

This is the pre-peer reviewed version of the following article: Andrews, Michael J.; Historical patent data: A practitioner's guide; Journal of Economics & Management Strategy 30,2, pp 368-397 (2021); <https://doi.org/10.1111/jems.12414>, which has been published in final form at <https://doi.org/10.1111/jems.12414>. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions. Access to this work was provided by the University of Maryland, Baltimore County (UMBC) ScholarWorks@UMBC digital repository on the Maryland Shared Open Access (MD-SOAR) platform.

Please provide feedback

Please support the ScholarWorks@UMBC repository by emailing scholarworks-group@umbc.edu and telling us what having access to this work means to you and why it's important to you. Thank you.

Historical Patent Data: A Practitioner's Guide*

Michael Andrews[†]

August 11, 2020

Abstract

I provide a primer on six recent large-scale historical patent datasets for use in innovation research. I discuss how each dataset is constructed, the types of patent information included in each, and the quality and the completeness of each. Throughout, I emphasize when our knowledge of the history of invention is dependent on the data source used and provide recommendations about which dataset is most likely to be best for different contexts. Overall, these datasets paint a remarkably consistent picture of the history of U.S. invention. When the datasets do disagree, these differences tend to be minor, although I highlight some important exceptions. I further describe several “niche” historical patent datasets that allow researchers to study institutional contexts that cannot be studied using modern data. Finally, I discuss features of patent data that are not available for the historical patents but are available for modern patents.

JEL Classification: O30, O34, N00, Y1

*I am very grateful to the Balzan Foundation, the Northwestern Center for Economic History, and the NBER for financial support. I further thank seminar participants at the Social Science History Association meetings and Enrico Berkes, Dan Gross, Adam Jaffe, Jim Miller, Sergio Petralia, Elisabeth Perlman, Gaétan de Rassenfosse, and Nicolas Ziebarth for graciously sharing data and/or providing thoughtful feedback.

[†]National Bureau of Economic Research. *Email:* mandrews@nber.org

1 Introduction

Patent data has become an indispensable tool for innovation researchers. Over the last half decade, numerous researchers have embarked on data construction projects to obtain more information about patents (Balsmeier et al., 2018; Graham, Marco, & Miller, 2018; Graham, Marco, & Myers, 2018; Jaffe & de Rassenfosse, 2019; Wasserman & Frakes, 2019) and to link patent data to other data sources (Argente, Baslandze, Hanley, & Moreira, 2019; Baron & Pohlmann, 2018; Bell, Chetty, Jaravel, Petkova, & Reenen, 2019; Graham, Grim, Islam, Marco, & Miranda, 2018; Marx & Fuegi, in press).

Each of these above studies uses data on patents that issued within the last 45 years. In 1975, the United States Patent and Trademark Office (USPTO) began recording patent data in a digital format, greatly facilitating the use of patent data by researchers. Of course, U.S. invention did not begin in 1975. Recently, several research teams have begun constructing large-scale historical datasets to extend access to easy-to-use patent data backwards in time. These datasets come from various sources of raw patent data and rely on different techniques to parse relevant patent information from text that is often of poor quality.

In this paper, I seek to provide a practitioners' guide for the burgeoning new historical patent datasets. I summarize six accessible historical patent datasets, describing how they are constructed, what types of patent-level information they contain, and the strengths and weaknesses of each. Understanding what can be accomplished with these existing data prevents researchers from "reinventing the wheel" and building historical datasets from scratch, duplicating previous efforts.

Historical patent data have become an increasingly valuable tool for scholars of innova-

tion. The use of historical data provides more opportunities to exploit natural experiments as well as to track the long term effects of innovation policies. To list just a few examples making use of historical patents, Rosenberg and Nelson (1994), Mowery, Nelson, Sampat, and Ziedonis (2001), Mowery and Sampat (2001), B. N. Sampat (2006), and Andrews (2020b) explore the changing role of universities in patenting throughout the 20th century; Acemoglu, Moscona, and Robinson (2016) and Berkes and Nencka (2019) study the effect of other institutions such as post offices and libraries on invention; Furman and MacGarvie (2009, 2007) study the role of new organizational forms on patenting in the early pharmaceutical industry; Trajtenberg (1990), in a now classic paper, examines computed tomography scanner patents dating to the early 1970s to document the correlation between patent citations and patent value; Moser (2005, 2011) uses data from historical World Fairs to study the effect of patent protection on the rate and direction of invention; Lampe and Moser (2010, 2012, 2013) use the early sewing machine industry to illustrate the effects of patent pools; and Andrews (2020a) exploits alcohol prohibition in the first half of the 20th century to investigate the importance of informal social interactions for invention. In all of these studies, the results are only as credible as the patent data used. The credibility and feasibility of future work likewise depends on understanding the quality and availability of historical patent data.

The six datasets examined in detail in this paper are:¹

¹I am aware of other groups of researchers attempting to construct comprehensive historical patent datasets: a dataset constructed by Tom Nicholas (see, for example, Akcigit, Grigsby, and Nicholas (2017a, 2017b)), a dataset constructed by Mikko Packalen and Jay Bhattacharya (Packalen & Bhattacharya, 2015a, 2015b), and separate projects by Dimitris Papanikolaou and coauthors (Kelly, Papanikolaou, Seru, & Taddy, 2020a, 2020b) and Tania Babina and coauthors (Babina, Bernstein, & Mezzanotti, 2020) that use the CUSP for validation and as a complement; these datasets share many features of the CUSP data, and so I do not discuss them separately here. At present I have been unable to examine the underlying data or aggregate summary statistics from any of these datasets, at least two of which are still in varying stages of being “under construction,” and so I do not include them in the analysis below.

1. **Comprehensive Universe of U.S. Patents (CUSP):** Described in Berkes (2018), this dataset contains U.S. patents issued from 1836-2015. The dataset is constructed from a number of sources, primarily high-quality USPTO patents images.
2. **HistPat:** The first version of this dataset is described in Petralia, Balland, and Rigby (2016). The data contains issued U.S. patents filed from 1790-1978, also collected from USPTO-digitized patent images.
3. **Sarada-Andrews-Ziebarth (SAZ):** Described in Sarada, Andrews, and Ziebarth (2019), this dataset contains U.S. patents issued from 1870-1942. The data is collected from the Annual Reports of the Commissioner of Patents and Annual Indices of Patents.
4. **Jim Shaw:** This dataset contains U.S. patents issued from 1836-1873 and was compiled from the Subject-Matter Index of Patents for Inventions Issued by the United States Patent Office from 1790 to 1873 (Leggett, 1874). The data from the Subject-Matter Index were transcribed by hand by Dr. Jim Shaw of Hutchinson, KS.
5. **Kogan-Papanikolaou-Seru-Stoffman (KPSS):** Described in Kogan, Papanikolaou, Seru, and Stoffman (2017b), this dataset contains U.S. patents from 1926-2019 linked to the Center for Research in Security Prices (CRSP)-Compustat merged data and provides estimates of each patent's private value constructed from the response of assigned firms' stock market response to news about the patent issuance.
6. **USPTO Historical Patent Data File (HPDF):** Described in Marco, Carley, Jackson, and Myers (2015), this dataset, which is constructed from USPTO internal patent

records, contains all known utility patents from 1790 to 2014, along with patent classifications for each. The dataset lacks inventor, assignee, and geographic data available in the other datasets.

I show that these six datasets paint a remarkably consistent picture of invention along many dimensions. This is true whether examining absolute patent counts, inventors' collaboration behavior, invention by firms, or the geographic concentration of invention. By showing aggregate statistics across several datasets, this paper also provides an overview of U.S. patenting through time. I highlight when our knowledge of historical invention is particularly dataset-dependent, as well as when changes in Patent Office policy have affected the availability or reliability of different kinds of data and hence our ability to draw conclusions about the history of invention.

While each of these datasets document similar aggregate patterns, they are not identical. I document how often a particular patent number appears in one dataset but not another, as well as how often a given patent contains different information in one dataset relative to another. For example, I show that the issue date of a given patent is almost always the same regardless of which dataset one examines, while inventors' location information is more likely to be recorded differently in different datasets. When differences across datasets do occur, I document common causes. These discrepancies are relatively rare and hence are unlikely to affect researchers' conclusion in many contexts. For instance, I conclude that the CUSP dataset has correct inventor location information for more than 95% of its patents and the HistPat has correct location information for about 97.5% of patents. Nevertheless, I suggest that researchers ensure their results are robust to using alternative historical patent

datasets, especially given the accessibility and ease of use of these datasets.

Next, I briefly survey several historical patent datasets that cover unique but limited aspects of patenting. First, I provide information on design and plant patents. Second, I describe available data on America's first patents, the records for many of which were destroyed in a fire at the USPTO in 1836. These patents are interesting both because they show the inventiveness of a young nation and because, prior to 1836, the U.S. had a patent registration system rather than a patent examination system. Third, I describe data on patents issued by the Confederate States Patent Office, which operated in southern states during the U.S. Civil War. These alternative datasets provide a window on invention under unique institutions and policies, several of which are not available to study today.

Finally, I discuss features of patent data that are available for modern patents but are not found in the historical patent data. In particular, micro data on patent applications as well as information on patent lawyers and details on the patent prosecution process are not available for historical patents. For many periods through history, common measures of patent quality such as citations are either not available or are of questionable usefulness.

This paper is organized as follows. Section 2 describes the sources for the underlying raw historical patent data and describes how each of the six historical patent datasets are constructed. Section 3 describes what features of patents are included in each patent dataset and presents aggregate statistics across the datasets for each of these features. In Section 4, I explore the differences between datasets in more detail, quantifying when information in a given patent is recorded differently in different datasets. Section 5 describes additional historical patent datasets that highlight unique features of historical invention. Section 6 highlights the data limitations of historical patent data relative to modern patent data.

Section 7 briefly concludes.

2 Sources of Historical Patent Data

Figure 1 is an image of a historical U.S. patent document, in this case for the first rotating wheel can opener, invented by William Lyman of Meriden, CT, in 1870. Lyman’s can opener is a straightforward patent for a fairly simple invention. The patent consists of only one page (excluding figures), one claim, one inventor, and no assignees. Other patents can be much more complicated: they can be longer, have highly variable and frequently much worse image quality, contain multiple inventors or assignees, and reference prior patents, among other possible complications. All of the datasets described in this paper are attempts to collect relevant information originally recorded in patent documents like this one, and then to make that information available in a usable format.

Three datasets (the CUSP, HistPat, and KPSS) are built on high quality scans of patent documents such as Figure 1.² A problem common to all of these datasets is that patent images are often of poor quality, especially for older patents. Figure 2 shows a screenshot from the Google Patents record for William Lyman’s can opener patent. In the Google Patents screenshot, William W. Lyman’s name is recorded as “wmi’iAM w; LYMAN” and Meriden, Connecticut as “Mnifninllv.‘ ooNNEoTloU’r;” neither of which is particularly accurate. In

²In addition to these patent images, the CUSP also utilizes several local databases, including the Cincinnati Inventors Database, Iowa Inventors Database, Nevada Inventors Database, Oklahoma Inventors Database, South Carolina Inventors, the Portal to Texas History, and the Wyoming Inventors Database. I do not discuss these other databases here, except to note that in many cases, a great deal of information about inventions and inventors from specific geographic regions may be available, often through local universities, public libraries, or historical societies (Berkes, 2018). I use the November, 2019 version of CUSP (stable version 2.0). To access the CUSP, contact Enrico Berkes. For this paper, I use HistPat Version 8.0 from January, 2019, downloadable at <https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/BPC15W>. The KPSS data is downloadable at <https://github.com/KPSS2017/Technological-Innovation-Resource-Allocation-and-Growth-Extended-Data>.

attempts to minimize these problems, all three datasets run their own OCR algorithms on the patent images which improve on the OCR conducted by Google Patents or the USPTO. The CUSP also makes use of higher quality scans than are available in the PDFs provided by Google Patents. In spite of these best efforts, OCR errors are inevitable.³

The SAZ data, in contrast, are built from a source of raw patent data that do not contain the full text of each patent, and so avoid these challenges. The Annual Reports of the Commissioner of Patents and Annual Indices of Patents (The Commissioner of Patents, Various Years) forms the basis of the SAZ data.⁴ These Annual Reports list only identifying patent information such as the inventors' names and locations, rather than containing the entire patent text. An example of a page from the Annual Report of the Commissioner of Patents from 1888 is shown in Figure 3. Sarada et al. (2019) use OCR to convert these images to text files to create the SAZ dataset.

The Jim Shaw data is built from the Subject-Matter Index of Patents for Inventions Issued by the United States Patent Office from 1790 to 1873 (Leggett, 1874). The Index lists every known USPTO patent granted between 1790 and 1873 alphabetically by invention name, and was transcribed by hand by Dr. Jim Shaw of Hutchinson, KS, for all years from 1836 onward (Shaw, n.d.-b).⁵ An example from the Index is presented in Figure 4. Because

³The goal of the authors of the CUSP, HistPat, and KPSS datasets is to parse key patent and inventor information, but the body of patent text itself could be used for other forms of textual analysis; see Kelly et al. (2020a, 2020b), Perlman (2015), Packalen and Bhattacharya (2015a), and Packalen and Bhattacharya (2015b) for examples of this kind of work. None of these cited papers use the HistPat or KPSS data. Kelly et al. (2020a) uses CUSP data in conjunction with other patent data scraped and processed by the authors. Perlman (2015) uses Google Patent data collected by Tom Nicholas. The papers by Mikko Packalen and Jay Bhattacharya use a dataset constructed by those authors. At present I do not have access to these datasets or any detailed information regarding how they were constructed, so I do not discuss them in this paper.

⁴SAZ data are available at <https://www.openicpsr.org/openicpsr/project/120556/version/V1/view>. The Annual Reports were downloaded from <https://catalog.hathitrust.org/Record/002138126> and <https://library.si.edu/digital-library/book/annual-report-commissioner-patents-year>.

⁵The Jim Shaw data was downloaded from <http://www.ptrca.org/history> on August 5, 2020. The Patent and Trademark Resource Center Association (PTRCA) site contains a number of other interesting

Dr. Jim Shaw transcribed this dataset by hand, it does not suffer from the OCR issues that occur when parsing the other raw data sources, although the index can often be difficult to decipher even for a human reader.

There is one important difference between the raw data used for the SAZ and Jim Shaw datasets on the one hand, and raw data used for the CUSP, HistPat, and KPSS datasets on the other. The former contain only select pieces of relevant patent information—inventor names, locations, etc.—and *nothing else*, while the latter contain the entire text of each patent. The former have a clear syntactic structure that makes it relatively easier to parse out the relevant inventor information. The latter are unstructured text and therefore harder to identify particular pieces of information. To see this, consider attempting to identify the inventor’s name from Figure 1. The patent text contains not only the name of the inventor, William Lyman, but also names of witnesses, in this case Ratcliffe Hicks and R. H. Foster. If the inventor’s name is not legible elsewhere, CUSP searches the end of the text for a name, introducing the possibility that a witness’s name may be erroneously recorded (although I should note that in the case of Lyman’s can opener, the inventor name is recorded correctly). Likewise, some patent texts may mention other towns, counties, or states that are different from the location of the invention, which HistPat may assign a high probability of being the correct location. But the fact that the CUSP, HistPat, and KPSS datasets make use of the entire patent data means they can potentially pull out many more pieces of patent information or conduct textual analysis; SAZ and Jim Shaw are limited to the pieces of information included in the Annual Reports and Subject Matter Index, respectively. Finally,

facts about the history of patenting in the US, and the members of the association have been very helpful in providing details about the construction of various data sources. Details on the creation of the Jim Shaw data come from personal communications with Jim Miller, a PTRCA member and the resident expert on the Jim Shaw patent data. To date, I have been unable to get in contact with Dr. Jim Shaw directly.

because the SAZ and Jim Shaw data are presented in a much more compact form, with many entries per page, poor image quality on one page will make multiple patent records illegible; see Appendix B. When instead using the entire patent text, key information may be recorded at different places on the patent document; this redundancy minimizes the risk that poor image quality on part of a page will result in lost information.

The final historical patent dataset is the USPTO Historical Patent Data File (HPDF).⁶ The HPDF contains administrative data created by the Patent Office to facilitate examination and prosecution of the patent, including patent classification and relevant filing, issuance, and expiration dates. Marco et al. (2015) describe the features of the HPDF in more detail.

3 Comparing Aggregate Patent Statistics Across the Datasets

The previous section describes how each patent dataset is created. Here, I describe the data that they contain. Table 1 presents the features of patents that are available in each dataset. In general, the CUSP and SAZ datasets contain inventor and assignee names and locations; the HistPat is focused primarily on location information and does not contain names; the KPSS is focused on patent quality measures and assignee information and does not contain data on inventors or locations; the HPDF only contains information on patent classes and does not have information on inventors, assignees, or location. The CUSP also contains data

⁶The HPDF is downloadable at <https://www.uspto.gov/learning-and-resources/electronic-data-products/historical-patent-data-files>. The data used in this paper were downloaded on July 7, 2020.

on patent classifications and citations. I describe each of these features in more detail below.

3.1 Patent Issue Dates

Perhaps the most basic question to ask about the historical patent datasets is what time periods do they cover? All of these datasets contain the year in which each patent issues. The CUSP contains information on patents issued from 1836 to 2015. The HistPat contains patents issued from 1790 until 1975, when the USPTO began keeping digital records. The SAZ data contains patents issued from 1870 to 1942; Annual Reports are available for other years as well, although they are not yet incorporated into the SAZ.⁷ Because the Jim Shaw data is based on the Subject-Matter Index published in 1874, it only includes patents from 1836 to 1873. The KPSS dataset is designed to match patents to firm names in the CRSP-Compustat merged data, which contains names of publicly traded firms from 1926 onward. For this reason, the KPSS data begins in 1926 and continues to 2019. Finally, the USPTO HPDF contains information on patents issued from 1790 through 2014.

While the year of issuance is a key piece of information, it may also be useful to have the exact date on which a patent issues.⁸ The CUSP, KPSS, Jim Shaw, and HPDF datasets contain the day, month, and year of patent issuance, whereas the SAZ and HistPat datasets only contain the year.

When visualizing the information in the patent datasets below, it is often useful to plot aggregate patenting outcomes over time. I do this for the years 1837 to 1975, even though some of the datasets include patents issued outside of this range. I begin in 1837 for two

⁷Sarada et al. (2019) are unable to locate an Annual Report from 1874, so that year is missing in the SAZ data.

⁸Kogan et al. (2017b) is one example of an application in which the exact issue date is important, as is Gans, Hsu, and Stern (2008) using recent patent data.

reasons. First, 1836 marked a major change in the patent law, in which the U.S. started examining patent applications for novelty and non-obviousness. Second, a major fire in 1836 destroyed many of the earliest patents. Thus 1837 is the first full year with the modern patent system and complete patent records. I end in 1975 because the USPTO began keeping digitized records the following year.

3.2 Patent Counts

The next most obvious dimension on which to compare these datasets is simply how many unique patents they contain. When the USPTO grants a patent, it issues a unique patent number. This patent number is included in every dataset except for the SAZ dataset, for which it proved impossible to parse patent numbers from the Annual Reports. Instead, the SAZ data uniquely identify patents based on the line in which they appear in the Annual Reports.⁹

I plot the number of unique patents for each of the six datasets in panel (a) of Figure 5. For most years, the HPDF contains the largest absolute number of patents, which is not surprising since this is data provided by the USPTO and is based off of administrative records. For many years, the difference in number of patents between the CUSP and HPDF is small and difficult to observe in Figure 5. HistPat contains fewer patents than the CUSP and HPDF for all years, although in early years these differences are typically only a few

⁹As Figure 3 shows, each record in the Annual Reports contains a patent number, patent grant date, and information on how to find the patent in the USPTO monthly publications for each patent. Unfortunately, these fields are recorded in columns separated by thick printed vertical lines. These lines confuse the OCR software, making it impossible to link, for instance patent number with the inventor name and location. The problem is compounded by the fact that cells are often left empty, so that one cannot simply match by row numbers. We have experimented with several different OCR software tools, and this problem is endemic to all. Because of this, the SAZ data does not include patent numbers.

dozen patents. The Jim Shaw data contains roughly the same number of patents as does the HPDF in all of the years for which it has data; it actually contains a few more patents than does the HPDF for most years and never has more than eight fewer patents than the HPDF. The SAZ dataset typically has fewer patents than the HPDF, CUSP, HistPat, or Jim Shaw datasets for common years. The fewest patents are in the KPSS data. This is also not surprising because KPSS only contains patents assigned to publicly traded firms and so is not designed to contain the universe of patents.

Across all of the datasets, the cyclicity of patenting noted by Griliches (1990) is apparent. The absolute number of issued patents tends to grow over time, but each time series exhibits declines during economic downturns, wars, and periods of decline in the number of patent examiners working at the USPTO. Schmookler (1966) also highlights a change in attitudes toward patenting and a shift towards industries that relied more on trade secrecy than patents to explain the large decline and then stagnation in the number of issued patents observed in the CUSP, HistPat, and HPDF datasets from 1930 to about 1960 (with the large drop in the early 1940s due to World War II).¹⁰

To put these patent counts into perspective, in panel (b) of Figure 5, I plot the share of total U.S. patents contained in each dataset and each year. Aggregate patenting counts in each year are provided by the USPTO.¹¹ In most years, the HPDF contains exactly the same number of granted patents as reported in the aggregate counts. In years where the numbers are not exactly the same, they differ by only a few patents. Thus the HPDF is indeed close to complete. Close to 100% of patents from 1836 to 1873 are included in the

¹⁰The SAZ shows a large decline in patents relative to the other datasets from about 1910 to 1930; rather than reflecting a real decrease in patenting during this period, this appears to be driven by poor image quality in the Annual Reports for those years.

¹¹See https://www.uspto.gov/web/offices/ac/ido/oeip/taf/h_counts.htm.

Jim Shaw dataset as well. The CUSP is similarly complete, containing more than 99.4% of all U.S. patents in every year. The fraction of aggregate patents contained in the HistPat is lower than these two datasets but still impressive, ranging from a high of 95.8% in 1947 to a low of 51.6% in 1974, with most years having more than 95% of all patents and the fraction falling below 90% only for 1973-1975. In all but two years, SAZ has a smaller share of aggregate patents than either CUSP or HistPat, with the share varying from a high of 95.1% in 1909 to a low of 41.3% in 1915. The KPSS dataset contains a lower share of patents than any of these other datasets, although I again emphasize that the KPSS was not constructed to contain all patents.

3.3 Inventors

In many contexts, we want to know something about the individuals who create patents. The CUSP, SAZ, and Jim Shaw datasets contain information on inventors' names; the HistPat, KPSS, and HPDF datasets do not.¹²

As noted in Section 2, the CUSP dataset uses the entire patent text to determine inventor names. The CUSP first searches for the end of the patent body text, which contains the inventors' names in all capital letters, minimizing the chance of OCR-induced errors. If this information cannot be located, CUSP searches each patent's header string; here, the inventor's name should proceed the word "of" followed by a town and state, as in "WILLIAM W. LYMAN, OF MERIDEN, CONNECTICUT" in Figure 1. If this cannot be located either, CUSP finally searches the body of the patent description, which begins with a declaration

¹²Even though inventor name is not included in HistPat, the authors have parsed name information. Sergio Petralia has graciously shared their inventor name raw data with me upon request. As the inventor names are still quite raw and are not included in the current downloadable release, I do not use them in this analysis.

that includes the inventors' names and location information. For example, William Lyman's can opener patent begins: "Be it known that I, WILLIAM W. LYMAN, of Meriden, in the county of New Haven and State of Connecticut, have invented a new Improvement in Can-Opener..."

Inventor names are easier to locate in the raw data used for the SAZ and Jim Shaw datasets. For SAZ, each row in the Annual Reports begins with the inventor's name, followed by their location of residence and the name of the invention (Figure 3). The SAZ dataset was designed with the goal of linking inventor names to historical U.S. census records (see Sarada et al. (2019) for complete details). A set of matched patent-census data is available from these authors, providing a resource for researchers who want to study inventor demographics at different times and places without conducting the computationally intensive linking exercise themselves.¹³

In the Subject-Matter Index used to construct the Jim Shaw dataset, the inventor's name for each patent is listed in the second column (Figure 4). The Subject-Matter Index, and therefore the Jim Shaw dataset, does have one key drawback relative to the Annual Reports and the full patent text: only the first letter of each inventor's first and middle name is provided. This renders the Jim Shaw data unusable when name-matching is required, for instance to match patentees to the census as in Sarada et al. (2019).

While the CUSP, SAZ, and Jim Shaw datasets contain inventor names, they have not been disambiguated. That is, the data do not contain unique individual identifiers that would allow one to easily count how many patents a given individual has obtained over their life.

¹³These census matched data are available at <https://www.openicpsr.org/openicpsr/project/109970/version/V1/view>.

Numerous studies present different approaches to disambiguating modern patent records, typically relying on machine learning algorithms to assess the probability that two records belong to the same inventor; see, for instance, Trajtenberg, Shiff, and Melamed (2006), Li et al. (2014), and Monath and McCallum (2015). In some cases, the information used to train these algorithms is not available in historical patents (see a discussion of this in Section 6 below). The fact that historical patents can be linked to other historical records (like the U.S. census in Sarada et al. (2019)) provides an interesting alternative for disambiguation of historical patent records: once a patent has been linked to a census record, the same individual can be located in prior or subsequent censuses. Babina et al. (2020) adopt a similar approach for inventors during the Great Depression, and work is underway using these methods on the entirety of the CUSP and SAZ datasets.

Patents may have more than one inventor, and collaboration patterns may be particularly interesting to investigate. All of the datasets that record inventor names attempt to find names for all listed inventors. Even the HistPat data, which does not record inventor names, can be used to identify cases when a patent has more than one inventor: in the HistPat data, each inventor's location is listed as a separate row in the data file, and so patents with multiple rows have multiple inventors.

