or collected historical data, such as the historical decennial Census of Population, researchers have started moving their attention to pre-1976 data. In the past few years, the efforts to digitalize and extract meaningful information from historical patents have multiplied. The lack of a single data set that offers all the variables of interest collected with a consistent methodology and the fact that these data are sometimes not share with the rest of the community might constitute an important barrier for researchers who do

not have access to them. This paper fills this gap and describes a freely available newly assembled data set of historical patents containing all the variables usually employed in the literature. I anticipate that some issues might surface at the beginning when using them for actual research, the same way I found and fixed some problems while writing Section 4. Based on the feedback I will receive, I expect to make the data set more reliable over time and potentially include additional variables. The comparison with HistPat performed in Section 3 validates the data at least from a coverage and geographical points of view. Finally, some of the stylized facts presented in Section 4 show that the data are able to replicate some already well-known trends in the literature and gives a novel historical perspective to them. Some others are new and could spur ideas for future research.

References

[1] Akcigit, Ufuk, John Grigsby, and Tom Nicholas. 2017. "Immigration and the Rise of American Ingenuity." *American Economic Review: Papers and Proceedings*, 107(5): 327–331.

[2] Alcácer, Juan, Michelle Gittelman, and Bhaven Sampat. 2009. "Applicant and examiner citations in U.S. patents: An overview and analysis." *Research Policy*, 38(2): 415–427.

[3] Andrews, Michael. 2017. "Comparing Historical Patent Datasets." https://sites.google.com/site/michaeljeffreyandrews/research.

[4] Berkes, Enrico, Ricardo Dahis, and Marti Mestieri. 2018. "Technology and City Cycles." Mimeo.

[5] de Solla Price, Derek. 1963. Little Science, Big Science. New York: Columbia University Press.

[6] Grullon, Gustavo, Yelena Larkin, and Roni Michaely. 2017. "Are U.S. Industries Becoming More Concentrated?" https://ssrn.com/abstract=2612047.

[7] Hall, Bronwyn H., Adam B. Jaffe, and Manuel Trajtenberg. 2001. "The NBER Patent Citations Data File: Lessons, Insights and Methodological Tools." NBER Working Paper No. 8498.

[8] Jones, Benjamin F. 2009. "The Burden of Knowledge and the "Death of the Renaissance Man": Is Innovation Getting Harder?" *The Review of Economic Studies*, 76(1): 283–317.

[9] Moser, Petra. 2005. "How Do Patent Laws Influence Innovation? Evidence from Nineteenth-Century World's Fairs." *American Economic Review*, 95(4): 1214–1236.

[10] Nicholas, Tom. 2010. "The Role of Independent Invention in U.S. Technological Development, 1880-1930 ." *Journal of Economic History,* 70(1): 57–82.

[11] Packalen, Mikko, and Jay Bhattacharya. 2015. "Cities and Ideas." NBER Working Paper No. 20921.

[12] Petralia, Sergio, Pierre-Alexandre Balland, and David L. Rigby. 2016. "Unveiling the geography of historical patents in the United States from 1836 to 1975." *Scientific Data*, 3.

[13] Sarada, Michael Andrews, and Nicolas L. Ziebarth. 2017. "The demographics of inventors in the historical United States." Unpublished.

[14] Wuchty, Stefan, Benjamin F. Jones and Brian Uzzi. 2007. "The Increasing Dominance of Teams in the Production of Knowledge." *Science*. 316(5827): 1036-1039.