But just because a particular patent dataset records inventor information does not mean that it successfully records information for *all* inventors; for instance, inventor names that do not appear first may be harder to parse and more error prone. This issue is especially prevalent in the Annual Reports and Subject-Matter Index used to build the SAZ and Jim Shaw datasets. Because page space is at a premium in these reports, multiple inventors with the same surname are often listed together, for instance as “Smith, John and Mark,

Chicago, IL”. This makes it difficult to both successfully identify and parse the second inventor’s name.¹⁴

Figure 6 plots changes in inventor collaborations in each dataset over time. Panel (a) plots the share of patents that list more than one inventor. Prior to about 1920, the vast majority of patents had only one inventor, regardless of which dataset is used. In the CUSP and SAZ datasets, about 90% of patents had a single inventor in early years. HistPat reports that almost 100% of patents had only one inventor prior to the twentieth century, which is likely a result of a failure to record names of all but the first listed inventor in the early years. The Jim Shaw data reports rates of collaboration in between HistPat and the CUSP. The pattern changes in 1920, with both the CUSP and HistPat datasets showing fairly steep increases in the fraction of patents with more than one inventor.¹⁵

Panel (b) plots changes in the average number of inventors per patent. Together with the results in panel (a), the data for CUSP suggest that not only were collaborations becoming more common, but team sizes were growing larger as well. By the early 1970s, the CUSP data reports that the average patent had more than two inventors; since about 60% of patents still had only one inventor, many patents involved collaborations of more than two inventors. These findings are consistent with research on team size and the complexity of inventions (Jones, 2009; Wuchty, Jones, & Uzzi, 2007). In the other datasets, and in particular in HistPat, the number of inventors does not appear to increase by as much. From spot-checking patents, it appears exceedingly rare that any of the historical patent

¹⁴This formatting also makes it difficult to differentiate second inventor information from assignee and location information. I present results on these features of the patent data below.

¹⁵The SAZ data also shows a dramatic increase beginning around 1920, although this is preceded by a drop in the share of patents with more than one inventor around 1910. This is likely driven by the difficulty of separating out second inventor names in the Annual Reports mentioned above. A change in formatting and a substantial drop in image quality in the Annual Reports around 1910 exacerbated these issues.

datasets record more inventors than are actually present on a patent; it is far more common to omit an inventor. Thus, the number of inventors in the HistPat and Jim Shaw datasets are likely under-counted. For investigating collaborations between inventors, the CUSP likely provides the most complete picture.

3.4 Assignees

Prior research suggests that the same factors that were driving greater collaboration among inventors—increasing technological complexity and larger fixed capital costs to inventing—were also driving inventive activity inside firms (Babina et al., 2020; Chandler, 1990; Lamoreaux & Sokoloff, 2005; Mowery, 1990; Nicholas, 2010). Researchers typically proxy invention that occurs inside the firm by patents that are assigned to a firm at the time the patent issues. Panel (a) of Figure 7 plots the number of patents that list an assignee in the CUSP, HistPat, SAZ, and KPSS datasets; the Jim Shaw and HPDF datasets do not record patent assignment. All four datasets show increases over time, although the SAZ data is highly variable from year to year. The HistPat data does a poor job of recording assignments prior to 1900, but by 1920 reports numbers very similar to the CUSP.

Panel (b) plots the share of patents in each dataset that are assigned. The share in the CUSP and HistPat datasets are likewise very similar after 1920. The SAZ data does not show as dramatic an increase in the share of assigned patents. Both the CUSP and HistPat data show that the modal patent was assigned around 1930, consistent with Nicholas (2010).

The KPSS dataset only contains patents that are assigned to publicly traded firms, so the share of assigned patents is equal to one in the KPSS for all years. The linking between

patents and publicly traded firms is the key benefits of the KPSS dataset. The assignee names in KPSS have been disambiguated, in contrast to the inventor names in the CUSP, SAZ, and Jim Shaw data, as well as the assignee names in the CUSP and SAZ data. Determining when a reported assignee name refers to a particular firm is not trivial. Large firms such as General Electric, which owns more than 43,000 patents in the KPSS data, may report their name numerous different ways on a patent, in addition to spelling and OCR errors (Kogan, Papanikolaou, Seru, & Stoffman, 2017a, p. 2). The authors harmonize assignee names and then link them to firm names in CRSP-Compustat.¹⁶ These linked data have proven very useful to researchers, with the KPSS data cited more than 750 times as of this writing.

If we are willing to believe that the CUSP and HistPat datasets do a reasonably good job of counting all patent assignments, and that the KPSS data does a good job of identifying publicly traded firms that appear in Compustat, then the difference between the number of assigned patents in the CUSP or HistPat and the number of assigned patents in KPSS gives the number of patents that are assigned to individuals or non-Compustat firms. This is potentially an important feature of these data. Today, almost all patents are assigned to publicly traded firms and consequently a large literature examining corporate innovation has emerged (e.g., Tian and Wang (2014), Sunder, Sunder, and Zhang (2017)), but there is little research that examines patent ownership by small firms or individuals.

¹⁶The CRSP-Compustat data can be accessed at <http://www.crsp.org/products/research-products/crspcompustat-merged-database>. The name matching builds on an algorithm by Bessen (2008).

3.5 Location

Where invention takes place may be as important as who is doing it (Moretti, 2012). The KPSS and HPDF datasets do not contain inventor location information, but the other four datasets do.

CUSP finds inventor location information in the same way it finds the inventor names. But even once the inventor location string has been identified in the patent text, it often contains errors due to the OCR. CUSP therefore proceeds in three more steps to assign a final inventor location. First, if the town-state pair string matches exactly to known town-state names, this is assumed to be the correct location. Second, if the inventor’s town-state is similar, but not identical to existing town-state pairs, the pair with the most other patents attributed to it is assumed to be correct. Third, if the inventor’s town-state pair still does not match to any known town-state, CUSP searches for the same inventor’s name on other patents and finds the most likely patent location from that set. In addition to the inventor’s town and state, the patent text also often lists each inventor’s county of residence. Regardless of whether the county appears in the patent text, the CUSP data assigns each inventor to its current county based on the latitude and longitude of the town.

HistPat uses complementary approaches to infer inventor locations. HistPat first assembles a list of “candidate location” names within the body of the patent text based on known town, county, and state names. After manually assigning correct locations to a subset of patents to act as a training set, the HistPat authors implement neural network, probit, and k -th nearest neighbor algorithms to determine which of the candidate locations is most likely to be the correct location. For any remaining patents for which location cannot be inferred,

HistPat searches for other patents with the same inventor name and nearby year and patent class, similar to the final step of the CUSP procedure. Similarly to the CUSP data, the HistPat data records inventors' town, county, and state of residence.¹⁷

For the SAZ and Jim Shaw datasets, inventors' locations can be parsed from each row similarly to the inventor names. Both the Annual Reports and the Subject-Matter Index list inventor towns and states; unlike the CUSP and HistPat data, they do not contain counties. This omission is unfortunate, since much geographic data from U.S. history is recorded at the county level.¹⁸ While the Subject-Matter Index on which the Jim Shaw data is based lists names of multiple inventors, the Index typically only lists one location. For instance, in only one of the 7,877 patents with multiple inventors does the Jim Shaw data list more than one state of residence.¹⁹ Over the period that the Jim Shaw data cover, the CUSP data records 423 cases in which inventors on the same patent reside in different states. The Jim Shaw data therefore does not appear to be ideal for analyzing inventor collaboration across space.

One concern is that these different ways of recording and parsing inventor locations may lead to different conclusions about where U.S. invention takes place. To assess whether this is the case, in Figure 8 plots the geographic concentration of patenting over time for the four datasets that include inventor location information. I calculate concentration using a simple Herfindahl-Hirschman Index (HHI). Panel (a) plots HHI at the state level, panel

¹⁷The "baseline" HistPat dataset contains information only for inventors residing in the U.S. To this I append information from inventors residing outside the U.S. from a separate dataset published March, 2019, and available for download at <https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/QT40JS>. See also Petralia (2019).

¹⁸For instance, see the National Historical Geographic Information System (Manson, Schroeder, Riper, & Ruggles, 2019).

¹⁹This is patent #3,852, with inventors living in Illinois and Michigan.

(b) plots concentration at the county level, and panel (c) plots concentration at the town level. Because Annual Reports and the Subject-Matter Index on which the SAZ and Jim Shaw datasets are based, respectively, do not contain inventors' counties, panel (b) contains results only for the CUSP and HistPat datasets. In Appendix E, I discuss different methods to assign SAZ and Jim Shaw patents to counties, and in Appendix F I plot the county-level concentration for the SAZ and Jim Shaw data after matching to counties.

At the state level, all four datasets paint a remarkably consistent picture, with the geographic concentration of patenting falling from 1837 to about 1910, with a jump during the Civil War, a small rise between 1910 and 1940, and then a slight decline to 1975. County-level concentration results are similar, although the CUSP shows less concentration than the HistPat for all years, with a difference that is especially pronounced prior to 1900. The temporal pattern is again similar at the town level, and the CUSP again demonstrates less concentration than does the HistPat; in this case, the town-level concentration measures in the Jim Shaw and SAZ data are almost identical to that in the CUSP prior to 1900. This suggests that the HistPat data are attributing patents to too few towns relative to the other datasets, resulting in an artificially high concentration at the town level. The temporal pattern of geographic concentration presented in all three figures and across all four datasets—with a steep decline in concentration until the early 1900s followed by a slight rise until about 1940 and then further declines following World War II—are consistent with the patterns of geographic concentration of patenting documented by Andrews and Whalley (2020), who use a different measure of concentration based on Ellison and Glaeser (1997).

Appendix F presents several other results comparing the location of patenting across datasets. In sum, while some differences exist, these tend to be modest in magnitude.

Overall, in all of the datasets patents are distributed similarly across space throughout U.S. history.

3.6 Application Dates

While the date on which a patent issues is no doubt important, in many contexts it may be more useful to know when a patent was filed. Only the CUSP and HPDF data contain patent application dates. But the HPDF appears to contain a number of errors on this dimension. For 20,875 patents, the HPDF lists an application date that is later than the issue date; this should be impossible. Nor is this simply the case of, say, “1800” being mis-recorded as “1900”; I spot checked a number of records and found no relationship between the actual patent filing date and the application date listed in the HPDF.²⁰ For the analysis here, I record as missing any filing dates for which the application date is later than the issue date. After this adjustment, the HPDF contains no filing dates before 1920. Even in the CUSP, patent filing dates are not available for all years; prior to 1873, patent filing dates were not printed on the patent text (Berkes, 2018, p. 4).

Figure 9 plots the mean and median difference between patent issue year and filing year in both the CUSP and HPDF data. For both datasets and all years, the mean and median patent grant delay track each other closely, although the mean delay is usually greater than the median, indicating that some patents had very lengthy prosecution processes. Prior to 1920, most patents issued in the same year that they were filed or in the following year, with the median delay growing to three years by the mid-1940s in the CUSP data. The HPDF

²⁰For example, patent 441,324 was filed on June 30, 1890 while the HPDF lists June 1, 1965 as the application date; patent 1,307,087 was filed on November 8, 1916 while the HPDF lists June 1, 1945; and patent 3,760,914 was filed on March 24, 1972 while the HPDF lists June 1, 1975.

data records patent grant delays that are highly variable from year to year, with the median delay for instance jumping from nine years in 1954 to zero years in 1955. In light of this, as well as the other issues mentioned above, I do not recommend using the patent filing dates reported in the HPDF.

3.7 Patent Classes

It is often important to understand what types of technologies a patent covers. Today, the USPTO assigns each patent to a US Patent Classification (USPC) code when assigning a patent application to examiners. These USPC codes have been in use since at least the late 20th century, and the system has undergone numerous revisions.²¹ Both the CUSP and HPDF datasets contain information on the USPC patent classes and subclasses; the CUSP also contains international patent classification (IPC) and Cooperative Patent Classification (CPC) codes for each patent.

One drawback to these Patent Office-assigned classifications is that they are designed for the benefit of patent examiners to determine the similarity of technologies. Often, researchers are interested in the industries that will use patents, which may be unrelated to the technology class. Hall, Jaffe, and Trajtenberg (2001) constructed a classification based on industry of use, known as the NBER patent classification, that has been heavily used by researchers.²² In the HPDF dataset, the PTO economists developed a probability-matching algorithm to apply the definitions of the NBER patent classes to historical patents (Marco et al., 2015, p. 7-9). Thus, researchers looking for classifications related to the industry of

²¹<https://www.uspto.gov/page/classes-g9b-49> lists USPC codes still in use and the date at which they were established; the earliest are in 1899. US Patent and Trademark Office (2013) lists the many revisions made to the classification system from 1947 to 2013.

²²As of this writing, Hall et al. (2001) has almost 4,000 citations on Google Scholar.

use should turn to the HPDF dataset.

3.8 Invention Names

The SAZ and Jim Shaw datasets also contain each patent's name or title. Similar to the patent's classification, the title provides some information about what type of technology the invention covers. The invention's name may be more descriptive than the patent's class and may also give some hint about the invention's intended use. At the same time, the invention's name is far less descriptive than an abstract or the full patent text, nor is each title unique. For instance, the Jim Shaw data contain 30 inventions named "Can Opener" just from 1866 to 1873. The invention titles, especially in the SAZ data, are also subject to numerous OCR errors and so should be treated with care.

3.9 Citations

Of course, not all patents are created equal; some patents are more important than others, and many patents are not important at all. For decades, scholars have argued that the number of citations a patent receives is correlated with the value of the patent (Hall et al., 2001; Trajtenberg, 1990). Today, each patent lists citations to prior patents on its first page, acting similarly to a list of references in a scholarly article.

Both the CUSP and KPSS datasets contain the number of these citations received by each patent, giving researchers a patent-level proxy for patent quality. The mechanics of how CUSP and KPSS capture these first-page citations are detailed in Berkes (2018, p. 7) and Kogan et al. (2017a, p. 11), respectively. In both cases, the datasets are able to

identify a larger number of forward citations than are listed in Google Patents. The CUSP also indicates when citations are added by the patent examiner; using modern patent data, several researchers have used examiner-added citations to investigate knowledge flows and patent prosecution practices (Frakes & Wasserman, 2017; Kuhn, Younge, & Marco, 2019; Thompson, 2006).

For historical patents, care should be taken when using citations. Citations were not printed as a separate section of the patent until 1947. While a large fraction of patents issued before 1947 were cited after 1947, these are likely the highest quality patents. In the left tail of the quality distribution, there is no variation in citations since all patents receive zero. But just because citations are not listed on a patent's first page does not mean that a patent does not cite the prior patent literature. The CUSP parses each patent's description for any mentions of prior patents. CUSP is therefore the only one of these datasets that contains citation data before 1947. For the pre-1947 years, however, citations to other patents are rare and tend to be "self-cites," that is, citations to other patents by the same inventor.²³

3.10 Private Patent Values

While forward citations are correlated with private patent value (Trajtenberg, 1990), they are measuring something slightly different: the extent to which a patent contributes to future knowledge, capturing both the private and social value of a patent. In many contexts, especially in finance and accounting, having a measure of the purely private value of a patent is important. On this dimension, the KPSS data shines.

²³Ongoing work such as Kelly et al. (2020a), which builds a novel historical patent dataset making use of the patent text, uses the citation counts data from the CUSP.

Building on work such as Pakes (1986), Hall, Jaffe, and Trajtenberg (2005), and Nicholas (2008), Kogan et al. (2017b) use changes in stock prices for the firms in the CRSP-Compustat data following patent issuance to estimate the private value of obtaining a patent. The estimated value of each patent, constructed under multiple assumptions, is available in the KPSS data.

4 Exploring Differences Between Datasets

The previous section shows that, over many dimensions and with few exceptions, all six datasets present similar aggregate pictures of the history of U.S. patenting. But just because the datasets appear similar in aggregate does not mean they do not differ in important ways. In this section, I explore the differences between datasets in more detail.

4.1 Do the Datasets Contain the Same Patents?

First, I ask whether the datasets contain the same patents. One might think that, so long as the datasets appear similar in aggregate, it does not matter if they are made up of different underlying patents. This may be true in some contexts, for instance if one is only interested in establishing broad stylized facts. But in analyses in which leverage can come from a small number of particular patents, small differences across datasets may matter.

Every dataset except for SAZ contains each patent's number, making it easy to see if a patent number appears in one dataset but not another. Table 2 lists pairwise comparisons between the CUSP, HistPat, Jim Shaw, and HPDF datasets (I omit the KPSS dataset because it is designed to contain only a subset of patents). Each cell lists the number and

fraction of patent numbers that are contained in the row dataset but that do not appear in the column dataset. Because the Jim Shaw data ends in 1873, I compare Jim Shaw to the CUSP, HistPat, and HPDF datasets only for the years 1837 to 1873; the other comparisons are over the years 1837 through 1975.

The first thing to note from this table is how complete the Jim Shaw data are: from 1837 to 1873, neither HistPat nor HPDF contain any patent numbers that do not also appear in the Jim Shaw data, and the CUSP contains only one. The HPDF has similarly complete coverage over an even longer period of time: one patent number is in the CUSP but not the HPDF, ten are in HistPat but not the HPDF, and 303 are in the Jim Shaw but not the HPDF. The CUSP data also has impressive coverage, with only 39 patent numbers in the HistPat, 326 in the Jim Shaw data, and 42 in the HPDF that are not also in the CUSP. The CUSP contains 606,124 patent numbers that do not appear in the HistPat, the HPDF has 605,795 patent numbers, and Jim Shaw 11,954 patent numbers. To put this into perspective, about 15% of the patent numbers in both the CUSP and HPDF do not appear in the HistPat; about 8% of patent numbers in the Jim Shaw dataset do not appear in HistPat. To a first approximation, the HistPat dataset is a subset of the CUSP, HPDF, and (for the relevant years) Jim Shaw datasets.

Next, I drill down into the HistPat and CUSP datasets, manually inspecting patents that are in one dataset but not the other. I first look up the 39 patent numbers in the HistPat data that are not in the CUSP data. For 12 of the 39 patents, I cannot locate a patent image file in Google Patents; for these cases it is likely that the HistPat patent number was recorded with error. For six of the 39 patents, the location information recorded in HistPat differed from that in the Google Patents images, raising the possibility that the patent number was

recorded incorrectly for these patents as well.

Second, I draw a random sample of 150 patent numbers that are in the CUSP but not the HistPat. I was able to locate the patent image file in Google Patents for all 150 of these patent numbers, and in only about 3% of these patents does the location information in the CUSP differ from the Google Patents images; hence these do not appear to be “bad” patent numbers. In 12% of the patents in the CUSP but not the HistPat, the CUSP does not record any inventor or location information. The sample of 150 patents overwhelmingly contains inventors who reside outside the U.S.: 74% of the sample of patent numbers in the CUSP but not the HistPat had a foreign inventor, compared to 12% in the entire HistPat and 21% in the entire CUSP for the years they overlap. Thus, the HistPat data appears to undercount foreign inventors; see also Appendix G for more on foreign inventors.

4.2 Do the Patents Have the Same Information Across Datasets?

Even if a given patent number appears in multiple datasets, the information about that patent may differ. In this section, I examine how often a patent number appears in multiple datasets but other patent information differs across the datasets.

In panel (a) of Figure 10, I plot the number of patent numbers with different issue years between the CUSP and HistPat data, the CUSP and Jim Shaw data, and the HistPat and Jim Shaw data. Rarely does a single patent number list a different issue year in different datasets. With the exception of patents issued in 1926 in the CUSP, when 75 patent numbers record a different issue year in HistPat, fewer than 20 patents in each year record different issue years. In panel (b), I plot the fraction of CUSP patents in which the HistPat or Jim

Shaw patents have different issue dates than the CUSP, and the fraction of HistPat patents in which the Jim Shaw patents have a different issue date than the HistPat; in all cases these amount to far less than 0.1% of patents.²⁴

Figure 11 plots the number and fraction of patents in which the state of residence of the first listed inventor is different between the CUSP and HistPat data, the CUSP and Jim Shaw data, and the HistPat and Jim Shaw. In contrast to issue dates, in which only a few patents differ across datasets in any given year, inventors' state of residence differs more frequently. Over all years in common, about 6% of the patent numbers in the CUSP report different inventors' states than the same patent number in the HistPat. In some years, more than 8% of patent numbers have different inventors' states between the CUSP and HistPat.

Given the unstructured nature of the patent text used by the CUSP and HistPat datasets described in Section 2 and the different approaches taken to parse out location information described in Section 3.5, it is not surprising that location data is more likely to differ across datasets. Nevertheless, 6% is a non-trivial fraction of patents, and it is worth understanding why these differences occur so often. I randomly sample 750 patents in which the inventors' state of residence in the CUSP disagrees with the state of residence in the HistPat, oversampling from the two periods with exceptionally high rates of disagreement, from 1845 to 1872 and 1950 to 1970. I then manually check the patent image files for these patents to see if the first inventor's true location matches the location reported in the CUSP, the HistPat, or neither. In 203 of the 750 cases (about 27%) the state in the CUSP was correct and the

²⁴In Appendix H, I compare the issue years in the HPDF to the CUSP, HistPat, and Jim Shaw data. While application dates in the HPDF are unreliable, the HPDF issue years are very similar to those in the CUSP. Appendix H also presents results of several other dimensions along which patents with the same patent number can have different information, including how often datasets record different numbers of inventors, different patent classes, or different inventor countries or counties of residence.

HistPat was in error. In the majority of cases (477 patents or about 64%), the state in the HistPat was correct and the CUSP was in error. Finally, for 71 patents (about 9%), the state was incorrect in both the CUSP and HistPat. Assuming that the state is correct in all cases for which the CUSP and HistPat report the same state, these results imply that patents in the CUSP have an incorrect state about 4.3% of the time, while the HistPat have an incorrect state about 2.5% of the time.²⁵ These aggregate percentages mask substantial change over time. For instance, among the patents that report different states between the CUSP and the HistPat, the share of patents for which the CUSP state is correct increases by about 0.15 percentage points per year; the CUSP is correct in these cases about 18% of the time from 1845 through 1875, but about a third of the time from 1950 through 1970.

What explains these errors? At least 15% of the mistaken locations in the CUSP data occur when an inventor resides in a town name that appears in multiple states; when this happens, recall from Section 3.5 that the CUSP places the patent in the state with the town that has the largest population. For example, the CUSP mis-records inventors residing in Riverside, CT, as residing in Riverside, CA, and inventors in Springfield, VA, as residing in Springfield, MA. In the HistPat data, the most common location errors come from reporting the assignee's location as the inventor's location; this accounts for more than 36% of the mistakes in the HistPat locations. For most errors, however, it is not clear why the state in the CUSP or HistPat differs from the location in the patent images.

One possibility is that discrepancies are caused by errors in recording patent numbers,

²⁵Conditional on the state disagreeing between the CUSP and the HistPat, the HistPat is more likely to be correct. However these disagreements make up a larger share of the HistPat dataset. Just under 6% of patents in the CUSP have different states than the HistPat, and in 73% of these, the CUSP is incorrect. 6.8% of the patents in the HistPat have different states than the CUSP, and in 36% of these the HistPat is incorrect.

rather than errors in recording inventors' states of residence. Patent numbers are often difficult to read, even for a human; it is especially difficult to distinguish between 3, 6, 8, and 9, and especially for patents in early years. To lend some credence to this, in Figure 12 I compare the distribution of reported states for the patents with incorrect states to the distribution of states for all patents. If incorrect states are the result of random OCR errors, then the distribution of incorrect states should be approximately uniform. If, instead, incorrect states are largely driven by errors in recording patent numbers, then the “wrong” state is the correct state for a different patent, and hence the distribution of incorrect states should be similar to the overall distribution of patents by state. For both the CUSP (panel (a)) and HistPat (panel (b)), the distributions of incorrect states are similar to those of all patents. Of course, the distributions are not identical; Alabama and Massachusetts are listed as incorrect states in the CUSP more often than expected based on the distribution of states over all patents, and California, New York, and Pennsylvania are listed as incorrect states less often than expected. Nevertheless, for both the CUSP and HistPat, Kolmogorov-Smirnov tests fail to reject the equality of the distribution of incorrect states to the distribution of states for all patents, while strongly rejecting the null of a uniform distribution.²⁶ If indeed many of these errors are caused by mistakes in patent numbers, then researchers should be advised against merging the CUSP and HistPat datasets on patent numbers to obtain larger sample sizes. Doing so may “double count” patents for which the patent number is mis-recorded in at least one of the datasets.

²⁶These results are available upon request. Results are similar if I limit the sample of incorrect states to those for which I cannot identify the cause of the error.

4.3 Comparing the SAZ Data to Other Datasets

While it is relatively straightforward to compare patent information across the datasets that contain patent numbers, the same cannot be said for comparing these datasets to the SAZ dataset. Recall that SAZ does not contain patent numbers; while patent numbers are recorded in the Annual Reports, they were difficult to parse and so are not included in the SAZ data. This makes it impossible to simply check whether the SAZ contains “the same patents” as, say, the CUSP or HistPat data.

Instead, I look for patents with a given characteristic in the SAZ data, manually record patent numbers for these patents, and then verify whether patents with the same characteristics in the CUSP or HistPat also have the same patent numbers. In particular, I find in the Annual Reports records for all SAZ patents with inventors residing in the states of Vermont and Wisconsin in 1900 and 1940. I choose these two states because in both years they account for a similar share of patents in the SAZ, CUSP, and HistPat datasets (see Appendix F) but had few enough patents that it was possible to manually check all of them. Using patent numbers from the Annual Reports, I see whether these same patents appear in the CUSP and HistPat datasets in the same year. I then count how many patent numbers are in the CUSP or HistPat datasets in these states and years that do not appear in these states and years in the SAZ data. This procedure is essentially the reverse of the procedure in Section 4.2. Above, I see if records with the same patent numbers have the same information; here, I check if records with the same information also have the same patent numbers.

Results are presented in Table 3. In the first row of panel (a), I compare the number

of Wisconsin and Vermont patents in 1940 and 1900 to the CUSP dataset.²⁷ There are 1,276 patents in the SAZ dataset in those states and years for which I can identify a patent number. For 1,083 of these 1,276 patents, or about 85%, I am able to locate the same patent number in the CUSP dataset; in 977 of these, or about 77%, the patent in the CUSP data lists the same inventor's state of residence as the SAZ patent. In 193 patents, or about 15% of the SAZ patents from these states and years, I am unable to locate the patent number in the CUSP dataset. Finally, in the CUSP data there are 441 patents in Wisconsin and Vermont for which patent numbers were not found in those states in the SAZ data; this is about 34% as many patents as are found in these states in the SAZ dataset. Row 2 of panel (a) lists results comparing the SAZ to the HistPat dataset. I find similar magnitudes.

For the patents that appear in both the SAZ dataset and either the CUSP or HistPat dataset but record a different state in different datasets (column 2 in panel (a)), it is possible to drill down to determine why these discrepancies occur. Different states could be the result of an OCR error when reading the Annual Reports to create the SAZ data, an OCR error when reading a patent image file to create the CUSP or HistPat data, or a typographical or clerical error that causes the Annual Reports and the individual patent image file to report different information. I manually inspect both the Annual Reports and each patent image file for these patents to see which of these stories is true.

In panel (b), I divide the 106 patents that disagree between the SAZ and CUSP and the 99 patents that disagree between the SAZ and HistPat into those three categories. In 33 of

²⁷In parentheses, also present the fraction of Annual Reports patents from Wisconsin and Vermont in 1940 and 1900. Because column 4 consists of patents not in the Annual Reports, the sum of the fractions is greater than one. There are 26 patents in Wisconsin and Vermont in 1940 and 1900 in the SAZ data for which I was unable to locate in the Annual Reports, and hence was unable to find a patent number for these patents. These 26 patents are not counted in the denominator in Table 3.

the 106 cases (31%) in which the SAZ and CUSP disagree, and 32 of the 99 cases (32%) in which the SAZ and HistPat disagree, the reason is due to an error in reading the inventor's state of residence from the Annual Report when creating the SAZ data. In the vast majority of these cases (27 out of 33 or 32, respectively, or about 82-84%), this error is caused by erroneously reporting the assignee's state as the inventor's state of residence. The remaining cases are due to mis-recognizing a character or some similar miscellaneous error.

In about 25% of the cases in which the SAZ and CUSP or HistPat disagree, the mistake appears to be in the CUSP or HistPat, respectively. Errors in these datasets occur for the same reasons identified in Section 4.2. For about 15% of the CUSP errors, the CUSP data locates the inventor in a city with same name and a larger population in a different state (e.g., erroneously reporting "Grand Rapids, WI," as "Grand Rapids, MI"). In 24% of the HistPat errors, the HistPat erroneously replaced an inventor residing in Wisconsin or Vermont with the location of either the assignee or another inventor.

The plurality of cases in which the SAZ and either the CUSP or HistPat data disagree is due to a disagreement between the state reported in the Annual Reports and the state reported in the individual patent image files (this is 42% of the disagreements between both the SAZ and CUSP and the SAZ and the HistPat). While there are likely some cases where the Annual Reports mis-record an inventor's state, I believe most of these come from difficulty in reading the patent number in the Annual Report: as difficult as it is for a human reader to decipher a patent number from scanned patent text, it is even harder to read from the Annual Reports, especially in 1900.

5 Other Types of Historical Patent Datasets

5.1 Design Patents

The data used in this paper are for U.S. utility patents, but the USPTO also grants other types of patents. One of these is the design patent, which protects the ornamental appearance of an article. Or, as the USPTO puts it, “a ‘utility patent’ protects the way and article is used and works...while a ‘design patent’ protects the way an article looks” (US Patent and Trademark Office, 2020a). Consequently, design patents are a poor measure of new useful inventions, but may be useful as proxies for new products or to measure changes in the rate of creation of ornamental designs. Design patents have been relatively understudied in the innovation, economics, and management literatures, even when using recent design patent data (Chan, Mihm, and Sosa (2018) is a rare exception), although they have been studied in the legal literature (e.g., Schwartz and Giroud (2020), Lee and Sunder (2013), Ferrill and Tanhehco (2011)).

The first design patent was issued in 1842, not long after the first utility patents studied in this paper.²⁸ The SAZ data includes design patents for the years 1870-1940, although the dataset was primarily constructed to record information on utility patents and so coverage of design patents varies widely from year to year depending on how design patents were recorded in the Annual Reports (see Appendix I). Jim Shaw has also transcribed a list of design patents granted between 1842 and 1873 (Shaw, n.d.-a).²⁹

²⁸See Quinn (2014) for an informative overview of the history of design patents.

²⁹The Jim Shaw design patent data is available for download at <https://ptrca.org/history>.

5.2 Plant Patents

Another type of patent granted by the USPTO is the plant patent. Following the passage of the Plant Protection Act of 1930, inventors could apply for a plant patent for novel varieties of plants that reproduce asexually.³⁰ For more detail on what plant patents cover, see US Patent and Trademark Office (2020b).

Relative to other types of patents, few plant patents are granted. In fact, prior to 1975, the USPTO has granted more than 200 plant patents in a single year only once (in 1974), and for every year before 1950 fewer than 100 were issued in any given year.³¹ But plant patents do provide researchers with a rare opportunity to observe the effects of introducing a novel form of intellectual property to cover a distinct type of technology. For instance, Moser and Rhode (2012) examine how the introduction of plant patents affected the rate of innovations in rose varieties. The SAZ data includes plant patents, although as in the case of design patents, coverage is limited and varies from year to year.

5.3 X-Patents

As noted above, I start the analysis in this paper with patents issued in 1836 or later. But U.S. patents did not begin in 1836. Patents were explicitly listed as a power of the federal government in the U.S. Constitution, written in 1787. In 1790, the U.S. passed its first patent law, under which the Attorney General and the Secretaries of War and State would examine patent applications. The examination process quickly became too burdensome for

³⁰Plants that reproduce sexually could be protected by utility patents beginning in the 1980s. For overviews of the evolution of patent law for plants with a focus on more recent changes, see Clancy and Moschini (2017) and Campi and Nuvolari (2015).

³¹Data on aggregate plant patenting is from https://www.uspto.gov/web/offices/ac/ido/oeip/taf/h_counts.htm.

these officials, especially Secretary of State Thomas Jefferson, and so in 1793 the U.S. passed a new patent law. From 1793 through most of 1836, the U.S. had a patent registration system rather than an examination system. This makes it difficult to trust the novelty of pre-1836 patents, which is one reason to begin the quantitative analysis of patents in 1836. A second reason is that an 1836 fire at the Patent Office destroyed many of the early patent records. While copies of some of these patents have been recovered, many have not, and the earliest patents were especially likely to be lost. For instance, only five of the 57 patents examined and granted by Thomas Jefferson survive (Risch, 2012, p. 1282). Of those patents that survived, many are handwritten, making them difficult to read and nearly impossible to digitize with OCR software. Moreover, these early patents did not have claims, making it difficult to determine what each patent covers.

While not all of these early patents survive, several of those that did are historically important. These include the first patent issued by the U.S. government, for a method of making potash; Eli Whitney's cotton gin; an early breech loading firearm; the first spring mattress; and several patents for asbestos fire proofing and lead paint. Risch (2012) provides an entertaining description of a number of these early patents. Thus these first patents may be of interest to scholars of technology, as may the study of nascent patent systems.

Although the text of every pre-1836 patent is not available, several sources have compiled lists of all known issued patents; see for instance Ellsworth (1840), Burke (1847), and MacMurray (1985). Collectively, the pre-1836 patents are known as the "X-patents," or sometimes as "name-and-date patents" because they were not initially issued with a patent number (making it difficult to know if these lists of issued patents are complete) but instead only listed inventors' names and date of issuance. After the 1836 fire, the Patent Office

retrospectively numbered these patents with an “X” followed by a number determined by the patents’ chronological order, hence their intriguing name.

“X-patents” data are currently available as part of the HistPat dataset; they are easily identifiable by the patent number. Jim Shaw has also posted a list of known “X-patents” (Shaw, n.d.-c), likely transcribed from one of the lists mentioned above.³² Several authors have studied the “X-patents”, including Sokoloff (1988), Risch (2012), and dos Santos Nascimento, dos Santos, Almedia Alves, and Ferreira Nascimento (2018).

5.4 Confederate Patents

One challenge with examining the time series of U.S. invention is that thirteen states—Alabama, Arkansas, Florida, Georgia, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee, Texas, and Virginia, along with parts of the territories that became Arizona, New Mexico, and Oklahoma—seceded from the United States and formed the Confederate States of America (CSA) from 1861 to 1865. These southern states accounted for relatively few patents until after World War II, so this issue is unlikely to have a major effect on aggregate patenting counts, but it does make interpreting changes in regional patenting difficult.³³

Promoting invention was seen as vital for the success of the South’s war effort, and so the CSA established a patent office in 1861. Knight (2011) provides a detailed analysis of the Confederate States Patent Office (CSPO), and lists of all Confederate patents are available from US Patent and Trademark Office (n.d.).³⁴ In Figure 13, I plot the number of patents

³²The Jim Shaw “X-patent” data is available for download at <https://ptrca.org/history>.

³³For this reason, some analyses such as Sarada et al. (2019) and Andrews and Whalley (2020) begin after the Civil War.

³⁴I thank H. Jackson Knight for pointing me to this data.

in the states and territories that made up the CSA from 1855 to 1870.³⁵ The CSPO issued a relatively small number of patents, only 266 during its entire existence, but ignoring those patents understates the amount of patented invention in the CSA states by more than half during the Civil War years. Knight (2011) provides a wealth of details about what those patents contained as well as the identities of their inventors.

6 What Historical Patents Don't Have

If one wants counts of patent numbers or inventor and assignee names and locations, then the historical patents are of comparable completeness to data on recent patents. But for other features of patents, data is not available for many historical periods. I describe several of these features here.

First, patent applications were not published prior to a change in the patent law in 2000 (US Patent and Trademark Office, 2000). While the USPTO's aggregate patenting data lists the aggregate number of patent applications received in each year going back to 1840, there is no way to disaggregate these data geographically or to examine individual applications. The reason historical patent applications have not been published is that inventors may have abandoned patent applications and decided to protect their inventions via trade secrecy instead; publishing the historical applications would disclose these inventions and make the trade secrets worthless.³⁶

³⁵In addition to patents by inventors who were residents of the CSA, the CSPO also issued several patents to inventors residing in states that did not officially succeed, namely Kentucky and Missouri, as well as to an inventor residing in Bavaria.

³⁶I appreciate conversations with several current and former members of the USPTO Office of the Chief Economist for clarifying the rationale for this policy decision. The HPDF dataset contains information on the flows of publicly available utility patent applications (that is, those that are referenced by another patent or have been voluntarily published by their owners) dating to 1981.

Second, historical patent data lacks information related to the patent prosecution process. More specifically, to the best of my knowledge there is no information on the patent attorney or agent and, prior to 1965, no information on the patent examiner associated with each patent for most years; there is no equivalent to the USPTO's Patent Application Information Retrieval (PAIR) system for historical patents.³⁷ This makes analysis that exploits assignment of patents to examiners, as in B. Sampat and Williams (2019), Gaulé (2018), Righi and Simcoe (2019), or Feng and Jaravel (2020), infeasible with historical patent data.

Third, patent quality is difficult to measure in traditional ways for historical patents. As mentioned in Section 3.9 above, patent citations are not recorded on the face of a patent until 1947. Another approach to measuring patent quality is to use the construction of patent claims. More specifically, Kuhn and Thompson (2019) and Marco, Sarnoff, and deGrazia (2019) argue that the length of a patent's first claim is a proxy for patent breadth, with longer claims indicating a narrower and more circumscribed patent. Enrico Berkes has constructed measures of the length of each patent's first claim, available upon request along with the CUSP data. But claim length is unlikely to be informative for many historical patents. Prior to the late nineteenth century, most U.S. patents were written with "central claiming," in which the invention description defines the scope of the invention and claims are either not precisely written or not present at all; see, e.g., Fromer (2009) and Sawicki (2018). While a change in the patent law in 1870 required peripheral claiming, for several decades many inventors continued to write claims that were vague and little more informative than central claims. Until the early twentieth century, claims that read "I claim the invention as described above," or some similar variation, are common (Anderson & Menell, 2015). There

³⁷See <https://portal.uspto.gov/pair/PublicPair>.

is very little variation in identifiable claim structure for these patents, and so claim length or related measures do not serve as good proxies for patent scope or quality. Perhaps the most promising approach for constructing measures of patent quality that are available for both historical and modern patents are text-based measures; patent descriptions have always been a part of every published patent, and while image quality and changing language may make textual comparisons over long periods of time difficult, these concerns are not insurmountable (Kelly et al., 2020a, 2020b).

7 Conclusion

What is the best historical patent dataset for a researcher to use? As the analysis above clearly shows, the answer to this question can only be: “it depends.” The best dataset depends on the what features of patents are most salient to the researcher as well as on how a researcher weights completeness and accuracy. In some cases, the decision of which dataset to use is obvious. If one wants to study patenting by publicly traded firms or wants a measure of patent values based on the movement of stock prices, then the KPSS dataset is the only choice. If one wants patent data linked to the U.S. census to track inventor demographics, the SAZ dataset should be used. For in-text patent citations, the CUSP is the answer. Often the time period of study answers the question as well: the Jim Shaw data cannot be used for patents issued after 1873, the KPSS data cannot be used for patents issued before 1926, and the SAZ data cannot be used for patents issued before 1870 or after 1942. But in other contexts, researchers face tradeoffs. For instance, the CUSP contains a larger number of patents than does the HistPat, although it has a slightly higher rate of

patents with incorrect location information (4.3% in the CUSP versus 2.5% in the HistPat). My overall impression is that, for most questions, the CUSP is currently the gold standard both in terms of completeness and scope of the types of patent information it contains, with any differences in error rates being immaterial in most contexts. But other researchers may draw different conclusions based on their specific situations.

Perhaps the strongest message of this paper is that for many questions, it does not matter which patent dataset is used; all will give similar answers. For most years and most pieces of patent information, these six datasets paint a remarkably consistent picture of the history of U.S. invention. It is worth pointing out the few cases where the datasets do not agree, especially when it is possible to determine why discrepancies occur and which dataset is more likely to contain inaccurate information. I do not recommend using the patent application dates reported in the HPDF data. The HistPat dataset does not appear to be a good source for information on collaborations between multiple inventors. Finally, the SAZ dataset, as well as the HistPat dataset for patents issued before 1900, appear to miss substantial portions of patent assignments. But in most cases, researchers have several acceptable options.

How should researchers handle cases in which multiple datasets could plausibly do the job? For instance, both the CUSP and HistPat are suitable in many situations that require wide coverage and information about patent locations. In these cases, one might be tempted to combine datasets to ensure maximal coverage. One could simply merge the CUSP and HistPat datasets by patent number, getting “the best of both worlds.” While at first blush this may sound like an effective strategy to minimize measurement error and missing data, I recommend against this approach. When a patent number appears in one dataset but not another, this often appears to be due to an error in recording that patent number. Thus, just

because a patent number is in the HistPat but not the CUSP does not mean that that patent is omitted from the CUSP, and vice versa. By merging two datasets together, researchers may therefore double count some patents.

Instead, I recommend simply running all analyses using multiple patent datasets. This should be especially feasible given that most of these datasets are easy to locate online, download, and use (especially after reading this paper). If the results are robust to using multiple datasets, this gives confidence that errors in constructing the patent datasets are immaterial in a researcher's present context. If the results are not robust, I encourage users to figure out when and where the multiple datasets differ; doing so will aid the entire research community in improving the quality and completeness of these data sources.

There are some features of patent data that are simply not available for historical patents, namely those related to patent applications and to the prosecution and examination process. But at the same time, history provides unique opportunities to examine nascent patent systems (e.g., from the Confederate States Patent Office) or patents issued under alternative institutional regimes (e.g., the patent registration system in place during the period of the X-patents). These are settings that cannot be investigated using modern patent data.

Finally, it is important to note that the results in this paper represent the quality and completeness of these historical patent datasets at a particular point in time. As work continues to improve these datasets, future researchers will hopefully have even higher quality data to work with. By documenting differences across the datasets and identifying some causes of these discrepancies, my hope is that this paper will contribute to that goal.

References

- Acemoglu, D., Moscona, J., & Robinson, J. A. (2016, May). State capacity and American technology: evidence from the nineteenth century. *American Economic Review: Papers & Proceedings*, 106(5), 61-67.
- Akcigit, U., Grigsby, J., & Nicholas, T. (2017a, May). Immigration and the rise of American ingenuity. *American Economic Review: Papers & Proceedings*, 107(5), 327-331.
- Akcigit, U., Grigsby, J., & Nicholas, T. (2017b). *The rise of American ingenuity: innovation and inventors of the golden age*. (Unpublished, University of Chicago)
- Anderson, J. J., & Menell, P. S. (2015, January). Informal deference: a historical, empirical, and normative analysis of patent claim construction. *Northwestern University Law Review*, 108(1), 1-84.
- Andrews, M. J. (2020a). *Bar talk: informal social interactions, alcohol prohibition, and invention*. (Unpublished, National Bureau of Economic Research)
- Andrews, M. J. (2020b). *How do institutions of higher education affect local invention? Evidence from the establishment of U.S. colleges*. (Unpublished, NBER)
- Andrews, M. J., & Whalley, A. (2020). *150 years of the geography of innovation*. (Unpublished, NBER)
- Argente, D., Baslandze, S., Hanley, D., & Moreira, S. (2019). *Patents to products: innovation, product creation, and firm growth*. (Unpublished, Penn State)
- Babina, T., Bernstein, A., & Mezzanotti, F. (2020). *Crisis innovation*. (Unpublished, Columbia University)
- Balsmeier, B., Assaf, M., Chesebro, T., Fierro, G., Johnson, K., Johnson, S., ... Fleming, L. (2018, Fall). Machine learning and natural language processing on the patent corpus: data, tools, and new measures. *Journal of Economics and Management Strategy*, 27(3), 535-553.
- Baron, J., & Pohlmann, T. (2018, Fall). Mapping standards to patents using declarations of standard-essential patents. *Journal of Economics and Management Strategy*, 27(3), 504-534.
- Bell, A., Chetty, R., Jaravel, X., Petkova, N., & Reenen, J. V. (2019, May). Who becomes an inventor in America? The importance of exposure to innovation. *Quarterly Journal of Economics*, 134(2), 647-713.
- Berkes, E. (2018). *Comprehensive universe of U.S. patents (CUSP): data and facts*. (Unpublished, Ohio State University)
- Berkes, E., & Nencka, P. (2019). *Knowledge access: the effects of Carnegie Libraries on innovation*. (Unpublished, Ohio State University)
- Bessen, J. (2008). *Name matching tool*. (<https://sites.google.com/site/patentdatapoint/Name-matching-tool>, accessed August 3, 2020)
- Burke, E. (1847). *List of patents for inventions and designs, issued by the United States, from 1870 to 1847, with the patent laws and notes of decisions of the courts of the United States for the same period: compiled and published under the direction of Edmund Burke, commissioner of patents*. Washington, DC: J. & G. S. Gideon.
- Campi, M., & Nuvolari, A. (2015, May). Intellectual property protection in plant varieties: a worldwide index (1961-2011). *Research Policy*(4), 951-964.
- Chan, T. H., Mihm, J., & Sosa, M. E. (2018, March). On styles in product design: an

- analysis of U.S. design patents. *Management Science*, 64(3), 983-1476.
- Chandler, A. D. (1990). *Scale and scope: the dynamics of industrial capitalism*. Cambridge, MA: Belknap Press.
- Clancy, M. S., & Moschini, G. (2017, October). Intellectual property rights and the ascent of proprietary innovation in agriculture. *Annual Review of Resource Economics*, 9, 53-74.
- dos Santos Nascimento, H. H., dos Santos, C. N., Almedia Alves, C. H., & Ferreira Nascimento, M. L. (2018, June). The X-patents. *World Patent Information*, 53, 1-13.
- Ellison, G., & Glaeser, E. L. (1997, October). Geographic concentration in U.S. manufacturing industries: a dartboard approach. *Journal of Political Economy*, 105(5), 889-927.
- Ellsworth, H. L. (1840). *A digest of patents, issued by the United States, from 1790 to January 1, 1839: published by Act of Congress under the superintendence of the Commissioner of Patents, Henry L. Ellsworth, to which is added the present law relating to patents*. Washington, DC: Peter Force.
- Feng, J., & Jaravel, X. (2020, January). Crafting intellectual property rights: implications for patent assertion entities, litigation, and innovation. *American Economic Journal: Applied Economics*, 12(1), 140-181.
- Ferrill, E., & Tanhehco, T. (2011, Spring). Protecting the material world: the role of design patents in the fashion industry. *North Carolina Journal of Law & Technology*, 12(2), 251-300.
- Frakes, M. D., & Wasserman, M. F. (2017, July). Is the time allocated to review patent applications inducing examiners to grant invalid patents? Evidence from microlevel application data. *Review of Economics and Statistics*, 99(3), 550-563.
- Fromer, J. C. (2009, Spring). Claiming intellectual property. *University of Chicago Law Review*, 76(2), 719-796.
- Furman, J. L., & MacGarvie, M. (2009, October). Academic collaboration and organizational innovation: the development of research capabilities in the US pharmaceutical industry, 1927-1946. *Industrial and Corporate Change*, 18(5), 901-928.
- Furman, J. L., & MacGarvie, M. J. (2007, August). Academic science and the birth of industrial research laboratories in the U.S. pharmaceutical industry. *Journal of Economic Behavior & Organization*, 63(4), 756-776.
- Gans, J. S., Hsu, D. H., & Stern, S. (2008, May). The impact of uncertain intellectual property rights on the market for ideas: evidence from patent grant delays. *Management Science*, 54(5), 982-997.
- Gaulé, P. (2018, June). Patents and the success of venture-capital backed startups: using examiner assignment to estimate causal effects. *Journal of Industrial Economics*, 66(2), 350-376.
- Graham, S. J. H., Grim, C., Islam, T., Marco, A. C., & Miranda, J. (2018, Fall). Business dynamics of innovation firms: linking U.S. patents with administrative data on workers and firms. *Journal of Economics and Management Strategy*, 27(3), 372-402.
- Graham, S. J. H., Marco, A. C., & Miller, R. (2018, Fall). The USPTO Patent Examination Research Dataset: a window on patent processing. *Journal of Economics and Management Strategy*, 27(3), 554-578.
- Graham, S. J. H., Marco, A. C., & Myers, A. F. (2018, Fall). Patent transactions in

- the marketplace: lessons from the USPTO Patent Assignment Dataset. *Journal of Economics and Management Strategy*, 27(3), 343-371.
- Griliches, Z. (1990, December). Patent statistics as economic indicators: a survey. *Journal of Economic Literature*, 28(4), 1661-1707.
- Hall, B. H., Jaffe, A., & Trajtenberg, M. (2005, Spring). Market value and patent citations. *RAND Journal of Economics*, 36(1), 16-38.
- Hall, B. H., Jaffe, A. B., & Trajtenberg, M. (2001). *The NBER patent citation data file: lessons, insights and methodological tools*. (NBER Working Paper)
- Jaffe, A. B., & de Rassenfosse, G. (2019). Patent citation data in social science research: overview and best practices. In P. Menell & D. Schwartz (Eds.), *Research handbook on the economics of intellectual property law: Volume 2: Analytical methods*. Northampton, MA: Edward Elgar Publishing.
- Jones, B. F. (2009, January). The burden of knowledge and the “death of the renaissance man”: is innovation getting harder? *Review of Economic Studies*, 76(1), 283-317.
- Kelly, B., Papanikolaou, D., Seru, A., & Taddy, M. (2020a). *Measuring technological innovation over the long run* (Vol. Forthcoming).
- Kelly, B., Papanikolaou, D., Seru, A., & Taddy, M. (2020b). *Technological change and occupations over the long run* (Vol. Forthcoming).
- Knight, H. J. (2011). *Confederate invention: the story of the Confederate States Patent Office and its inventors*. Baton Rouge, LA: LSU Press.
- Kogan, L., Papanikolaou, D., Seru, A., & Stoffman, N. (2017a). *Online appendix to “technological innovation, resource allocation, and growth”*. (Online appendix)
- Kogan, L., Papanikolaou, D., Seru, A., & Stoffman, N. (2017b, May). Technological innovation, resource allocation, and growth. *Quarterly Journal of Economics*, 131(2), 665-712.
- Kuhn, J. M., & Thompson, N. C. (2019). How to measure and draw causal inferences with patent scope. *International Journal of the Economics of Business*, 26(1), 5-38.
- Kuhn, J. M., Younge, K. A., & Marco, A. C. (2019). Patent citations reexamined. *RAND Journal of Economics*, Forthcoming.
- Lamoreaux, N. R., & Sokoloff, K. L. (2005). *The decline of the independent inventor: a Schumpeterian story*. (NBER Working Paper 11654)
- Lampe, R., & Moser, P. (2010, December). Do patent pools encourage innovation? Evidence from the 19th-century sewing machine industry. *Journal of Economic History*, 70(4), 898-920.
- Lampe, R., & Moser, P. (2012). Patent pools: licensing strategies in the absence of regulation. *Advances in Strategic Management*, 29, 69-86.
- Lampe, R., & Moser, P. (2013, Winter). Patent pools and innovation in substitute technologies: evidence from the U.S. sewing machine industry. *RAND Journal of Economics*, 44(4), 757-778.
- Lee, P., & Sunder, M. (2013, Fall). Design patents: law without design. *Stanford Technology Law Review*, 17, 277-303.
- Leggett, M. D. (1874). *Subject-matter index of patents for invention issued by the United States Patent Office from 1790 to 1873, inclusive* (Vol. 1-3). Washington, DC: Government Printing Office.
- Li, G.-C., Lai, R., D’Amour, A., Doolin, D. M., Sun, Y., Torvik, V. I., ... Fleming, L.

- (2014, July). Disambiguation and co-authorship networks of the U.S. Patent Inventor Database (1975-2010). *Research Policy*, 43(6), 941-955.
- MacMurray, R. P. (1985, June). Technological change in a society in transition: work in progress on a unified reference work in early American patent history. *Journal of Economic History*, 45(2), 299-305.
- Manson, S., Schroeder, J., Riper, D. V., & Ruggles, S. (2019). *IPUMS national historical geographic information system: version 13.0*. Minneapolis, MN: University of Minnesota. (<http://doi.org/10.18128/D050.V14.0>)
- Marco, A. C., Carley, M., Jackson, S., & Myers, A. F. (2015, June). *The USPTO historical patent data files: two centuries of invention*. (Unpublished, USPTO Economic Working Paper No. 2015-1)
- Marco, A. C., Sarnoff, J. D., & deGrazia, C. A. W. (2019, November). Patent claims and patent scope. *Research Policy*, 48(9).
- Marx, M., & Fuegi, A. (in press). Reliance on science in patenting: USPTO front-page citations to scientific articles. *Strategic Management Journal*.
- Monath, N., & McCallum, A. (2015). *Discriminative hierarchical coreference for inventor disambiguation*. (PatentsView Inventor Disambiguation Technical Workshop)
- Moretti, E. (2012). *The new geography of jobs*. New York: Houghton Mifflin Harcourt Publishing Company.
- Moser, P. (2005, September). How do patent laws influence innovation? Evidence from nineteenth-century world's fairs. *American Economic Review*, 95(4), 1214-1236.
- Moser, P. (2011, June). Do patents weaken the localization of innovations? Evidence from World's fairs. *Journal of Economic History*, 71(2), 363-382.
- Moser, P., & Rhode, P. W. (2012). Did plant patents create the American rose? In J. Lerner & S. Stern (Eds.), *The rate and direction of inventive activity revisited* (p. 413-438). Chicago: University of Chicago Press.
- Mowery, D. C. (1990, May). The development of industrial research in U.S. manufacturing. *American Economic Review*, 80(2), 345-349.
- Mowery, D. C., Nelson, R. R., Sampat, B. N., & Ziedonis, A. A. (2001, January). The growth of patenting and licensing by US universities: an assessment of the effects of the Bayh-Dole Act of 1980. *Research Policy*, 30(1), 91-119.
- Mowery, D. C., & Sampat, B. N. (2001, August). University patents and patent policy debates in the USA, 1925-1980. *Industrial and Corporate Change*, 10(3), 781-814.
- Nicholas, T. (2008, September). Does innovation cause stock market runups? Evidence from the Great Crash. *American Economic Review*, 98(4), 1370-1396.
- Nicholas, T. (2010, March). The role of independent invention in U.S. technological development, 1880-1930. *Journal of Economic History*, 70(1), 57-82.
- Packalen, M., & Bhattacharya, J. (2015a). *Cities and ideas*. (NBER Working Paper)
- Packalen, M., & Bhattacharya, J. (2015b). *New ideas in invention*. (NBER Working Paper)
- Pakes, A. (1986, July). Patents as options: some estimates of the value of holding European patent stocks. *Econometrica*, 54(4), 755-784.
- Perlman, E. R. (2015). *Dense enough to be brilliant: patents, urbanization, and transportation in nineteenth century America*. (Unpublished, Boston University)
- Petralia, S. (2019). *Mapping the technology frontier*. (Unpublished, LSE)
- Petralia, S., Balland, P.-A., & Rigby, D. L. (2016, August). Data descriptor: unveiling the

- geography of historical patents in the united states from 1836-1975. *Nature: Scientific Data*, 3, 1-14. (Article number: 160074)
- Quinn, G. (2014). *A brief history of design patents*. IPWatchdog. (<https://www.ipwatchdog.com/2014/01/11/design-patent-history/id=47283/>, accessed July 3, 2020)
- Righi, C., & Simcoe, T. (2019, February). Patent examiner specialization. *Research Policy*, 48(1), 137-148.
- Risch, M. (2012, September). America's first patents. *Florida Law Review*, 64(5), 1279-1336.
- Rosenberg, N., & Nelson, R. R. (1994, May). American universities and technical advance in industry. *Research Policy*, 23(3), 323-348.
- Sampat, B., & Williams, H. L. (2019, January). How do patents affect follow-on innovation? Evidence from the human genome. *American Economic Review*, 109(1), 203-236.
- Sampat, B. N. (2006, July). Patenting and US academic research in the 20th century: the world before and after Bayh-Dole. *Research Policy*, 35(6), 772-789.
- Sarada, Andrews, M. J., & Ziebarth, N. L. (2019, October). Changes in the demographics of American inventors, 1870-1940. *Explorations in Economic History*, 74.
- Sawicki, A. (2018, March). The central claiming renaissance. *Cornell Law Review*, 103(3), 645-722.
- Schmookler, J. (1966). *Invention and economic growth*. Cambridge, MA: Harvard University Press.
- Schwartz, D. L., & Giroud, X. (2020). The secret world of design patents. *Alabama Law Review*, *Forthcoming*.
- Shaw, J. (n.d.-a). *Design patents to 1873*. Hutchinson, KS: Patent and Trademark Resource Center Association. (<http://www.ptrca.org/history>, accessed July 3, 2020)
- Shaw, J. (n.d.-b). *Utility patents* (Vols. 3 vols: 1-65000, 65000-100000, 100001-146119). Hutchinson, KS: Patent and Trademark Resource Center Association. (<http://www.ptrca.org/history>, accessed August 5, 2020)
- Shaw, J. (n.d.-c). *X-patents (with added fields)*. Hutchinson, KS: Patent and Trademark Resource Center Association. (<http://www.ptrca.org/history>, accessed July 1, 2019)
- Sokoloff, K. L. (1988, December). Inventive activity in early industrial America: evidence from patent records, 1790-1846. *Journal of Economic History*, 48(4), 813-850.
- Sunder, J., Sunder, S. V., & Zhang, J. (2017, January). Pilot CEOs and corporate innovation. *Journal of Financial Economics*, 123(1), 209-224.
- The Commissioner of Patents. (Various Years). *Annual report of the Commissioner of Patents for the year ...* Government Printing Office.
- Thompson, P. (2006). Patent citations and the geography of knowledge spillovers: evidence from inventor- and examiner-added citations. *Review of Economics and Statistics*.
- Tian, X., & Wang, T. Y. (2014, January). Tolerance for failure and corporate innovation. *Review of Financial Studies*, 27(1), 211-255.
- Trajtenberg, M. (1990, Spring). A penny for your quotes: patent citations and the value of innovations. *RAND Journal of Economics*, 21(1), 172-187.
- Trajtenberg, M., Shiff, G., & Melamed, R. (2006). *The "names game": harnessing inventors' patent data for economic research*. (NBER Working Paper)
- US Patent and Trademark Office. (n.d.). Appendix: List of all known patents issued by

- confederate patent office. *History of the United States Patent Office: The Patent Office Pony: A history of the Early Patent Office*. (<http://www.myoutbox.net/popchapx.htm>, accessed October 21, 2016)
- US Patent and Trademark Office. (2000). *USPTO wil begin publishing patent applicationsl*. (<https://www.uspto.gov/about-us/news-updates/uspto-will-begin-publishing-patent-applications>, accessed February 28, 2017)
- US Patent and Trademark Office. (2013). *Classification orders archival report (classifications orders 1 through 1919)* (Archival Report). Washington, DC: U.S. Patent and Trademark Office.
- US Patent and Trademark Office. (2020a). *1502 definition of a design [R-07.2015]*. Washington, DC: U.S. Patent and Trademark Office. ([https://www.uspto.gov/web/offices/pac/mpep/s1502.html#:~:text=In%20general%20terms%2C%20a%20%E2%80%9Cutility,171\).&text=Both%20design%20and%20utility%20patents,its%20utility%20and%20ornamental%20appearance.](https://www.uspto.gov/web/offices/pac/mpep/s1502.html#:~:text=In%20general%20terms%2C%20a%20%E2%80%9Cutility,171).&text=Both%20design%20and%20utility%20patents,its%20utility%20and%20ornamental%20appearance.), accessed July 3, 2020)
- US Patent and Trademark Office. (2020b). *1601 introduction: the act, scope, type of plants covered [R-11.2013]*. Washington, DC: U.S. Patent and Trademark Office. (<https://www.uspto.gov/web/offices/pac/mpep/s1601.html>, accessed July 3, 2020)
- Wasserman, M. F., & Frakes, M. D. (2019). Empirical scholarship on the prosecution process at the PTO. In P. Menell & D. Schwartz (Eds.), *Research handbook on the economics of intellectual property law: Volume 2: Analytical methods*. Northampton, MA: Edward Elgar Publishing.
- Wuchty, S., Jones, B. F., & Uzzi, B. (2007, 18 May). The increasing dominance of teams in production of knowledges. *Science*, *316*(5827), 1036-1039.

Tables

Table 1: Comparison of Features in Each Historical Patent Dataset

	CUSP	HistPat	SAZ	Jim Shaw	KPSS	USPTO HPDF
Years Covered	1836-2014	1790-1976	1870-1942	1836-1873	1926-2019	1790-2014
Issue Date	x	x	x	x	x	x
Patent Number	x	x		x	x	x
Inventor First Name	x		x			
Inventor Last Name	x		x	x		
Names of Multiple Inventors	x		x	x		
Assignee Name	x		x		x	
Inventor Town	x		x	x		
Inventor County	x	x	x			
Inventor State	x	x	x	x		
Assignee Location	x	x	x			
Application Date	x					x
Patent Class	x					x
Invention Name			x	x		
Citations	x				x	
Private Patent Values					x	

Notes: An “x” indicates that a given feature appears in the corresponding patent dataset. See the text for a detailed description of each feature in each dataset and some caveats.

Table 2: Number (and Fraction) of Patent Numbers in One Dataset That Are Not in Another

Patent #s In	CUSP	Patent #s Not In HistPat	Patent #s Not In Jim Shaw	Patent #s Not In HPDF
CUSP		606,124 (0.1545)	1 (0.0000)	1 (0.0000)
HistPat	39 (0.0000)		0 (0.0000)	10 (0.0000)
Jim Shaw	326 (0.0022)	11,954 (0.0819)		303 (0.0021)
HPDF	42 (0.0000)	605,795 (0.1515)	0 (0.0000)	

Notes: Number and fraction of patent numbers that appear in the dataset listed in the row, but do not appear in the dataset listed in the column. The analysis is for all patent numbers included in any of the CUSP, HistPat, Jim Shaw, and HPDF datasets for years that they datasets have in common. For each row, the number of patent numbers are listed on top and the fraction of the row patent numbers are listed on the bottom in parentheses.

Table 3: Comparing Patents in the SAZ Dataset to the CUSP and HistPat Datasets

(a) Number (and Fraction) of Patent Numbers in the SAZ and/or Another Dataset

	Pat # In Both		Pat # In SAZ,	Pat # In Other,
	Same State	Different State	Not Other	Not SAZ
CUSP	977 (0.77)	106 (0.08)	193 (0.15)	441 (0.35)
HistPat	963 (0.75)	99 (0.08)	214 (0.17)	441 (0.35)

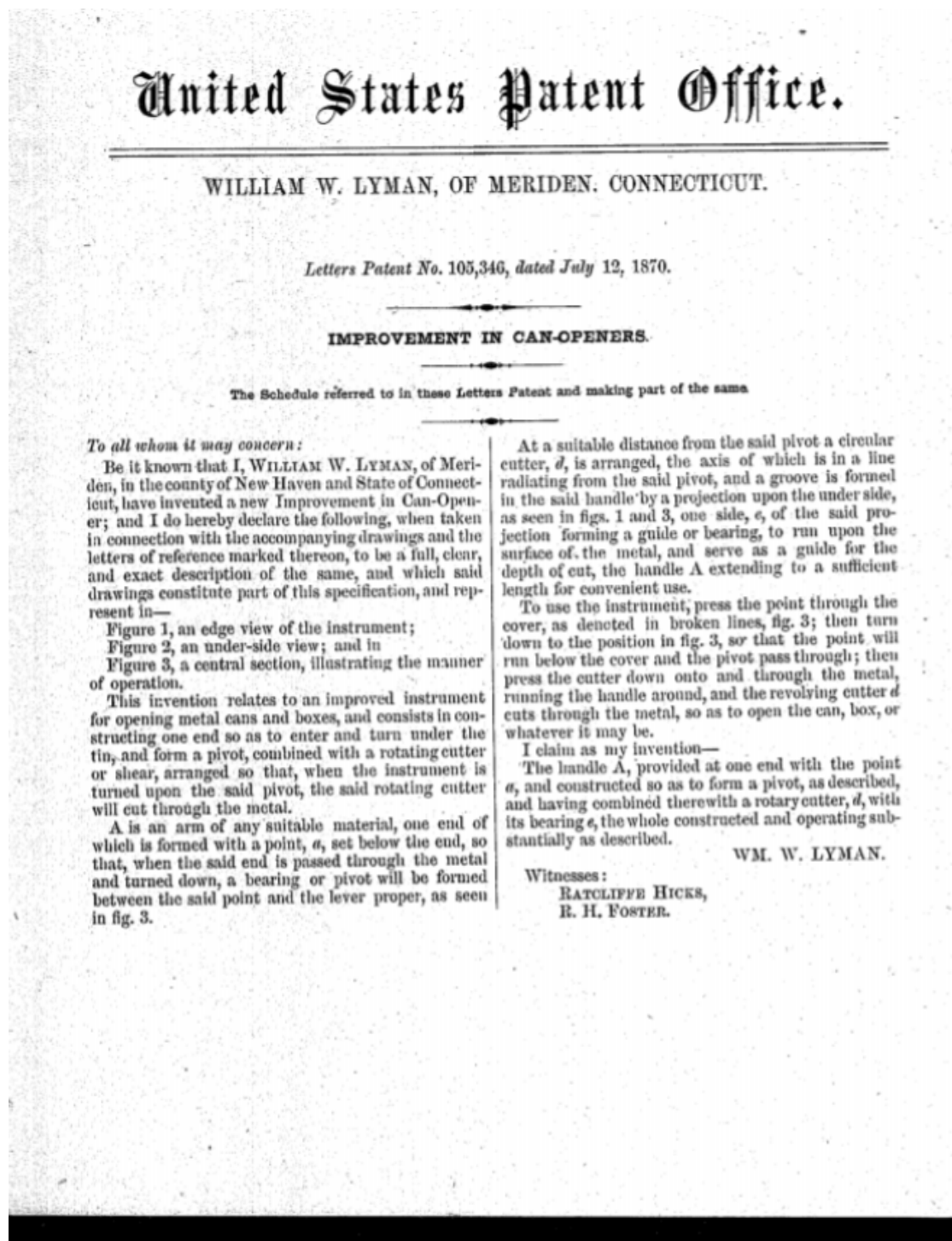
(b) Why Do Patent Numbers Record Different States between SAZ and Another Dataset?

	SAZ Different than Annual Report	Annual Report Different than Patent PDF	Other Different than Patent PDF
CUSP	33 (0.31)	28 (0.26)	45 (0.42)
HistPat	32 (0.32)	25 (0.25)	42 (0.42)

Notes: Comparison of the number and fraction of patent numbers from inventors residing in Wisconsin and Vermont in 1940 and 1900 in SAZ and either the CUSP or HistPat datasets. In Panel (a), Column 1 reports the number of patent numbers that are in both the SAZ and another dataset and have the same inventor state recorded in both datasets. Column 2 reports the number of patent numbers that are in both the SAZ and another dataset and have a different inventor state recorded in each dataset. Column 3 reports the number of patent numbers that are in the SAZ but not in another dataset. Column 4 reports the number of patent numbers that are in another dataset but are not in the SAZ. Panel (b) drills down into the patents from panel (a) that are in both datasets but report a different state. Column 1 of Panel (b) reports the number of patents with different states for which the SAZ data records a different state than does the corresponding record in the Annual Reports. Column 2 reports the number of patents with different states for which the other dataset records a different state than does the corresponding record in the patent PDF image file. Column 3 reports the number of patents with different states for which the Annual Reports records a different state than does the patent image PDF. In both panels, the first row compares SAZ to the CUSP and the second row compares SAZ to HistPat. For each row, the number of patents are listed on top and the fraction of the relevant SAZ patents are listed on the bottom in parentheses.

Graphs

Figure 1: Example of a Historical Patent



Notes: US patent 105,346 is for the first rotating wheel can opener, invented in 1870 by William Lyman.

Figure 2: Example of a Google Patent Record

Improvement in can-openers
US 105346 A

ABSTRACT [available in](#)

Publication number	US 105346 A
Publication type	Grant
Publication date	Jul 12, 1870
Export Citation	BIBTeX, EndNote, RefMan
Classifications (1)	
External Links:	USPTO, USPTO Assignment, Espacenet

IMAGES (1)



DESCRIPTION (OCR text may contain errors)

CLAIMS [available in](#)

@aient @Mire wmiIAM w; LYMAN, or Mnifnlnlv.' ooNNEoTloUf.

" Letters Patent No. 105,346, dated July 12, 1870.

p IMPROVEMENT -IN ymittelElmas.4

The Schedule ifereeo'i to' in these Letters Patent andmaking part of the sama To (all whom tt 'may conce/m: h p 'Be it known that I, WILLIAM W. LYMAN, of Meriden, in the coun ty of New Haven and State of'Gonnecticut, have invented a new Improvement in Gan-Opener; and I do herebydeclare the following, when taken in 'connection with the accompanying drawings and the letters of reference marked thereon,"to be a full, clear, and exact descriptionof the saine, and which said drawings constitute part ofths specification, und represent in--Figure 1, anedge view ofthe instrument;

Figure 2, an under-side view; and in Figure 3, a cent-ral section, ill sti-ating the manner of operation. f I, This invention 'relates toan improved instrument i for opening metal cans and boxes, and consists in constructing one end so as to enter and turn under the tin andl form a pivot, combined with a rotating cutter or shear, arranged 'so that, when the instrument is- -tuned lupon thexsaid pivot, the said rotating cutter will cut through the metal.

As an arm of anysuitable material, one end of which is formed with. a point, u, set below the end, so that,4 when thesaid end is passed through the metal and turned down, a bearing or vpivot will be formed betgveen the said point andthe lever proper, as seen in e. 3.

At a'suitable distance from the said pivot a circular cutter, d, is arranged, the axis of which is in a line radiating from lthe said pivot, and a groove is formed in the said lian'dleby a projection upon the under side, as seen in figs. 1 and 3, one side, c, of 4the said projection forming a guide or bearing, to run upon the surface otthe meta-l, and serve as a guide for the 'depth of out, the handle A extending to a suticientlength for convenient use.

'to use thel instrument, press the point through the cover, as denoted in broken lines, tig. 3; then turn 'down tothe position in tig. 3, so that the pointwill run belowthe cover and the pivotpass through; then press the cutter down onto andthrough the metal, ruiming the handle around, and the revolving cutter d cuts through the metal, so as to open the can, box, or whatever it may be.

Notes: A screenshot of the Google Patent record for William Lyman's can opener patent, <https://patents.google.com/patent/US105346A/en?q=US105346>.

Figure 3: A Page of the 1870 Annual Report of the Commissioner of Patents

<i>Alphabetical list of patentees for the year 1870—Continued.</i>		
No.	Name, residence, and invention or discovery.	Date.
103,274	Luppen, Luppe, Pekin, Ill. Shovel-plow.....	Oct. 11, 1870
103,275	Same..... same.....	Oct. 11, 1870
103,276	Same..... same.....	Oct. 11, 1870
110,054	Lupton, George, Indianapolis, Ind. Purifying benzine.....	Dec. 13, 1870
110,055	Same.....Lamp-burner. (Antedated Nov. 26, 1870).....	Dec. 13, 1870
105,953	Lushor, John, La Porte, Ind. Vegetable-cutter.....	Aug. 2, 1870
	Lusk, Andrew P., <i>et al.</i> (See Reinshagen, Peter W., assignor.)	
	Same..... (See Wiehl, Daniel, assignor.)	
	Lusk, Salmon B. (See Mallory, Orson E., assignor.)	
104,609	Luther, Henry C., Providence, R. I., and Celius E. Richards, North Attleborough, Mass. Picture-knob.....	June 21, 1870
	Luther, Jonathan. (See Burns, W. H., assignor)..... (Three cases.)	
	Same..... (See Long, Charles B., assignor.)	
104,970	Luther, Justus P., and Solon K. Back, Berlin, Wis. Clamp for making whips...	July 5, 1870
99,780	Luther, L. T., Oak Grove, Pa. Match-machine.....	Feb. 15, 1870
102,413	Luther, Ormel R., Waterbury, Conn. Adjustable foot for clock-cases.....	April 26, 1870
100,049	Lutz, Georg, John Schultheis, and Michel Florentin, Newark, N. J. Vegetable-cutter.....	Feb. 22, 1870
102,563	Lutz, George, Danforth H. Royce, Michael Trenor, and Robert Chadwick, Columbus, Ohio. Policeman's nippers.....	May 3, 1870
103,918	Lutz, John A., Waynesborough, Va. Cotton-chopper.....	Nov. 1, 1870
	Lutz, Joseph F. C., and John W. Pearman. (See Pearman & Lutz.)	
102,952	Lutz, Stimmel, Philadelphia, Pa. Fruit-can.....	May 10, 1870
101,478	Luxton, Charles, Hudson City, N. J. Peat-machine.....	April 5, 1870
	Lyman, Azel S., New York, N. Y. Method of cooling and ventilating rooms, &c..... (Extension).....	Mar. 24, 1870
103,478	Lyman, Chester C., Edinborough, Pa. Platform-scale.....	May 24, 1870
3,799	Lyman, John N., New York, N. Y. Stay-log for cutting veneers.... (Reissue).....	Jan. 18, 1870
100,164	Same.....Clamp for stay-logs.....	Feb. 22, 1870
100,774	Lyman, Myron W., Chicago, Ill. Animal-trap.....	Mar. 15, 1870
102,951	Lyman, William W., Meriden, Conn. Apparatus for preserving fruit.....	May 10, 1870
105,346	Same.....Can-opener.....	July 12, 1870
105,583	Same.....West Meriden, Conn. Can-opener.....	July 19, 1870
108,802	Lynch, Charles S., Boston, Mass. "Fix" for puddling-furnaces.....	Nov. 1, 1870
108,277	Lynch, George F., Milwaukee, Wis. Railway-car axle-box.....	Oct. 11, 1870
	Lynch, J. Augustus. (See Hantoon, Reuben K., assignor.)	
105,101	Lynch, John H., New York, N. Y. Ice-chamber for refrigerators.....	July 5, 1870
99,578	Lynch, Matthew, New York, N. Y. Hoisting-machine.....	Feb. 8, 1870
101,634	Lynch, Nathaniel K., New York, N. Y. Variable cut-off valve-gear.....	April 5, 1870

Notes: This page contains the William Lyman can opener patent.

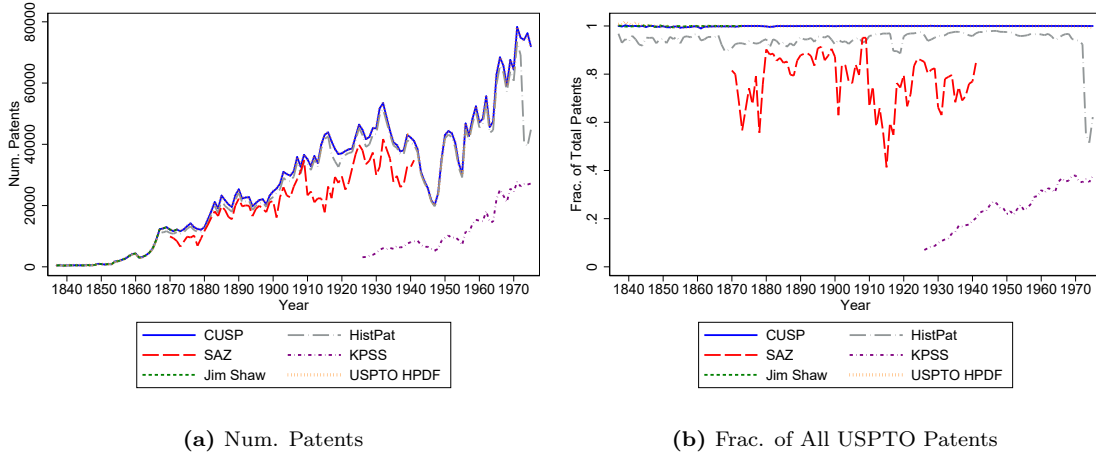
Figure 4: A Page of the 1874 Subject-Matter Index of Patents for Inventions

Index of patents issued from the United States Patent Office from 1790 to 1873, inclusive—Continued.

Invention.	Inventor.	Residence.	Date.	No.
Can—Continued.				
See Packing and atomizing can.				
Paint-can.				
Preserving-can.				
Rectangular can.				
Roofing-can.				
Roving-can.				
Safety-can.				
Scaled can.				
Sealing-can.				
Seamless can.				
Sheet-metal can.				
Shipping-can.				
Sirup-can.				
Spice-can.				
Sprinkling-can.				
Tea-can.				
Tin can.				
White-lead can.				
Wood-incased can.				
Wooden-cased can.				
Can	A. D. Armstrong	Pittsburgh, Pa.	Sept. 24, 1867	69, 154
Can	C. A. Murdock	Milwaukee, Wis.	May 20, 1873	139, 181
Can and bottle, Sealing	J. D. Willoughby	Carlisle, Pa.	Jan. 4, 1859	22, 535
Can and jar, Sealing	J. Bellerjean	Philadelphia, Pa.	Mar. 31, 1868	76, 149
Can-cap, Tuncut	J. I. Livingston	Pittsburgh, Pa.	Mar. 2, 1869	87, 415
Can-cover lock	H. W. Shepard	New York, N. Y.	Apr. 9, 1872	125, 622
Can-cover lock	H. W. Shepard	New York, N. Y.	Apr. 9, 1872	125, 623
Can-filler	R. Newton	Millville, N. J.	May 6, 1873	138, 574
Can filling and soldering apparatus	L. C. Straub	Pittsburgh, Pa.	June 20, 1871	116, 114
Can-handle	L. A. Sunderland	Madison, Ohio	Nov. 30, 1869	97, 326
Can-hook	G. Webber	Portland, Me.	Sept. 11, 1849	6, 702
Can-making die	J. L. Gray	Baltimore, Md.	Apr. 2, 1867	63, 503
Can-manufacture	J. T. Ackley and J. K. Trux	Philadelphia, Pa.	Apr. 10, 1866	53, 765
Can-opener	H. C. Alexander	New York, N. Y.	Nov. 16, 1869	96, 761
Can-opener	H. C. Alexander	New York, N. Y.	Jan. 25, 1870	99, 046
Can-opener	R. H. Atwell	Baltimore, Md.	Feb. 9, 1869	86, 626
Can-opener	A. Barker	Wyoming, Pa.	May 17, 1870	103, 125
Can-opener	F. G. Beach	Hartford, Conn.	Apr. 6, 1869	88, 536
Can-opener	W. M. Bleakley	Verplank, N. Y.	June 29, 1869	91, 902
Can-opener	W. M. Bleakley	Verplank, N. Y.	Oct. 19, 1869	95, 873
Can-opener	S. O. Church	West Meriden, Conn.	Jan. 15, 1867	61, 161
Can-opener	M. C. Davis	Folsom, Cal.	July 13, 1869	92, 520
Can-opener	E. F. Dewey	San Francisco, Cal.	Sept. 28, 1869	95, 205
Can-opener	E. M. Dewey	San Francisco, Cal.	Aug. 29, 1871	118, 593
Can-opener	W. L. Hubbell	Brooklyn, N. Y.	Oct. 23, 1867	69, 996
Can-opener	G. C. Humphreys	Washington, D. C.	Nov. 17, 1868	84, 122
Can-opener	G. G. Joyce	Baltimore, Md.	Aug. 10, 1869	93, 541
Can-opener	J. Kaufman	New York, N. Y.	Sept. 6, 1870	107, 061
Can-opener	O. J. Livermore	Worcester, Mass.	June 26, 1866	55, 878
Can-opener	W. W. Lyman	Meriden, Conn.	July 12, 1870	105, 346
Can-opener	W. W. Lyman	West Meriden, Conn.	July 19, 1870	105, 553
Can-opener	T. A. McFarland	Meadville, Pa.	May 21, 1867	64, 891
Can-opener	C. J. C. Petersen	Port Chester, N. Y.	June 17, 1873	140, 072
Can-opener	A. C. Platt	Sandusky, Ohio	Aug. 23, 1870	106, 723
Can-opener	J. J. Reed	Lyons, Iowa	Mar. 25, 1873	137, 149
Can-opener	C. F. Ritchel	Chicago, Ill.	May 12, 1868	77, 916
Can-opener	L. B. Smith	West Meriden, Conn.	June 17, 1873	140, 088
Can-opener	N. F. Stone	Chicago, Ill.	Apr. 14, 1868	76, 669
Can-opener	W. Thomas	Geneseo, Ill.	Nov. 26, 1872	133, 509
Can-opener	S. E. Totten	Brooklyn, N. Y.	Jan. 22, 1867	61, 424
Can-opener	E. J. Warner	Waterbury, Conn.	Jan. 5, 1858	19, 063
Can-opener	J. A. Wells	Holly Springs, Miss.	Aug. 10, 1869	93, 505
Can-opener	J. Wood	New York, N. Y.	July 8, 1873	140, 604
Can-opener	F. S. Wyman	Chicago, Ill.	July 28, 1868	80, 326
Can-opener and knife or fork, Combined	T. Kenderdine	Lisbon, Iowa	Aug. 26, 1873	142, 109
Can-opener and pipe-cutter	D. A. Barnes	Chicago, Ill.	Oct. 28, 1873	144, 051
Can-opening machine	W. H. Forker	Meadville, Pa.	Oct. 29, 1867	70, 188
Can-opening tool	G. A. Dickson	Woodcock Township, Pa.	Dec. 24, 1867	72, 464
Can-opening tool	M. T. McCormick	Meadville, Pa.	Apr. 7, 1868	76, 450
Can-opening tool	E. T. Orne	Chicago, Ill.	Nov. 6, 1866	59, 513
Can or bottle stopper	J. Drenton	Philadelphia, Pa.	Dec. 23, 1862	37, 221
Can or canister top	A. Bliss	Newark, N. J.	Oct. 21, 1851	8, 440
Can or flask	J. Duntun	Philadelphia, Pa.	Mar. 3, 1863	37, 843
Can screw-tap	L. R. Boyd	New York, N. Y.	June 13, 1871	115, 927

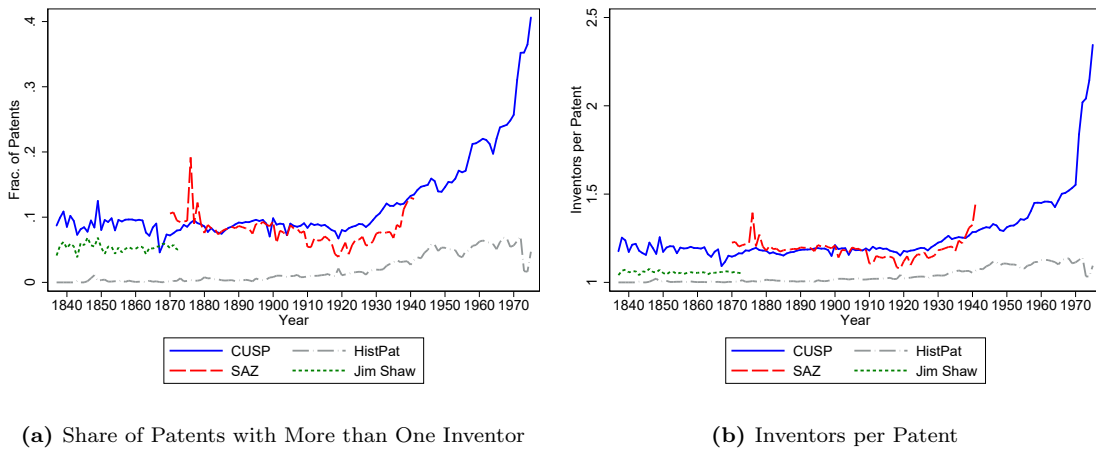
Notes: The Subject-Matter Index contains all patents issued from from 1790 to 1873. This page is from the second volume of the index. The page contains the William Lyman can opener patent.

Figure 5: Number of Patents



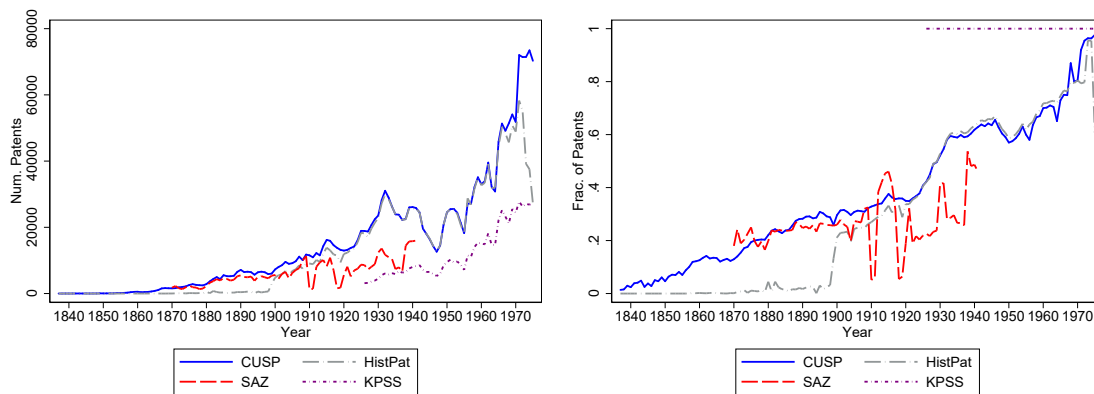
Notes: Panel (a) plots the number of patents in each year using each patent dataset. Panel (b) plots the fraction of the total USPTO aggregate patents in each year using each patent dataset.

Figure 6: Multiple Inventors



Notes: Panel (a) plots the share of patents with more than one inventor in each year using each patent dataset. Panel (b) plots the average number of inventors per patent in each year using each patent dataset.

Figure 7: Assigned Patents

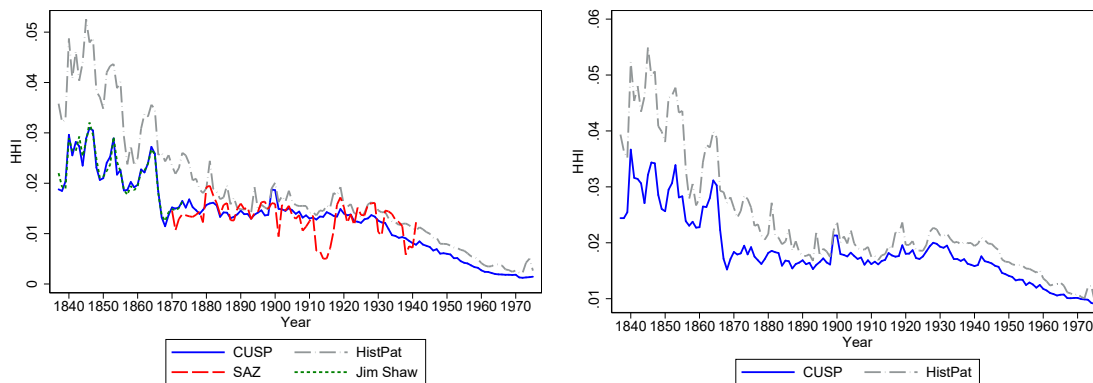


(a) Number of Assigned Patents

(b) Share of Assigned Patents

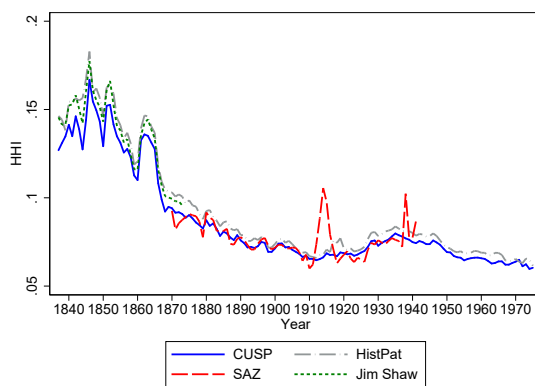
Notes: Panel (a) plots the number of patents that are assigned in each year using each patent dataset. Panel (b) plots the share of each dataset's patents that are assigned in each year using each dataset.

Figure 8: HHI of Patents by Location



(a) Town

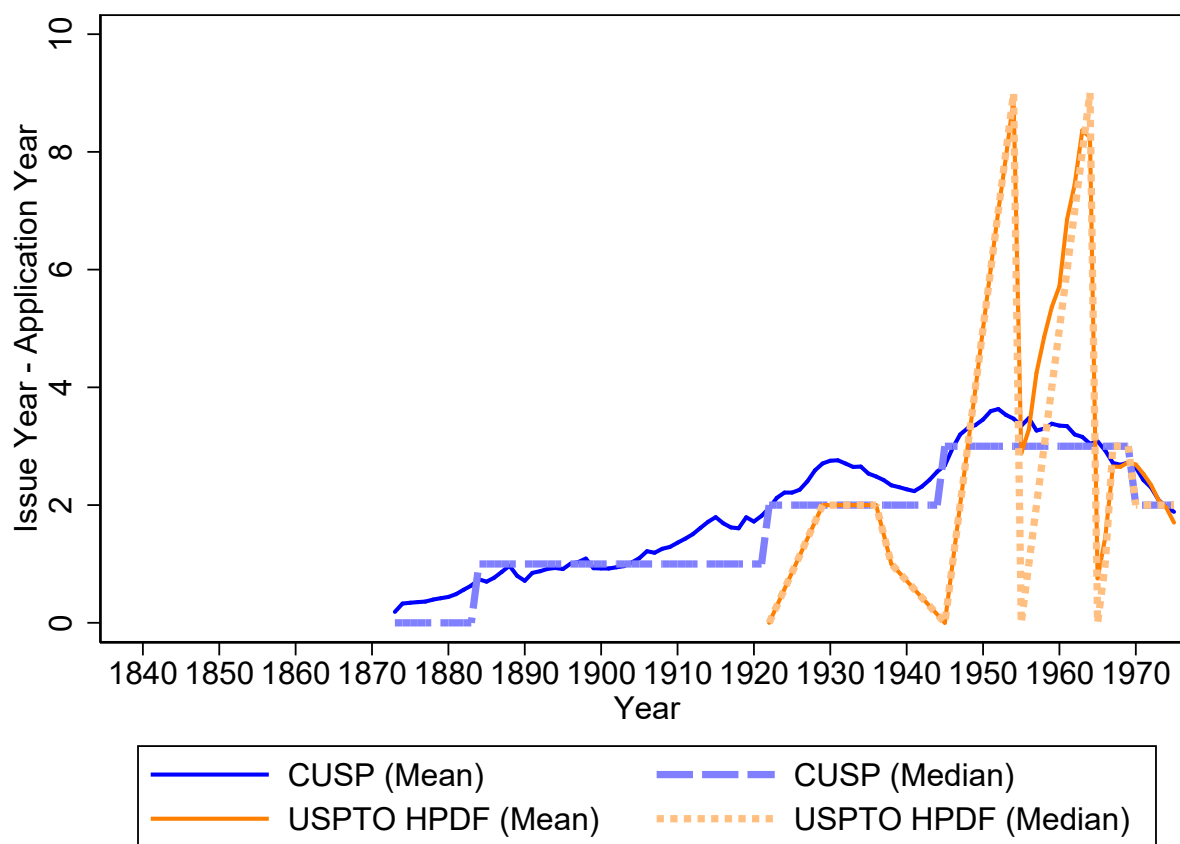
(b) County



(c) State

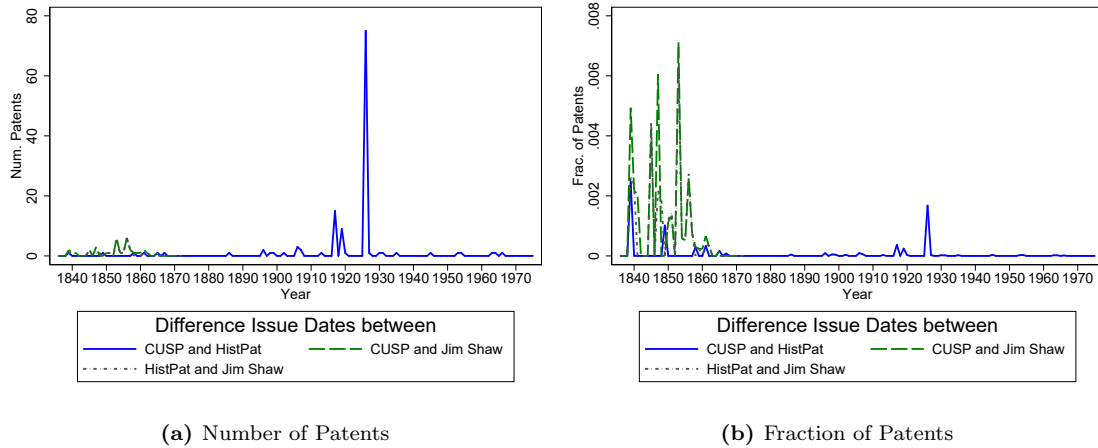
Notes: Herfindahl-Hirschman index of patenting by location each year using each patent dataset. Panel (a) plots HHI by town. Panel (b) plots HHI by county. Panel (c) plots HHI by state.

Figure 9: Patent Grant Delay



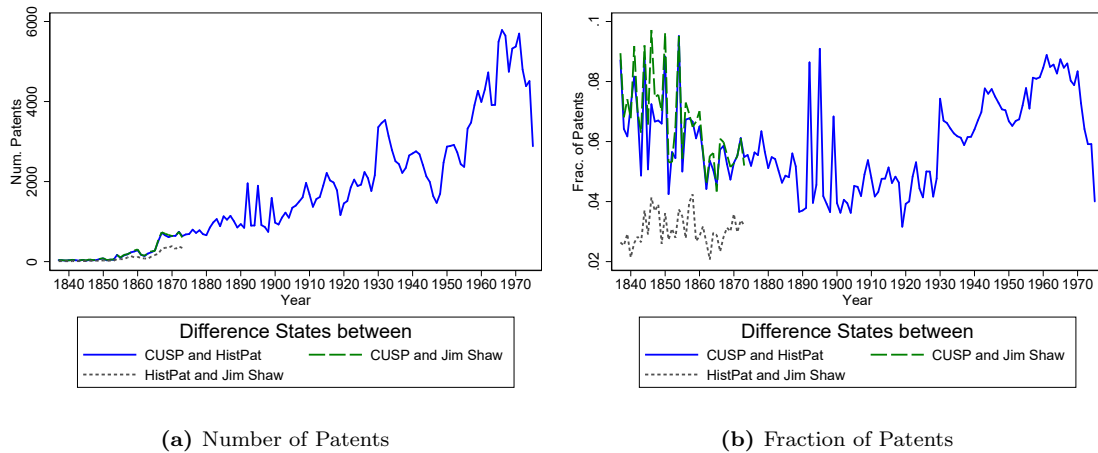
Notes: The difference between the issue year and the application year in each year using the CUSP and HPDF datasets. The mean difference between issue year and application year is plotted in the solid line, while the median difference is plotted in the dashed line.

Figure 10: Patents that Have Different Issue Years in Different Datasets



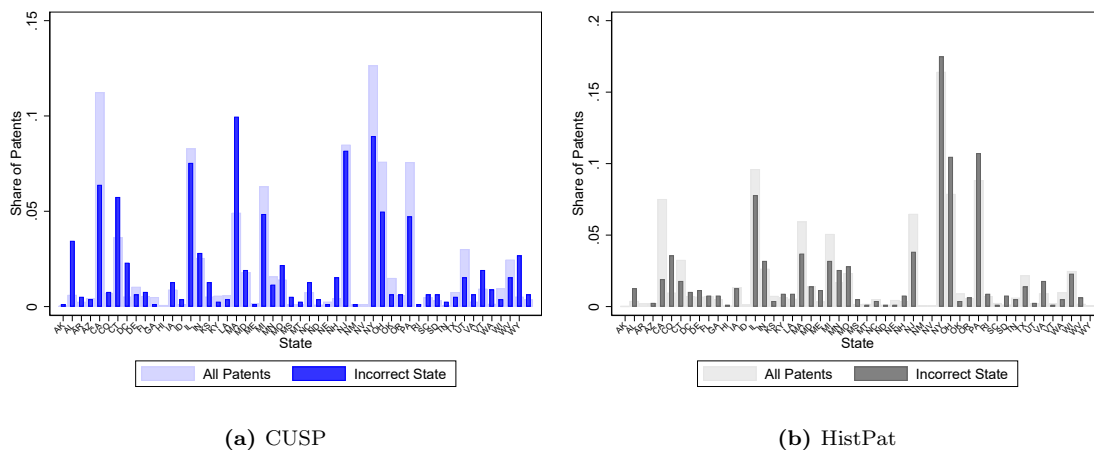
Notes: The number (panel a) and fraction (panel b) of patents that record a different issue year in one dataset compared to another.

Figure 11: Patents that Have Different Inventor States of Residence in Different Datasets



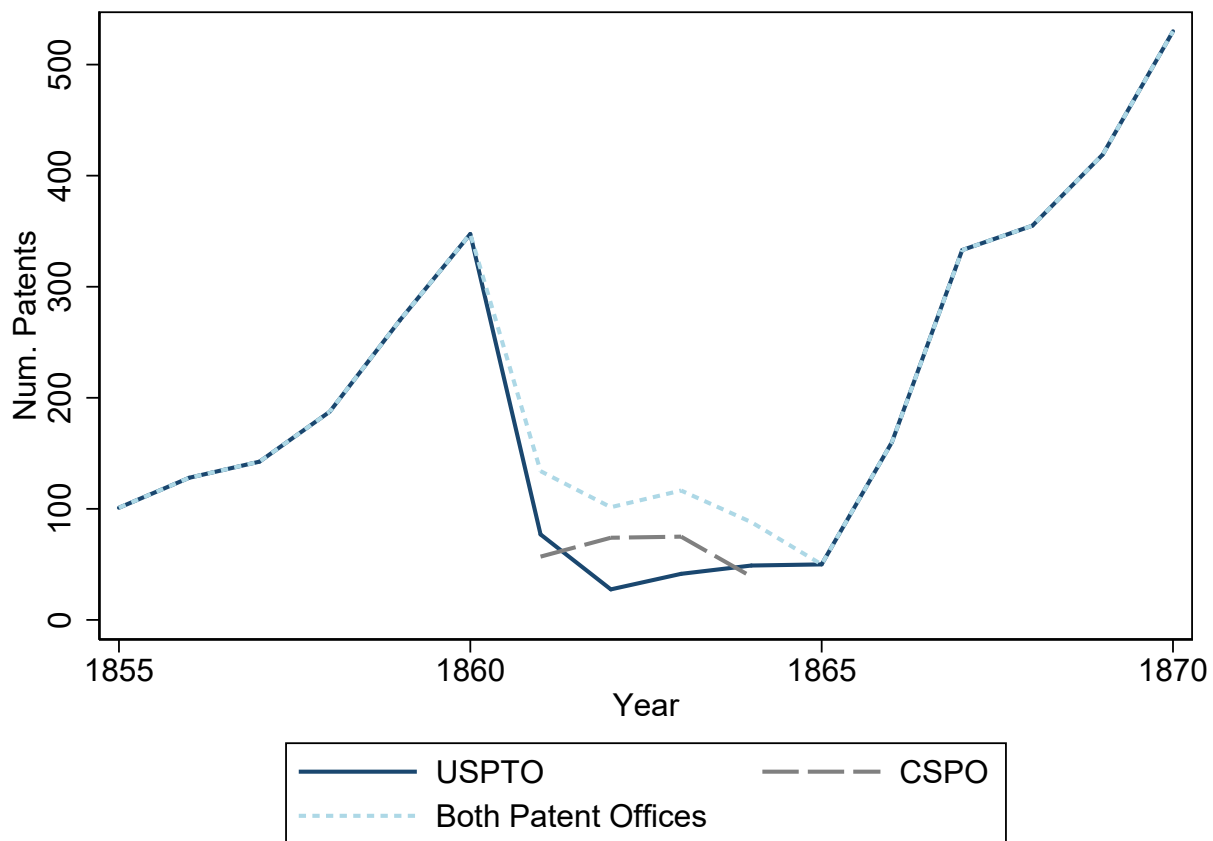
Notes: The number (panel a) and fraction (panel b) of patents for which the state of residence of the first listed inventor is different in one dataset compared to another.

Figure 12: Comparing the Distribution of Inventor States that Are Incorrect to Inventor States for All Patents



Notes: The distribution of inventors' states of residence for all patents in a dataset (the wide, light bars) and the distribution of inventors' states of residence for patents in which the inventor state of residence is recorded in error (thin, dark bars).

Figure 13: Patenting in Confederate States



Notes: The number of patents in the states and territories that made up the Confederate States of America; see the text for list of states and territories. The blue line plots patents in these places issued by the USPTO for the years 1855 to 1870. The gray dashed line plots patents in these places issued by the CSPO for the years 1861 to 1864. The light blue dotted line is the sum of the USPTO and CSPO patents for these places.

Online Appendix to:
Historical Patent Data: A Practitioner’s Guide

Michael Andrews*

August 11, 2020

A A Brief History of the USPTO Patent Images

The USPTO has made multiple attempts to digitize historical patents and make these images accessible to researchers and the wider public. The first attempts by the USPTO took place in the 1980s and 1990s using early OCR software. These are available at https://bulkdata.uspto.gov/data3/patent/grant/multipagetiff/1790_1999/. Not surprisingly given the time period in which the data were originally constructed, the quality of the image capture is quite low. This first attempt at OCR is affectionately known as “the Dirty File” within the USPTO (Raider, 2016).¹ Since then, the USPTO has entered into agreements with Google and Reed Tech to re-run newer versions of OCR software on the USPTO image files. These more successfully-digitized images are now hosted by the USPTO at <https://www.uspto.gov/learning-and-resources/bulk-data-products> and available for bulk download as

*National Bureau of Economic Research. 1050 Massachusetts Ave., Cambridge, MA 02138. *Email*: mandrews@nber.org.

¹Ran Raider a former president of the PTRC and currently works at the PTRC office in Dayton, OH.

TIFF or PDF files. These images are also hosted on Google Patents. In addition, Google has merged the OCR-translated patent text with bibliographic information from the European Patent Office Worldwide Bibliographic Database (PATSTAT).²

²Information on the process by which Google Patents digitized the USPTO patent images is obtained from Google Patents Team (2016). Many researchers use PATSTAT as the exclusive source of patent data, particularly when studying contemporary patenting. For historical patents, however, information on inventor names and/or locations is frequently missing or inconsistently recorded in PATSTAT (Billington & Hanna, 2020). The PATSTAT data can be obtained at <http://www.epo.org/searching-for-patents/technical/docdb.html#tab1>. See European Patent Office (2015) and Kang and Tarasconi (2016) for more details on using PATSTAT.

B Examples of Image Quality Issues in the Annual Reports

Figure A1: A Page of the 1870 Annual Report of the Commissioner of Patents with a Smudge

COMMISSIONER OF PATENTS.		23
<i>Alphabetical list of patentees for the year 1870—Continued.</i>		
No.	Name, residence, and invention or discovery.	Date.
102,472	Badger, William, Hastings-on-Hudson, N. Y. Washing-machine	May 3, 1870
	Baer, John H., et al. (See Deeds, William S., assignor.)	
99,132	Bagg, Albert G., New York, N. Y. Cheese-press	Jan. 25, 1870
107,211	Baggs, James T., Bridgeport, Ohio. Sawing-machine	Sept. 13, 1870
105,162	Bagley, Allen, Ypsilanti, Mich. Double-acting force-pump	July 12, 1870
101,412	Bailey, Albert R., Plantsville, Conn. Die for forging shear-bows	April 5, 1870
	Bailey, Alfred, Amesbury, Mass. Pegging-jack. (Extension)	July 23, 1870
	Bailey, Allen. (See Boyd, John G., assignor.)	
107,212	Bailey, Alonzo B., Cobalt, Conn. Collin-handle	Sept. 13, 1870
100,103	Bailey, Alson S., Chicago, Ill. Coal-elevator	Feb. 22, 1870
103,960	Bailey, Bennett C., Constitution, Ohio. Pump	June 7, 1870
100,719	Same. Washing-machine	Nov. 23, 1870
100,371	Bailey, Charles S., Mobile, Ala. Hydraulic stop-valve. (Antedated Nov. 19, 1870)	Nov. 22, 1870
106,984	Bailey, Charles W., Boston, Mass. Detachable boot and shoe heel	Sept. 6, 1870
99,747	Bailey, Fortune L., Freeport, Ind. Washing-machine	Feb. 15, 1870
	Bailey, George W., and A. H. Knapp. (See Knapp & Bailey.)	
107,213	Bailey, Henry Dwight Wright, Sterling, Ill. Spring-bed bottom	Sept. 13, 1870
101,367	Bailey, H. P., Amsterdam, N. Y. Tablet for toms	April 5, 1870
103,278	Bailey, Jacob B., New York, N. Y. Handle for knives	May 24, 1870
105,833	Bailey, John A., assignor to self, L. E. Webb, Joseph Lederle, Anthony Lederle, and John H. Wells, Detroit, Mich. Excavating-machine	Aug. 2, 1870
103,279	Bailey, John W., Aurora, Ind. Steam-engine	May 24, 1870
98,837	Bailey, Joseph R., Woonsocket, R. I. Tool-holder	Jan. 18, 1870
98,838	Same. same	Jan. 18, 1870
105,766	Bailey, Joseph B., assignor to self and Selden A. Bailey, Woonsocket, R. I. Box-scaper	July 26, 1870
105,767	Same. Bench-plane	July 26, 1870
4,057	Bailey, Ralph P., assignor of part interest to Platt D. Babbitt, Niagara Falls, N. Y. Machine for dressing marble. (Reissue)	July 5, 1870
106,765	Bailey, Timothy, Wyoming, Ill. Combined harrow and roller	Aug. 30, 1870
107,434	Bailey, William, Troy, N. Y. Hydrant	Sept. 30, 1870
4,155	Same. same. (Reissue)	Oct. 18, 1870
	Same. (See Kirk, Solomon W., assignor.)	
98,907	Bailey, W. N., Duplain, Mich. Combined latch and lock	Jan. 18, 1870
98,906	Baily, A. H., and W. G. Bratton, Marseilles, Ill. Leather channeling and folding tool	Jan. 18, 1870
108,231	Bainbridge, David, Philadelphia, Pa. Ayril-rod. (Two cases.)	Oct. 11, 1870
99,620	Bair, Brothers & Co. (See Bowker, James, assignor.)	
102,473	Baird, Francis R., Norfolk, Va. Hasp fastening for fruit-crates	Feb. 8, 1870
	Baird, James M., assignor to self, John McLure, James W. Ward, and John Boone McLure, Wheeling, West Va. Fastening for bedsteads, tables, &c. (Antedated April 30, 1870)	May 3, 1870
100,584	Baird, James R., Vincennes, Ind. Carriage wheel	Mar. 8, 1870
106,247	Baird, John, and Warren Fisher, Hamilton, Ohio. Mica chimney for lamps	Aug. 9, 1870
101,210	Baird, Joseph H., Oakville, Conn. Hinge machine	Mar. 29, 1870
106,985	Baird, Thomas W., Bowling Green, Ky. Fireplace	Sept. 6, 1870
	Baisly, George H., and John Wilson. (See Wilson & Baisly.)	
104,246	Baker, Albert C., assignor to self and George L. Laffin, Westfield, Mass. Steam-fanator	June 14, 1870
98,543	Baker, Ansel Granville, and George Moyer Eunis, New Bedford, Mass. Axle-box for carriages	Jan. 4, 1870
104,536	Baker, Benjamin, Addison, Mich. Grain-measuring attachment to thrashing-machines	June 21, 1870
	Baker, Charles J., and George A. Anderson. (See Anderson & Baker.)	
106,986	Baker, Charles L. W., Hartford, Conn. Mop. (Antedated August 25, 1870.)	Sept. 6, 1870
	Baker, C. R. (See Noyes, Walter B., assignor.)	
98,508	Baker, David H., and C. S. Southwick. (See Southwick & Baker.)	
	Baker, George W., Lincoln, Ill. Scaffold	Jan. 18, 1870
	Same. (See Morris, Lewis, assignor.)	
98,658	Baker, Haydn M., Washington, D. C. Process of cleaning cotton-waste and other filers from oil, &c.	Jan. 11, 1870
101,969	Same. Manufacture of steel from cast or pig-iron	April 19, 1870
102,906	Same. Manufacture of soap	May 10, 1870
108,028	Same. Process for the manufacture of soda, hydrochloric acid, &c.	Oct. 11, 1870
108,029	Same. Brooklyn, Eastern Division, N. Y. Manufacture of aluminate of soda	Oct. 11, 1870
105,884	Baker, Haydn M., assignor to John M. Smith, Washington, D. C. Preparation of paper stock	Aug. 2, 1870
100,090	Baker, Haydn M., Williamsburg, N. Y. Manufacture of nitric acid	Oct. 11, 1870
99,648	Baker, Henry, assignor to self and Christian G. Herr, Lancaster, Pa. Stove-pipe damper	Jan. 25, 1870
100,562	Baker, Henry P., Centreville, Ind. Drain-tile machine	Dec. 6, 1870
110,333	Same. Combined hand-roller and marker	Dec. 20, 1870
3,812	Baker, Henry N., assignor through mesne assignments to Gold and Stock Telegraph Company, New York, N. Y. Printing-telegraph. (Reissue)	Jan. 25, 1870
	Baker, Henry W., Binghamton, N. Y. Printing-telegraph. (Extension)	April 28, 1870
3,854	Baker, Horace, assignor to Richard K. Sanford, Cortland, N. Y. Hay-raker and loader. (Reissue)	Mar. 1, 1870
104,408	Baker, Isaac L., Prairie City, Kansas. Weaning-bit	June 21, 1870
104,056	Baker, Ismel, Tomar, Wis. Washing-machine	June 14, 1870
	Baker, James H., et al. (See Warriner, Baker & Slocum.)	

Notes: A dark mark about two-thirds of the way down the page makes the first names of inventors in four patent records unreadable to most OCR software.

Figure A2: A Page of the 1920 Annual Report of the Commissioner of Patents with Distortion

UNITED STATES PATENT OFFICE, 1920.
Alphabetical list of patents—Continued.

167

Field, Crosby, Brooklyn, N. Y., assignor to Chemical Machinery Corporation, New York, N. Y. Drying process and apparatus therefor. No. 1,358,431; Nov. 9; v. 280; p. 304.

Field, Edward A., assignor of one-half to E. A. Field, Jr., Chicago, Ill. Connecting-rod assembly. No. 1,329,889; Feb. 3; v. 271; p. 9.

Field, Edward A., Jr., Grand Rapids, Mich., assignor of one-fourth to E. A. Field, Sr., Chicago, Ill. Engine-lubricator. No. 1,329,888; Feb. 3; v. 271; p. 8.

Field, Edward A., Jr., Grand Rapids, Mich., assignor of one-half to E. A. Field, Sr., Chicago, Ill. Internal combustion engine. No. 1,333,611; Mar. 16; v. 272; p. 384.

Field, Edward A., Jr., Grand Rapids, Mich., assignor of one-half to E. A. Field, Sr., Chicago, Ill. Internal combustion engine. No. 1,339,471; May 11; v. 274; p. 234.

Field, Edward A., Jr., Oak Park, assignor of one-fourth to E. A. Field, Sr., Chicago, Ill. Governor. No. 1,354,608; Oct. 1; v. 279; p. 32.

Field, Edward A., Sr. (See Field, E. A., Jr., assignor.)

Field, Edward A., Sr. (See Field, Edward A., Jr., assignor.)

Field, Joseph C., Orange, N. J., assignor to Western Electric Company, Incorporated, New York, N. Y. Signaling system. No. 1,333,014; Mar. 9; v. 272; p. 312.

Field, Joseph C., Orange, N. J., assignor to Western Electric Company, Incorporated, New York, N. Y. Selectively-operated circuit-controlling device. No. 1,343,256; June 15; v. 275; p. 438.

Field, Joseph C., Orange, N. J., assignor to Western Electric Company, Incorporated, New York, N. Y. Impulse-transmitting device. No. 1,354,814; Oct. 5; v. 279; p. 14.

Field, Willard F., and C. D. Lanning, Boston, Mass., assignors, by mesne assignments, to Barber-Colman Company, Rockford, Ill. Mechanism for operating upon warp threads or the like. No. 1,341,705; June 29; v. 273; p. 871.

Field, Oscar S., Elizabeth, N. J., assignor to Hall Switch & Signal Co. Alternating current relay. No. 1,331,070; Mar. 9; v. 272; p. 222.

Field, Samuel (See Sulman, H. L. and Field.)

Field, Samuel, assignor to The Metals Extraction Corporation, Limited, London, England. Purification of zinc solutions. No. 1,331,334; Feb. 17; v. 271; p. 471.

Field, Samuel, assignor to The Metals Extraction Corporation, Limited, London, England. Purification of zinc solutions. No. 1,337,058; Apr. 13; v. 273; p. 318.

Field, Thaddeus S., Atlanta, Ga. Horseshoe. No. 1,347,023; July 29; v. 276; p. 453.

Field, William H., Jr., Winchester, Mass. Measuring instrument. No. 1,349,992; June 8; v. 275; p. 324.

Fields, Alpheus, et al. (See Parker, George C., assignor.)

Fields, Clara G., Huntington, W. Va. Baby-band. No. 1,359,344; Nov. 16; v. 280; p. 524.

Fields, Eliza E., Marion, Ohio. Combined curtain and shade support. No. 1,344,208; June 22; v. 273; p. 708.

Fields, James A., and J. J. White, San Antonio, Tex. Tire-chains. No. 1,334,263; Mar. 23; v. 273; p. 372.

Fields, William J. (See Gilbert, Battle B., assignor.)

Fiero, George M., Pittsburgh, Pa. Rail-clamp. No. 1,332,117; Feb. 24; v. 271; p. 639.

Fietzsch, John F., Forest Park, Ill. Device for measuring material for tooth-fillings. No. 1,353,407; Sept. 21; v. 278; p. 437.

Fife, Riley C., Toledo, Ohio. Garment-fastener. No. 1,369,282; Nov. 30; v. 280; p. 842.

Fife, Riley C., Toledo, Ohio. Garment-fastener. No. 1,369,283; Nov. 30; v. 280; p. 842.

Fife, Albert F. (See Barron, W. L. and Fife.)

Fife, Albert F., Newark, N. J., assignor to The Singer Manufacturing Company. Thread-severing mechanism for sewing-machines. No. 1,346,814; July 20; v. 276; p. 418.

Fifield, Chester C., Des Moines, Iowa. Non-freezable water-gate. No. 1,348,480; Aug. 3; v. 277; p. 70.

Fifield, Walter, Waterbury, Conn. Interchangeable-jawed tool. No. 1,338,005; Apr. 27; v. 273; p. 640.

Fifield, Irvin, assignor to the Winton, Edward, assignor.

Figli, Francesco, Genoa, Italy. Portable testing-machine. No. 1,321,491; Mar. 2; v. 272; p. 61.

Figueroa, William L., assignor of one-half to O. Heymann, Philadelphia, Pa. Bowling-alley. No. 1,329,235; Jan. 27; v. 279; p. 569.

Field, Nivola. (See Evans, William F., assignor.)

Figueroa, Maurice, assignor to Etablissements Continouza, Sides Anonyme, Paris, France. Star-wheel transmission. No. 1,338,823; May 4; v. 274; p. 32.

Figueroa, Peter F., New York, N. Y. Bathing-cabinet. No. 1,331,312; Feb. 10; v. 271; p. 212.

Figueroa, Peter F., New York, N. Y. Display-exhibitor. No. 1,341,119; May 25; v. 274; p. 669.

Filar Electric Heater, The. (See Smith, William H., assignor.)

Fils, Ira, Waterville, Me. Trap-setting device. No. 1,354,341; Nov. 9; v. 280; p. 242.

Films, Arthur J., Chicago, Ill. Smoke-jack. No. 1,349,666; Aug. 17; v. 275; p. 425.

Filkins, George H., Watervliet, assignor of one-half to J. P. Murray, Troy, N. Y. Automobile device. No. 1,328,954; Jan. 20; v. 270; p. 431.

Fills, James, Wilmington, Del. Pipe-coupling. No. 1,354,815; Oct. 5; v. 279; p. 74.

Filman, John H. (See Hays, J. M. and Filman.)

Finsberg, David H., St. Louis, Mo. Shoe-heel. No. 1,335,371; Mar. 30; v. 272; p. 822.

Finsberg, Harry D., Roxbury, and S. S. Levy, Dorchester, Mass. Heater. No. 1,355,220; Oct. 12; v. 279; p. 206.

Finsberg, Joseph, and C. T. Heller, Jr., St. Paul, Minn. Non-refillable bottle. No. 1,362,264; Dec. 14; v. 281; p. 361.

Fischer, Albert B., assignor to Belt Grip Pulley Company, Buffalo, N. Y. Caster. No. 1,342,600; June 8; v. 275; p. 226.

Fischer, Clifton R., Newark, N. J. Framework for display publicity by machine. No. 1,336,831; Apr. 13; v. 273; p. 273.

Fisch, Clifton R., Newark, N. J. Exhibiting or display device. No. 1,358,531; Nov. 9; v. 280; p. 321.

Fisch, Emory D., Atlanta, N. Y. Automobile-stop. No. 1,334,147; Mar. 16; v. 272; p. 485.

Fisch, Horner L. (See Winsor, W. G. and Fisch.)

Fisch, John S. (See Eames, G. M. and Fisch.)

Fisch, William. (See Arnett, Albert C., assignor.)

Fisch, William G. H., New York, N. Y. Wireless receiving and recording apparatus. No. 1,331,958; Feb. 17; v. 271; p. 429.

Fischer, Fritz, Essen, assignor to Fried. Krupp Aktiengesellschaft, Essen-on-the-Ruhr, Germany. Turntable for railway-guns. No. 1,360,977; Nov. 23; v. 280; p. 694.

Fischer, Fritz, Essen, assignor to Fried. Krupp Aktiengesellschaft, Essen-on-the-Ruhr, Germany. Gun with railway gun-carriage. No. 1,360,512; Nov. 30; v. 280; p. 807.

Fischer, Fritz, Essen, assignor to Fried. Krupp Aktiengesellschaft, Essen-on-the-Ruhr, Germany. Turntable for railway-guns. No. 1,360,513; Nov. 30; v. 280; p. 808.

Fischer, Fritz, Essen, assignor to Fried. Krupp Aktiengesellschaft, Essen-on-the-Ruhr, Germany. Support for the gun-carriages of railway-guns on turntables. No. 1,360,514; Nov. 30; v. 280; p. 808.

Finselsen & Kropf Mfg. Co. (See Edwards, William A., assignor.)

Finselsen & Kropf Manufacturing Company. (See Rayfield, Charles L., assignor.)

Findley, A. H., et al. (See Niland, Andrew, assignor.)

Findley, Charles L., Los Angeles, Calif. Oil-agitator. No. 1,365,992; Dec. 28; v. 281; p. 724.

Findley, William A., San Francisco, Calif. Tool. No. 1,331,847; Sept. 17; v. 271; p. 47.

Fine, Jacob, Louisville, Ky. Apparatus for applying fasteners. No. 1,355,457; Oct. 12; v. 279; p. 252.

Finger, J. P., Hummer, and F. Croghan, Scranton, Pa. Removable brass for incombustible boxings and locking means therefor. No. 1,358,596; Nov. 9; v. 280; p. 274.

Finger, Will, Gastonia, N. C. Whiffletree. No. 1,356,058; Oct. 19; v. 279; p. 409.

Fingerhut, Carl, Chicago, Ill. Picture-frame. No. 1,363,463; Dec. 28; v. 281; p. 627.

Finkbein, Thomas D., Kitchawan, N. Y. Attaching-plug. No. 1,331,634; Feb. 24; v. 271; p. 549.

Fink, Charles D., Toledo, Ohio. Hog-feeder. No. 1,333,800; Mar. 16; v. 272; p. 429.

Fink, Colin G., East Orange, N. J., assignor to General Electric Company, Albany, N. Y. No. 1,342,993; June 8; v. 275; p. 324.

Fink, Colin G., Yonkers, assignor to Chiles Exploration Company, New York, N. Y. Casting ferro-manganese. No. 1,361,036; Dec. 7; v. 281; p. 11.

Fink, Esther B., Charleston, W. Va. Button-clasp. No. 1,342,163; June 11; v. 275; p. 108.

Fink, Esther B., Charleston, W. Va. Shoe-lace tip. No. 1,344,044; Dec. 28; v. 281; p. 740.

Fink, Eugene L., Dubois, Iowa. Lion's seat. No. 1,347,296; July 29; v. 276; p. 563.

Fink, Israel J., New York, N. Y. Dental drill. No. 1,358,432; Nov. 9; v. 280; p. 361.

Fink, Louis, Philadelphia, Pa. Sammersdale expiring or salvaging vessel. No. 1,346,749; July 13; v. 276; p. 341.

Fink, William H., Diagonal, Iowa. Record-repeating device for talking-machines. No. 1,331,702; Feb. 24; v. 271; p. 561.

Fink, William H., Diagonal, Iowa. Record-repeating device. No. 1,332,691; Dec. 14; v. 281; p. 209.

Fink, William O., South Range, Wis. Cattle-guard. No. 1,338,908; Apr. 27; v. 273; p. 640.

Finktoener, William M., et al. (See Thompson, Harry W., assignor.)

Finkelstein, Joseph, Brooklyn, N. Y. Curving-machine. No. 1,334,086; Mar. 16; v. 272; p. 474.

Finkelstein, Morris F., New York, N. Y. Electrical fitting. No. 1,336,290; Apr. 6; v. 273; p. 110.

Finks, Abraham J. (See Johns, C. O., and Finks.)

Finko, Coleman, Dundee, Ill. Can-opener. No. 1,341,421; May 25; v. 274; p. 728.

Notes: This page was not laying flat when it was scanned. Not only are many of the lines close to the middle of the page illegible, but it becomes difficult to determine which information belongs on the same line.

Figure A3: A Page of the 1920 Annual Report of the Commissioner of Patents with a Dark Mark

UNITED STATES PATENT OFFICE, 1920.

403

Alphabetical list of patentees—Continued.

<p>Perry, Frederick H., Beverly, Mass., assignor to United Shoe Machinery Corporation, Paterson, N. J. Sole-fitting machine. No. 1,338,957; May 4; v. 274; p. 57.</p> <p>Perry, Frederick H., Beverly, Mass., assignor by mesne assignments, to United Shoe Machinery Corporation, Paterson, N. J. Grinding machine. No. 1,347,656; July 27; v. 276; p. 687.</p> <p>Perry, Frederick H., Beverly, Mass., assignor to United Shoe Machinery Corporation, Paterson, N. J. Stitch-separating machine. No. 1,357,511; Nov. 2; v. 280; p. 55.</p> <p>Perry, Frederick H., Beverly, Mass., assignor by mesne assignments, to United Shoe Machinery Corporation, Paterson, N. J. Machine for lasting stitchdown-shoes. No. 1,363,265; Dec. 28; v. 281; p. 590.</p> <p>Perry, George E., Chicago, Ill., assignor to Western Electric Company, Incorporated, New York, N. Y. Pneumatic brush or sprayer. No. 1,332,996; Mar. 9; v. 272; p. 208.</p> <p>Perry, George H., Potsdam, N. Y. Hydrant. No. 1,325,526; Mar. 30; v. 272; p. 550.</p> <p>Perry, George H., New York, N. Y. Furniture-leveler. No. 1,348,441; Aug. 3; v. 277; p. 61.</p> <p>Perry, Grover, Chicago, Ill. Marker. No. 1,329,428; Feb. 8; v. 271; p. 16.</p> <p>Perry, Grover, Chicago, Ill. Portable bookcase. No. 1,329,429; Feb. 8; v. 271; p. 16.</p> <p>Perry, James F., Glenora, N. Mex. Boiler. No. 1,350,565; Aug. 24; v. 277; p. 645.</p> <p>Perry, John. (See Leviston, W., and Perry.)</p> <p>Perry, Lindon A. (See Clifford, F. J., and Perry.)</p> <p>Perry, Melvin W., Altona, Wis. Spark-plug. No. 1,362,504; Dec. 14; v. 281; p. 349.</p> <p>Perry, Nathan G., C. E. Hulting, and C. X. Thompson, Douglas, Ariz. Turret. No. 1,333,596; Mar. 9; v. 272; p. 304.</p> <p>Perry, Ray F., Montclair, N. J., assignor to The Barrett Company, New York, N. Y. Molded form of bitumen. No. 1,327,354; Jan. 6; v. 270; p. 88.</p> <p>Perry, Ray F., Upper Montclair, N. J., assignor to The Barrett Company. Applying waterproofing to materials and product therefrom. No. 1,331,965; Feb. 17; v. 271; p. 477.</p> <p>Perry, Ray P., Upper Montclair, N. J., assignor to The Barrett Company. Making construction material. No. 1,339,037; Aug. 27; v. 275; p. 547.</p> <p>Perry, Ray P., Upper Montclair, N. J., assignor to The Barrett Company. Centrifugal machine and disintegrating material. No. 1,352,633; Sept. 14; v. 278; p. 211.</p> <p>Perry, Ray P., assignor to The Barrett Company. Upper Montclair, N. J. Producing felt. No. 1,352,687; Sept. 14; v. 278; p. 224.</p> <p>Perry, Ray P., Upper Montclair, N. J., assignor to The Barrett Company. Process and apparatus for making composite sheets of felt and the like. No. 1,369,313; Nov. 30; v. 280; p. 893.</p> <p>Perry, Thomas O., Oak Park, Ill. Friction-clutch. No. 1,345,100; June 23; v. 275; p. 928.</p> <p>Perry, Thomas O., Oak Park, Ill. Aircraft. No. 1,345,101; June 23; v. 275; p. 928.</p> <p>Perry & Webster Incorporated. (See Webster, F. W., and Boynton, assignors.)</p> <p>Perry, William H., assignor of one-half to Normandy Sea Food Company, San Diego, Calif. Drying apparatus. No. 1,363,431; Dec. 28; v. 281; p. 621.</p> <p>Perry, William F., Leistonstone, London, England. Apparatus for distilling carbonaceous material. No. 1,341,537; May 25; v. 274; p. 546.</p> <p>Pershall, Edward E., St. Louis, Mo. Distillation. No. 1,341,437; May 25; v. 274; p. 731.</p> <p>Pershing, Frank W., Moline, Ill. Holder for use in coating rods for electric welding. No. 1,354,418; Sept. 28; v. 278; p. 673.</p> <p>Persohn, Frank A., Baltimore, Md. Lens-mount. No. 1,339,974; May 11; v. 274; p. 324.</p> <p>Person, Axel. (See Erickson, August, assignor.)</p> <p>Person, Earl B., assignor to The Viscoid Company, Leominster, Mass. Machine for feeding cylindrical blanks to heading devices. No. 1,330,450; Feb. 10; v. 271; p. 238.</p> <p>Persons-Arter Machine Company, The. (See Arter, William, assignor.)</p> <p>Persons, Charles A., Worcester, Mass. Handle-grip. No. 1,345,565; July 6; v. 276; p. 95.</p> <p>Pesci, Peter, Newark, N. J. Broken-tap remover. No. 1,351,586; Feb. 17; v. 271; p. 478.</p> <p>Peschke, Rudolph J., York, Pa., assignor to The Dentist's Supply Company. Preventing distortion of precious metals temperatures higher than their fusing points and article produced thereby. No. 1,335,024; Mar. 30; v. 272; p. 702.</p> <p>Peschke, Slave, Butte, Mont. Safety device for mine-cages. (Reissue.) No. 14,864; May 25; v. 274; p. 768.</p> <p>Peschke, Slave, Butte, Mont. Safety device for mine-cages. No. 1,354,116; Mar. 16; v. 272; p. 480.</p> <p>Petaja, John W. (See Bannano, Peter, and Petaja.)</p> <p>Peter, John, Philadelphia, Pa. Ash-maze color-preserving screen. No. 1,346,346; May 18; v. 274; p. 432.</p> <p>Pet, Frank, McDonald, Pa. Clothesline. No. 1,360,144; Nov. 23; v. 280; p. 127.</p>	<p>Petalar, Adolph, Freeport, N. Y. Lifting-jack with rotatable cams. No. 1,352,472; Sept. 14; v. 278; p. 181.</p> <p>Peter, Alfred, assignor to Sheffield Car Company, Three Rivers, Mich. Starting and reversing mechanism. No. 1,348,560; Mar. 23; v. 272; p. 607.</p> <p>Peter, Bernard H., Westminster, England, assignor by mesne assignments, to The Union Switch & Signal Company, Swiserve, Pa. Signaling system. No. 1,327,570; Jan. 6; v. 270; p. 129.</p> <p>Petera, John. (See Green, N. F., and Petera.)</p> <p>Peterman, Floyd T., et al. (See Howard, Albertson, assignor.)</p> <p>Peterman, Frederick, New York, N. Y. Doll's eyes. No. 1,332,518; Mar. 2; v. 272; p. 66.</p> <p>Peterman, James. (See Peterman, Joseph F., assignor.)</p> <p>Peterman, Joseph F., Superior, Wis., assignor of one-third to J. Peterman, Chicago, Ill. Track-brace. No. 1,361,901; Dec. 14; v. 281; p. 231.</p> <p>Petermann, Joseph, Philadelphia, Pa. Asphalt cutter mechanism. No. 1,331,967; Feb. 17; v. 271; p. 478.</p> <p>Petermann, Joseph, Philadelphia, Pa. Apparatus for cutting and breaking asphalt or concrete. No. 1,344,590; June 22; v. 275; p. 700.</p> <p>Petermann, Otto, assignor to Corona Typewriter Company, Inc., Grotton, N. Y. Typewriting-machine. No. 1,356,825; Oct. 26; v. 279; p. 697.</p> <p>Petermann, Otto, assignor to Corona Typewriter Company, Inc., Grotton, N. Y. Typewriting-machine. No. 1,356,826; Oct. 26; v. 279; p. 697.</p> <p>Peters, Andrew, Oakland, Calif. Dough cutting and molding machine. No. 1,337,152; Apr. 13; v. 275; p. 334.</p> <p>Peters, Caroline A., Los Angeles, Calif. Hat. No. 1,342,132; June 1; v. 275; p. 162.</p> <p>Peters Cartridge Company, The. (See Peck, Harley T., assignor.)</p> <p>Peters, Edward A., and J. T. Welhofer, assignors to Drew Carriage Company, Waterloo, Wis. Hoop. No. 1,337,262; Apr. 5; v. 273; p. 490.</p> <p>Peters, Emil, and A. Lambert, Chicago, Ill. Legging. No. 1,343,657; May 27; v. 275; p. 657.</p> <p>Peters, Frank A., Chicago, Ill. Curtain-support. No. 1,327,961; Jan. 13; v. 270; p. 265.</p> <p>Peters, Frederick W., West Park, assignor to The Paragon Machine Tool & Manufacturing Company, Cleveland, Ohio. Reciprocating saw. No. 1,356,169; Oct. 19; v. 279; p. 452.</p> <p>Peters, Frederick W., West Park, assignor to The Peters Machine and Manufacturing Company, Cleveland, Ohio. Universal joint. No. 1,360,788; Nov. 30; v. 280; p. 923.</p> <p>Peters, George M., Kirkwood, Mo. Cushion-wheel. No. 1,343,613; June 8; v. 275; p. 618.</p> <p>Peters, George W., and O. F. Alkira, Charleston, Wash. Electric. No. 1,356,468; Oct. 19; v. 279; p. 490.</p> <p>Peters, John F., Edgewood Park, Pa., assignor to Westinghouse Electric and Manufacturing Company, Westcott, Pa. No. 1,347,299; June 23; v. 275; p. 850.</p> <p>Peters, John F., Edgewood Park, Pa., assignor to Westinghouse Electric and Manufacturing Company, Transformer for use with rotary converters. No. 1,347,219; July 27; v. 276; p. 688.</p> <p>Peters, John F., Pittsburgh, Pa., assignor to Westinghouse Electric & Manufacturing Co. Transformer. No. 1,351,061; Aug. 31; v. 277; p. 804.</p> <p>Peters, Louis J., Brooklyn, N. Y. Match-comb. No. 1,330,279; Feb. 10; v. 271; p. 206.</p> <p>Peters, Lyle, Denver, Colo. Valve-lock. No. 1,339,412; May 11; v. 274; p. 217.</p> <p>Peters Machine and Manufacturing Company, The. (See Peters, Frederick W., assignor.)</p> <p>Peters Machine and Manufacturing Company, The. (See Thieme, William H., assignor.)</p> <p>Peters, Orland W., Los Angeles, Calif. Shoe-protector. No. 1,345,875; July 6; v. 274; p. 107.</p> <p>Peters, Percy H. (See Davis, Verner E., assignor.)</p> <p>Peters, Robert H. (See Western, A. H., and Peters.)</p> <p>Peters, Theodore, assignor to The Pet-Son Pump Company, Ferdinand, Ind. Pump. No. 1,353,492; Sept. 21; v. 278; p. 442.</p> <p>Peters, Thomas J., Peters, Fla. Tomato and fruit sorter and distributor. No. 1,340,079; May 11; v. 274; p. 343.</p> <p>Peters, Walter R., assignor of one-half to F. R. Roddall, Philadelphia, Pa. Combination-tool. No. 1,331,368; Feb. 17; v. 271; p. 478.</p> <p>Peters, William E., Brownfield, Tex. Stove. No. 1,357,182; Oct. 26; v. 279; p. 676.</p> <p>Petersen, Anker, Moulle, Minn. Shaft-protector. No. 1,361,403; Dec. 7; v. 281; p. 85.</p> <p>Petersen, Anker, Winthrop, Mass., assignor by mesne assignments, to The America Wiremold Company, Hartford, Conn. Braiding-machine. No. 1,331,672; Feb. 24; v. 271; p. 556.</p> <p>Petersen, Arnold J., Cedar Rapids, Iowa. Light-bowl hanger. No. 1,338,958; Mar. 4; v. 274; p. 57.</p> <p>Petersen, August H. (See Lacroix, J. D., and Petersen.)</p> <p>Petersen, Christian M., and O. J. Bodin, San Francisco, Calif. Electrical apparatus. No. 1,346,606; July 13; v. 276; p. 314.</p> <p>Petersen, Hans, assignor to The Linograph Company, Davenport, Iowa. Mold. No. 1,342,586; June 8; v. 275; p. 245.</p>
---	--

Notes: A dark mark on the bottom left corner of the page makes most of the patents in the lower left corner illegible.

C Summary Statistics of Patent Features by Dataset

In the following tables, I plot summary statistics for each of the patent features plotted in Section 3. Note that differences in the summary statistics are in some cases driven by the fact that different datasets cover different years. For example, the Jim Shaw data only covers years before 1874, when both co-patenting and assignment were relatively rare. Summary statistics computed over common sets of years are available upon request.

Table A1: Summary Statistics for Number of Patents

	Mean	SD	Median	Min	Max
CUSP	28,230.381	20,065.529	27,118.000	405.000	78,316.000
HistPat	26,291.590	18,215.485	25,301.000	382.000	73,453.000
SAZ	23,301.746	8,805.143	22,424.000	6,584.000	41,721.000
KPSS	12,139.360	7,822.800	8,650.500	3,133.000	27,786.000
Jim Shaw	3,946.270	4,454.985	1,891.000	403.000	12,957.000
HPDF	28,222.863	20,046.887	27,118.000	411.000	78,251.000

Notes: Summary statistics for the number of patents in each year for each of the datasets. Rows list the datasets. The first column lists the average number of patents in each dataset-year, the second column the standard deviation, the third column the median, the fourth column the minimum, and the fifth column the maximum.

Table A2: Summary Statistics for the Fraction of Aggregate Patents

	Mean	SD	Median	Min	Max
CUSP	0.999	0.002	1.000	0.990	1.002
HistPat	0.941	0.060	0.952	0.516	0.980
SAZ	0.777	0.110	0.797	0.413	0.951
KPSS	0.243	0.093	0.244	0.070	0.379
Jim Shaw	1.000	0.002	1.000	0.993	1.004
HPDF	1.000	0.003	1.000	0.993	1.017

Notes: Summary statistics for the fraction of the total number of patents issued by the USPTO in each year for each of the datasets. Rows list the datasets. The first column lists the average fraction of patents in each dataset-year, the second column the standard deviation, the third column the median, the fourth column the minimum, and the fifth column the maximum.

Table A3: Summary Statistics for Fraction of Patents with More than One Inventor

	Mean	SD	Median	Min	Max
CUSP	0.120	0.064	0.093	0.046	0.406
HistPat	0.019	0.021	0.010	0.000	0.070
SAZ	0.081	0.023	0.079	0.039	0.192
Jim Shaw	0.054	0.006	0.054	0.038	0.069

Notes: Summary statistics for the fraction of patents that have more than one inventor in each year for each of the datasets. Rows list the datasets. The first column lists the average fraction of patents in each dataset-year, the second column the standard deviation, the third column the median, the fourth column the minimum, and the fifth column the maximum.

Table A4: Summary Statistics for Inventors per Patents

	Mean	SD	Median	Min	Max
CUSP	1.265	0.187	1.194	1.093	2.345
HistPat	1.038	0.042	1.020	1.000	1.142
SAZ	1.190	0.060	1.188	1.078	1.463
Jim Shaw	1.057	0.008	1.057	1.040	1.077

Notes: Summary statistics for the number of inventors per patent in each year for each of the datasets. Rows list the datasets. The first column lists the average number of inventors per patent in each dataset-year, the second column the standard deviation, the third column the median, the fourth column the minimum, and the fifth column the maximum.

Table A5: Summary Statistics for the Number of Assigned Patents

	Mean	SD	Median	Min	Max
CUSP	15,010.108	16,959.868	9,756.000	6.000	73,454.000
HistPat	12,453.072	14,751.920	7,323.000	0.000	58,278.000
SAZ	6,377.731	3,756.731	5,501.000	1,209.000	15,999.750
KPSS	12,139.360	7,822.800	8,650.500	3,133.000	27,786.000

Notes: Summary statistics for the number of assigned patents in each year for each of the datasets. Rows list the datasets. The first column lists the average number of assigned patents in each dataset-year, the second column the standard deviation, the third column the median, the fourth column the minimum, and the fifth column the maximum.

Table A6: Summary Statistics for the Fraction of Assigned Patents

	Mean	SD	Median	Min	Max
CUSP	0.371	0.249	0.313	0.014	0.976
HistPat	0.296	0.299	0.247	0.000	0.955
SAZ	0.265	0.097	0.253	0.052	0.537
KPSS	1.000	0.000	1.000	1.000	1.000

Notes: Summary statistics for the fraction of total patents that are assigned in each year for each of the datasets. Rows list the datasets. The first column lists the average fraction of assigned patents in each dataset-year, the second column the standard deviation, the third column the median, the fourth column the minimum, and the fifth column the maximum.

Table A7: Summary Statistics for the Geographic Concentration of Patents at the Town Level

	Mean	SD	Median	Min	Max
CUSP	0.013	0.007	0.014	0.001	0.031
HistPat	0.018	0.012	0.016	0.002	0.053
SAZ	0.013	0.003	0.014	0.005	0.019
Jim Shaw	0.022	0.005	0.022	0.012	0.032

Notes: Summary statistics for the Herfindahl-Hirschman Index of patents by town in each year for each of the datasets. Rows list the datasets. The first column lists the average HHI in each dataset-year, the second column the standard deviation, the third column the median, the fourth column the minimum, and the fifth column the maximum.

Table A8: Summary Statistics for the Geographic Concentration of Patents at the County Level

	Mean	SD	Median	Min	Max
CUSP	0.019	0.006	0.017	0.009	0.037
HistPat	0.024	0.010	0.020	0.009	0.055

Notes: Summary statistics for the Herfindahl-Hirschman Index of patents by county in each year for each of the datasets. Rows list the datasets. The first column lists the average HHI in each dataset-year, the second column the standard deviation, the third column the median, the fourth column the minimum, and the fifth column the maximum.

Table A9: Summary Statistics for the Geographic Concentration of Patents at the State Level

	Mean	SD	Median	Min	Max
CUSP	0.086	0.027	0.075	0.060	0.167
HistPat	0.092	0.031	0.079	0.062	0.183
SAZ	0.077	0.009	0.075	0.060	0.106
Jim Shaw	0.136	0.022	0.141	0.096	0.177

Notes: Summary statistics for the Herfindahl-Hirschman Index of patents by state in each year for each of the datasets. Rows list the datasets. The first column lists the average HHI in each dataset-year, the second column the standard deviation, the third column the median, the fourth column the minimum, and the fifth column the maximum.

Table A10: Summary Statistics for the Average Patent Grant Delay

	Mean	SD	Median	Min	Max
CUSP	1.927	1.003	1.973	0.186	3.632
HPDF	3.731	2.546	2.726	0.000	8.900

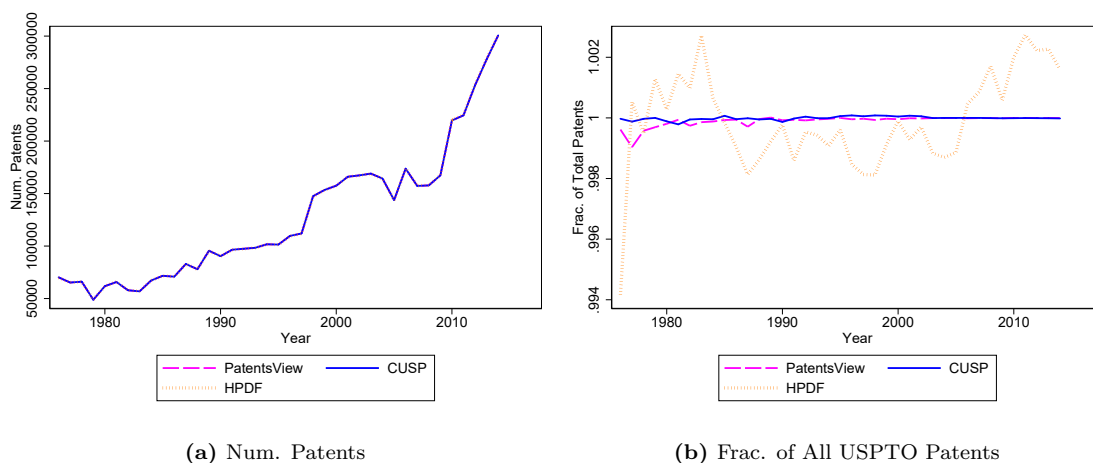
Notes: Summary statistics for the average difference between the year of patent issue and year of patent filing in each year for each of the datasets. Rows list the datasets. The first column lists the average difference in each dataset-year, the second column the standard deviation, the third column the median, the fourth column the minimum, and the fifth column the maximum.

D Modern Patent Data

The historical patent data have comparable levels of completeness as do modern patent records. Modern patents are available from several sources. First, the USPTO publishes PatentsView data for post-1976 in easily usable tab-separated files.³ The CUSP and HPDF datasets used in the paper also contain records for post-1976 patents.

In Figure A4, I repeat Figure 5 using these modern patent datasets. The first thing to note is that, even with modern patent data when all patent records have been digitized, these datasets do not contain the same numbers of patents nor do they contain 100% of aggregate patents. These modern datasets do typically contain a larger fraction of patents than the historical patent datasets, although the differences are small when using the CUSP, HPDF, or Jim Shaw data. For instance, the PatentsView data contains about 99.96% of patents in the late 1970s, a fraction that the CUSP obtains in more than 100 years between 1837 and 1975.

Figure A4: U.S.-Based Patents



Notes: Panel (a) plots the number of patents each year using each modern patent dataset. Panel (b) plots the fraction of the total USPTO aggregate patents each year using each modern patent dataset.

³These are available from <https://www.patentsview.org/download/>.

E Matching Patents to Counties

The raw data used to construct the Jim Shaw and SAZ datasets do not contain the county in which an inventor lives; instead, these data only include the town and state of a patentee. It is therefore necessary to first link towns to counties. I use the 100% U.S. decennial censuses to obtain a list of every town in each county. I next use a fuzzy matching algorithm to match the towns listed in the patent data to a town in the nearest census year, after blocking on state name. This same approach is used to match SAZ patents to counties in Andrews (2020). More precisely, I match using Stata's `relink` command, which is a modified bigram string comparator that returns a "distance" (match score) between two strings.⁴

The necessity of the fuzzy matching procedure depends on the underlying quality of the town strings. Since the Jim Shaw data was transcribed by hand, OCR errors were avoided and thus for most years the fuzzy matching procedure matches few additional towns to their counties relative to simply exactly comparing strings. I show this graphically in Figure A5, which plots the number of patents' towns successfully matched to a county with both the fuzzy matching algorithm as well as when town names are required to match exactly in order to record a successful match. The fuzzy matching is more valuable for the SAZ data, which is based off of OCR'd annual reports that are often of middling quality; in many years, the fuzzy matches assigns almost twice as many patents to counties as does exact matching. Results are similar using alternative weights in the fuzzy matching process.

⁴The same algorithm is used to match inventors to the US decennial census in Sarada, Andrews, and Ziebarth (2019).

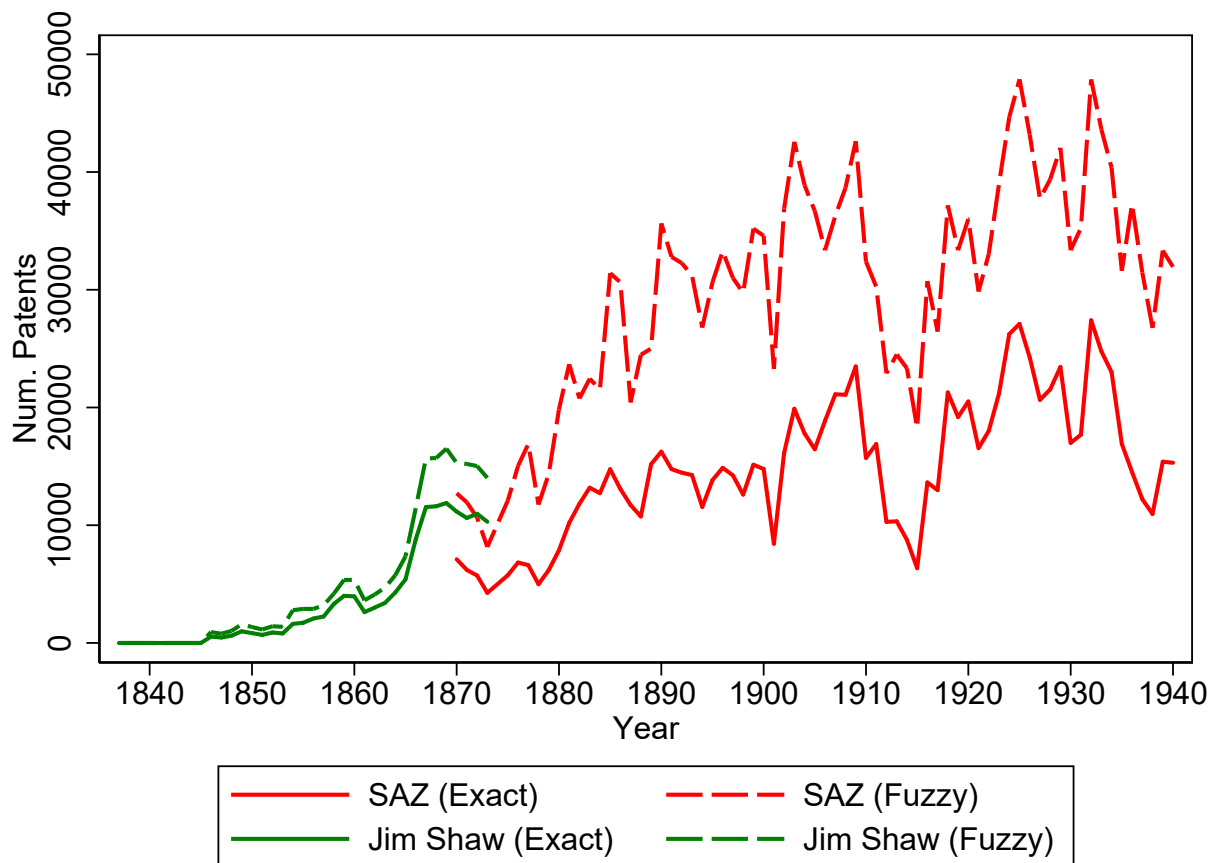


Figure A5: The number of patents' towns matched to counties in the SAZ and Jim Shaw datasets when towns are matched to counties using a fuzzy matching procedure versus when an exact match between town names is required to record a match. Dashed lines indicate fuzzy matching. Solid lines indicate exact matching.

F Other Results on Patent Locations

One might worry that the ways that the CUSP and HistPat data determine a patent's location (discussed in Section 3.5) may overstate the number of patents in highly populous counties relative to the SAZ or Jim Shaw data. The following three figures suggest that results may indeed sometimes be sensitive to how a patent's location is determined, but the most important determinant is how a patent's county is determined in the SAZ and Jim Shaw data, following the discussion in Appendix E above.

Figure A6 plots patenting concentration by county as in panel (b) of Figure 8, but using patents from the SAZ and Jim Shaw datasets that are matched to counties as described in Appendix E above. The Jim Shaw patents match to a very small number of counties, resulting in high concentration indices, for years before 1845, so I omit them for readability. These data are available upon request. Regardless of how patents are matched to counties, the SAZ and Jim Shaw datasets show a similar overall pattern to the CUSP and HistPat data: concentration falls until around the end of the Civil War and then stays roughly constant after that (the SAZ data does not show a decline as it starts after the large pre-Civil War decrease in concentration). For most years, fuzzy matching of towns to counties in the SAZ and Jim Shaw datasets results in concentration levels that are higher than those when using exact matching and that are comparable to the concentration levels found in the CUSP and HistPat datasets.

To give a sense of an “extensive margin” of locations where inventive activity occurs, Figure A7 plots the number of counties that have at least one patent in a given year in each dataset. I again plot results for the SAZ and Jim Shaw datasets using both methods to

match patents to counties. For most years, the exact-matched SAZ and Jim Shaw datasets report smaller numbers of counties with at least one patent than do the CUSP or HistPat datasets, while the fuzzy-matched SAZ and Jim Shaw datasets report a larger number of counties. Across all datasets, the number of counties in which patenting activity occurs grows over time until about 1920, when it falls until 1940. In the CUSP and HistPat datasets, the number of counties with at least one patent remains roughly stable from the 1940s until about 1980; these results are available upon request.

Similar to Figure A7, Figure A8 shows the number of patents from counties with small populations across various datasets. Panel (a) plots the number of patents from counties with populations of 2,500 or smaller. These are the smallest of counties. For all years, only a small number of patents come from these counties, and the number declines to almost zero in the 1940s. Panel (b) plots the number of patents from counties with populations of 50,000 or smaller. While this includes many patents in early years, the number of patents peaks around 1920 and declines thereafter. Panel (c) plots the number of patents from the smallest 20% of counties; by using population percentiles rather than an arbitrary level of population, the number of counties is not eroded by population growth. The pattern in panel (c) is very similar to that in panel (b). Finally, panel (d) plots the number of patents from the smallest 90% of counties; this effectively includes all counties except for those with a major city. The number grows until about 1920, then falls until 1940 before rising again after World War II. In all four panels, the highest number comes when the SAZ or Jim Shaw data is matched to counties using fuzzy-matching techniques, with the lowest number usually coming when the SAZ data is matched by exact name.

In Figures A6- A8, the results for the SAZ and Jim Shaw datasets are quite sensitive to

how patents' towns are matched to counties. These choices appear to dwarf differences in how patent locations are obtained from the raw patent data (see Section 2; in the CUSP and HistPat datasets the location must be inferred from largely unstructured patent text, while in the SAZ and Jim Shaw datasets the location is parsed from a syntactically simple row of text). Moreover, the difference in concentration at the town and state level is often larger between the CUSP and the HistPat than it is between the CUSP and the SAZ and the difference between the HistPat and CUSP is often larger than the difference between the HistPat and the Jim Shaw.

I next use patents assigned to counties to determine whether, across datasets, patents are likely to be coming from the same counties and counties with similar characteristics. I begin by describing the “average” county according to each patent dataset. To do this, I calculate the mean of several county characteristics, weighting by the number of patents in each county. County characteristics are from Manson, Schroeder, Riper, and Ruggles (2019). Results are presented in Table A11. Each panel computes the descriptive statistics of the “average” county over different sets of years, facilitating comparisons across the different datasets. On average, the counties in which patenting occurs appear quite similar across the different datasets, with the only exception that patents come from smaller counties on average in the SAZ and Jim Shaw data (when fuzzy-matched to counties) than in the CUSP and HistPat.

Table A12 shows how a one percent increase in patenting in a given county and year in one dataset is correlated with additional patents in the same county and year in another dataset. For instance, the CUSP and HistPat data are very highly correlated: a one percent increase in patenting in a given county and year in the CUSP is correlated with a .98% increase

in patenting in the same county in the CUSP. The correlations are smaller, although still large and statistically significant, between the SAZ data and the CUSP and HistPat data. Table A13 presents similar results, but all regressions include year fixed effects. Residual correlations are similar, although typically slightly smaller, and all are still highly statistically significant. In sum, counties and years that have more patents in one dataset also tend to have more patents in other datasets.

Table A14 estimates elasticities of patenting with respect to population in the four different datasets. A 10% increase in a county's population increases patenting by about 2.6% in the CUSP and HistPat and by about 1.4% in the SAZ. The elasticity in the Jim Shaw data is sensitive to how a patent's county is determined, ranging from 0.36% to 1.6%. In all cases, the correlations are highly statistically significant.

In Figures A9-A11, I take a different approach to see if patents are coming from the "same locations" across datasets. This allows a better visualization of which locations are accounting for more patents in different datasets and years. For each state s , I plot the ratio of the share of patents coming from s in a given year t in dataset i to dataset j . I also plot the distribution of patents by state in dataset j . When the ratio is close to one, the two datasets have similar shares of patents coming from a given state. When comparing SAZ to CUSP (Figure A9), SAZ to HistPat (Figure A10), or CUSP to HistPat (Figure A11), ratios are close to one for the states that account for a large share of patents. Large deviations occur in cases of states that have very small numbers of patents, such as Idaho or Wyoming. In these cases, when the share of patents in a given state are very small in each dataset, small deviations can result in ratios of shares that are very far from one.

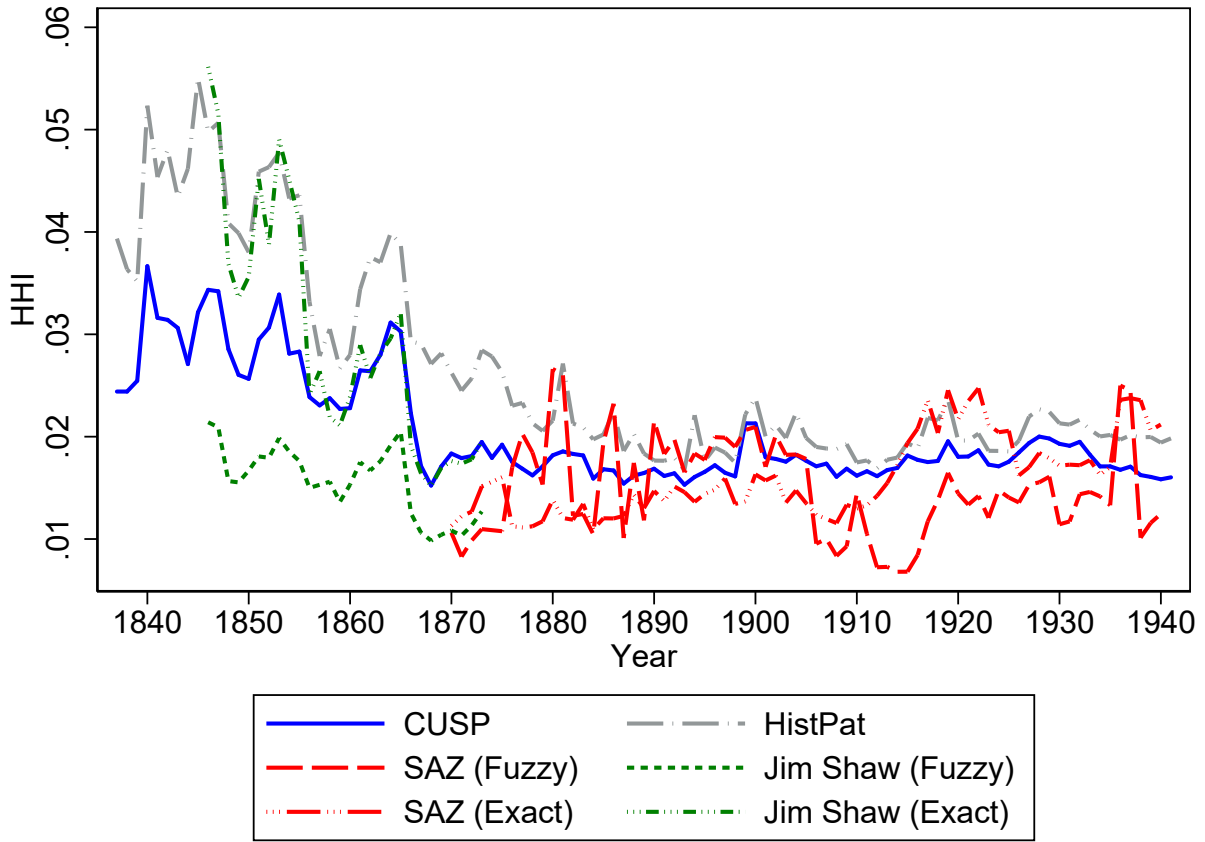
To further put these state results in perspective, in Tables A15-A20 provides the absolute

number of patents in each state and each of the CUSP, HistPat, and SAZ datasets for each year plotted in Figures A9-A11. Recall that, when a patent has multiple inventors and those inventors live in different states, each state will a fractional count equal to the fraction of that patent's inventors that live in that state; hence these counts need not be whole numbers.

One aspect of the state-by-state results deserves special mention. When comparing SAZ to both the CUSP and HistPat data in Figures A9-A11, the SAZ patents invariably have a much larger share of patents from the state of Georgia. Tables A15-A20 show that SAZ has a large absolute number of patents from Georgia as well. The SAZ patents erroneously record patents from Germany as being from Georgia. In many cases, the Annual Reports abbreviate state names using only the first two or three letters of a state's name. For this reason, the algorithm that identifies state names in SAZ codes any state that begins with "Ge" as Georgia. So almost all of the SAZ patents from Georgia are mis-located.

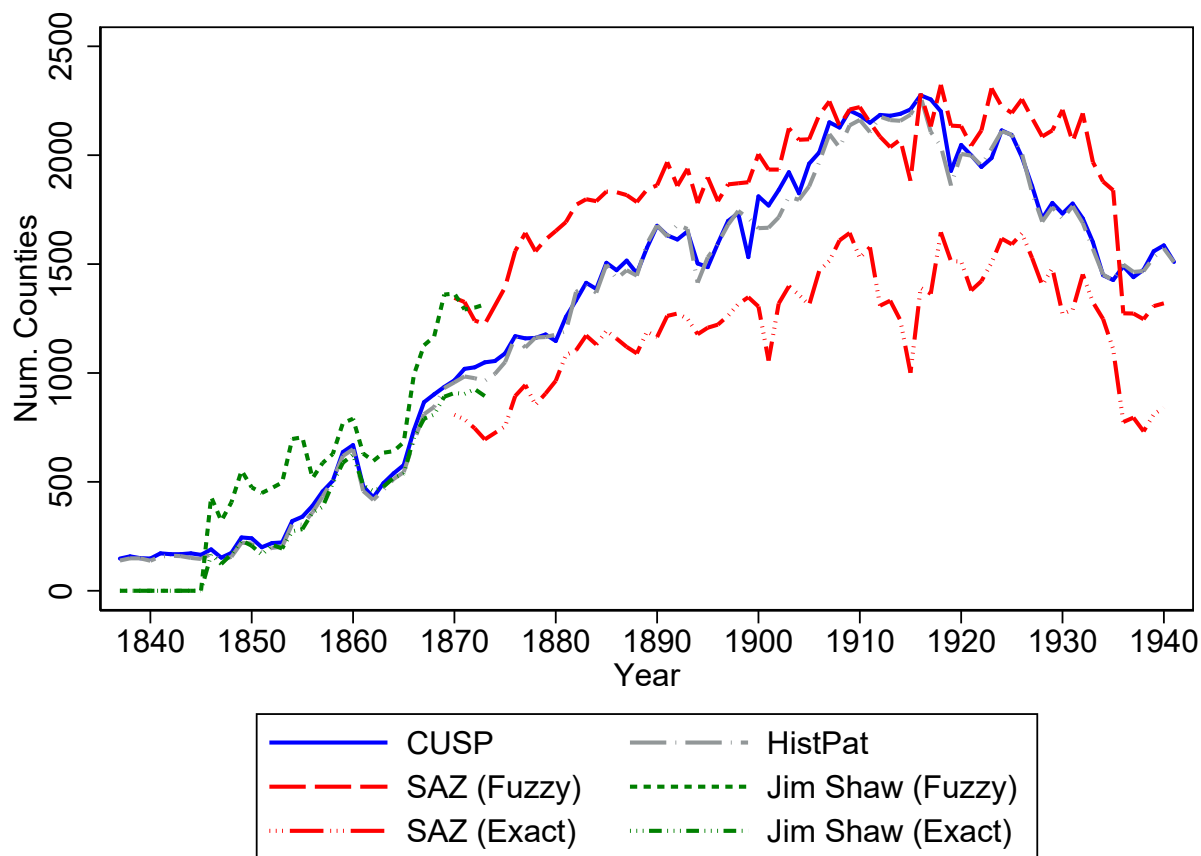
Additional correlations between patenting and locational characteristics across datasets are available upon request.

Figure A6: HHI of Patents by County



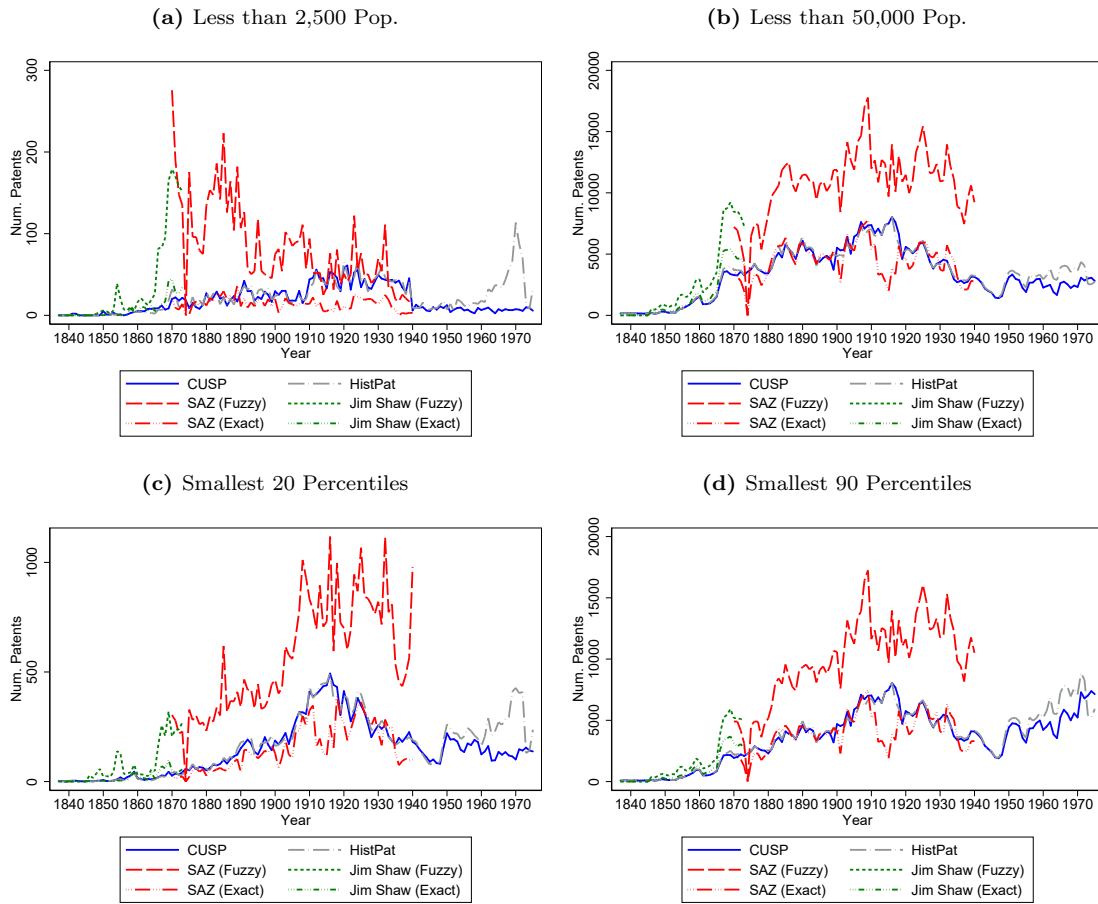
Herfindahl-Hirschman index of patenting by county each year using each patent dataset.

Figure A7: Number of Counties with At Least One Patent



The number of counties that contain one or more patents in a given year using each patent dataset.

Figure A8: Patents from Small Counties



Notes: The number of patents from counties with small population each year using each patent dataset. Panel (a) plots the number of patents in counties with fewer than 2,500 population. Panel (b) plots the number of patents in counties with fewer than 50,000 population. Panel (c) plots the number of patents in counties in the smallest 20% of counties. Panel (d) plots the number of patents in counties in the smallest 90% of counties.

Table A11: Descriptive Statistics of the “Average” County Across Datasets

(a) 1837-1873

	CUSP	HistPat	SAZ (Fuzzy)	Jim Shaw (Fuzzy)
log(Pop.)	11.225	11.267	10.892	10.795
Frac. Urban	0.462	0.472	0.382	0.323
Frac. Immigrant	0.188	0.191	0.197	0.155
log(Manuf. Output)	15.859	15.946	15.472	15.119
log(Farm Output)	14.368	14.391	14.391	14.366

(b) 1870-1940

	CUSP	HistPat	SAZ (Fuzzy)
log(Pop.)	12.252	12.295	11.555
Frac. Urban	0.488	0.489	0.404
Frac. Immigrant	0.286	0.284	0.228
log(Manuf. Output)	18.199	18.247	17.033
log(Farm Output)	14.635	14.634	14.448

(c) 1837-1975

	CUSP	HistPat
log(Pop.)	12.556	12.584
Frac. Urban	0.449	0.437
Frac. Immigrant	0.175	0.172
log(Manuf. Output)	18.095	18.138
log(Farm Output)	14.857	14.860

Notes: The mean of various county characteristics weighted by the number of patents in each column dataset. Each panel calculates the mean over a different set of years.

Table A12: Comparing Patents in the SAZ Dataset to the CUSP and HistPat Datasets

(a) CUSP			
	HistPat	SAZ (Fuzzy)	JimShaw (Fuzzy)
CUSP Patents	0.978*** (0.000472)	0.953*** (0.00147)	1.115*** (0.00186)
N	285392	140931	48346
Adj. R-sq	0.938	0.749	0.881
Years	1837-1975	1870-1940	1837-1873
(b) HistPat			
	CUSP	SAZ (Fuzzy)	JimShaw (Fuzzy)
HistPat Patents	0.959*** (0.000463)	0.956*** (0.00143)	1.088*** (0.00177)
N	285392	141929	48386
Adj. R-sq	0.938	0.758	0.887
Years	1837-1975	1870-1940	1837-1873
(c) SAZ			
	CUSP	HistPat	JimShaw (Fuzzy)
SAZ Patents	0.786*** (0.00121)	0.793*** (0.00119)	1.024*** (0.00371)
N	140931	141929	7926
Adj. R-sq	0.749	0.758	0.906
Years	1870-1940	1870-1940	1870-1873
(d) Jim Shaw			
	CUSP	HistPat	SAZ (Fuzzy)
Jim Shaw Patents	0.790*** (0.00132)	0.815*** (0.00132)	0.900*** (0.00108)
N	48346	48386	41235
Adj. R-sq	0.881	0.887	0.944
Years	1837-1873	1837-1873	1837-1873

Notes: Correlations between patents across datasets. The unit of observation is the county by year. In each cell, the inverse hyperbolic sine of patents in the column is regressed on the inverse hyperbolic sine of patents in the row. Results can be interpreted as elasticities. Panel (a) shows how patenting in each dataset varies with changes in patenting in the CUSP, panel (b) the HistPat, panel (c) the SAZ (when matching patents to counties using the fuzzy matching procedure), and panel (d) the Jim Shaw (when matching patents to counties using the fuzzy matching procedure). Standard errors are in parentheses. Stars indicate statistical significance: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A13: Comparing Patents in the SAZ Dataset to the CUSP and HistPat Datasets, Including Year Fixed Effects

(a) CUSP			
	HistPat	SAZ (Fuzzy)	JimShaw (Fuzzy)
CUSP Patents	0.977*** (0.000489)	0.948*** (0.00148)	1.090*** (0.00197)
N	285392	140931	48346
Adj. R-sq	0.939	0.755	0.884
Year FE	Yes	Yes	Yes
Years	1837-1975	1870-1940	1837-1873

(b) HistPat			
	CUSP	SAZ (Fuzzy)	JimShaw (Fuzzy)
HistPat Patents	0.955*** (0.000478)	0.950*** (0.00144)	1.065*** (0.00188)
N	285392	141929	48386
Adj. R-sq	0.939	0.763	0.889
Year FE	Yes	Yes	Yes
Years	1837-1975	1870-1940	1837-1873

(c) SAZ			
	CUSP	HistPat	JimShaw (Fuzzy)
SAZ Patents	0.786*** (0.00123)	0.793*** (0.00120)	1.026*** (0.00368)
N	140931	141929	7926
Adj. R-sq	0.753	0.762	0.908
Year FE	Yes	Yes	Yes
Years	1870-1940	1870-1940	1870-1873

(d) Jim Shaw			
	CUSP	HistPat	SAZ (Fuzzy)
Jim Shaw Patents	0.792*** (0.00143)	0.816*** (0.00144)	0.884*** (0.00139)
N	48346	48386	41235
Adj. R-sq	0.882	0.887	0.945
Year FE	Yes	Yes	Yes
Years	1837-1873	1837-1873	1837-1873

Notes: Correlations between patents across datasets. The unit of observation is the county by year. In each cell, the inverse hyperbolic sine of patents in the column is regressed on the inverse hyperbolic sine of patents in the row, while including a year fixed effect. Results can be interpreted as elasticities. Panel (a) shows how patenting in each dataset varies with changes in patenting in the CUSP, panel (b) the HistPat, panel (c) the SAZ (when matching patents to counties using the fuzzy matching procedure), and panel (d) the Jim Shaw (when matching patents to counties using the fuzzy matching procedure). Standard errors are in parentheses. Stars indicate statistical significance: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A14: Elasticity of Patenting with Respect to County Population Across Datasets

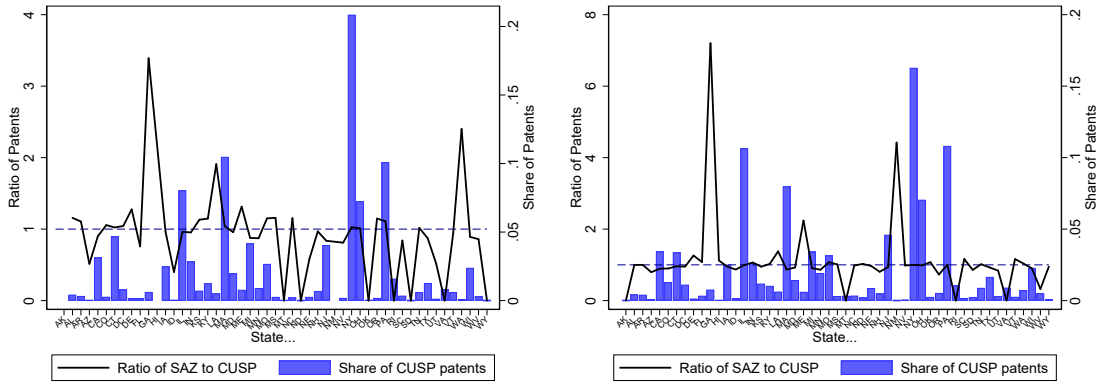
	CUSP	HistPat	SAZ (Fuzzy)	SAZ (Exact)	JimShaw (Fuzzy)	JimShaw (Fuzzy)
lhs_total_pop	0.259*** (0.0148)	0.260*** (0.0147)	0.110*** (0.00951)	0.103*** (0.00772)	0.0436*** (0.00348)	0.0355*** (0.00284)
N	301103	307932	257236	257236	147498	147498
Adj. R-sq	0.776	0.773	0.764	0.744	0.708	0.683
County FE	No	No	No	No	No	No
Year FE	No	No	No	No	No	No

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

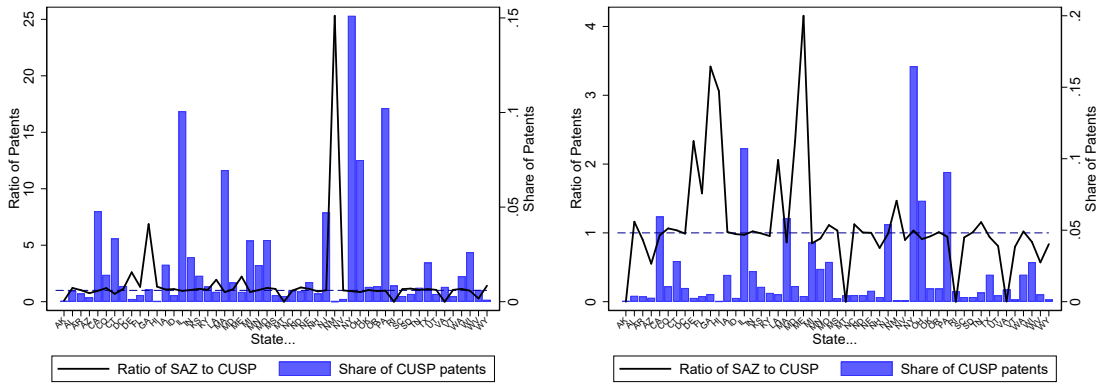
Notes: Correlations between patenting and county population across datasets. The unit of observation is the county by year. In each cell, the inverse hyperbolic sine of patents in the column is regressed on the inverse hyperbolic sine of county population, while including county and year fixed effects. Results can be interpreted as elasticities.

Figure A9: Distribution of Patents by State in CUSP and SAZ



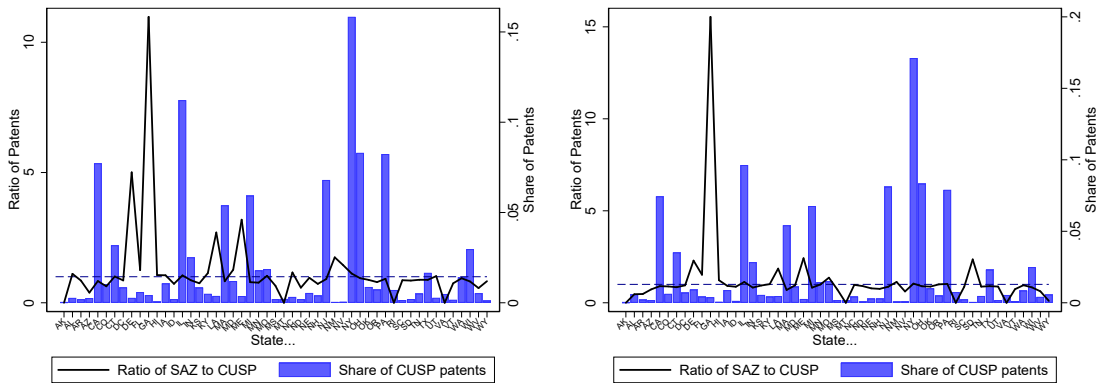
(a) 1880

(b) 1900



(c) 1910

(d) 1920

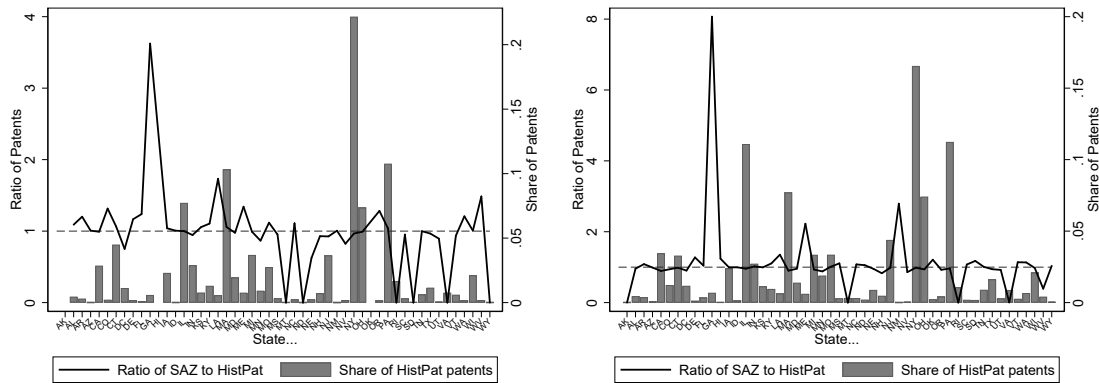


(e) 1930

(f) 1940

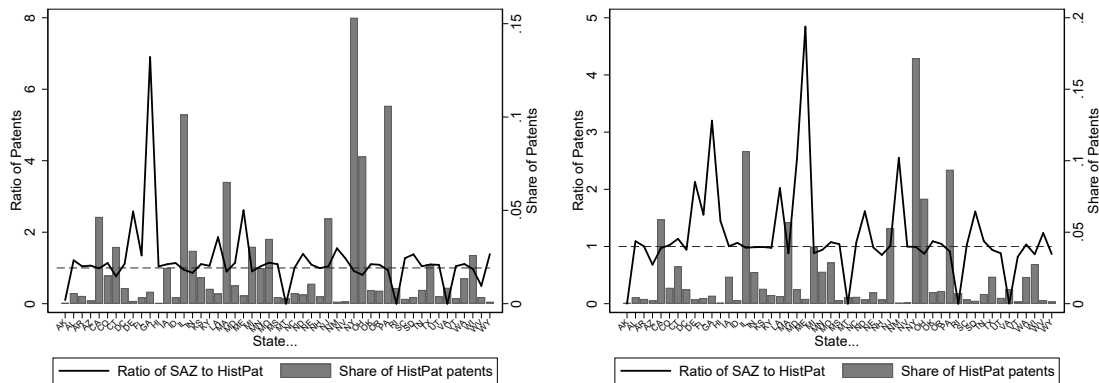
Notes: The black line plots the ratio of the share of patents in each state in SAZ to the share of patents in each state in the CUSP. A ratio of 1 is indicated by the dashed dark blue line. The blue bars plot the distribution of patents in each state in the CUSP. Results are plotted separately for the years 1880, 1900, 1910, 1920, 1930, and 1940.

Figure A10: Distribution of Patents by State in CUSP and HistPat



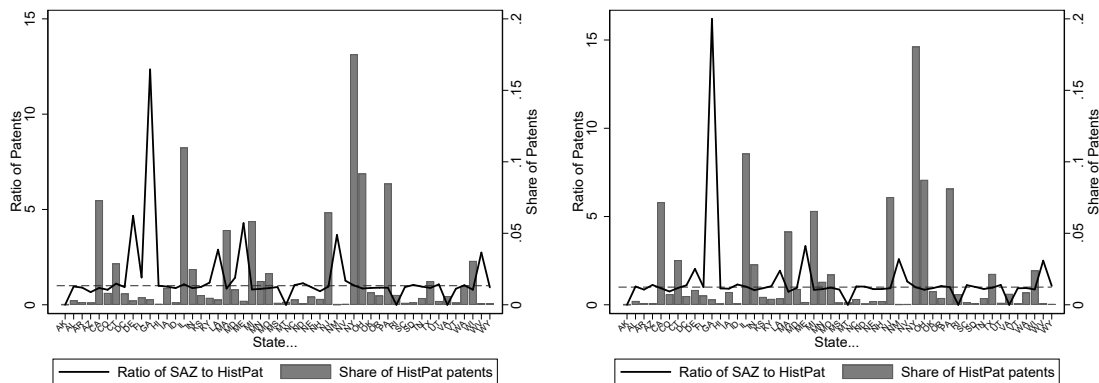
(a) 1880

(b) 1900



(c) 1910

(d) 1920

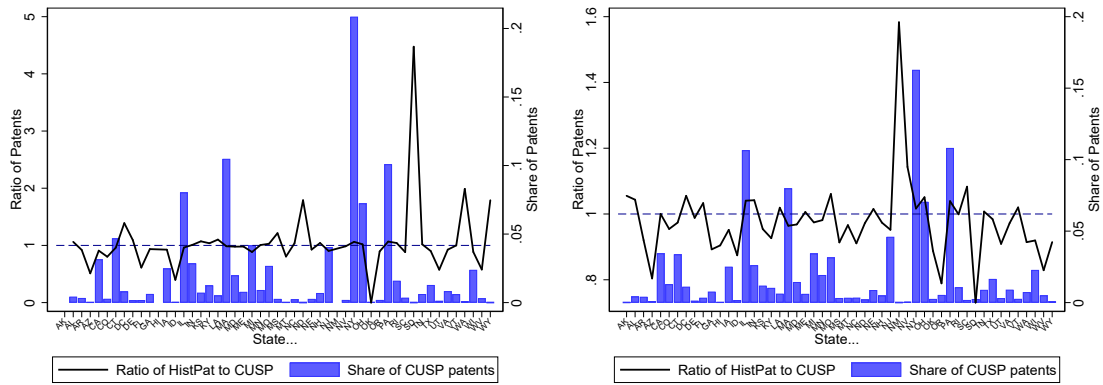


(e) 1930

(f) 1940

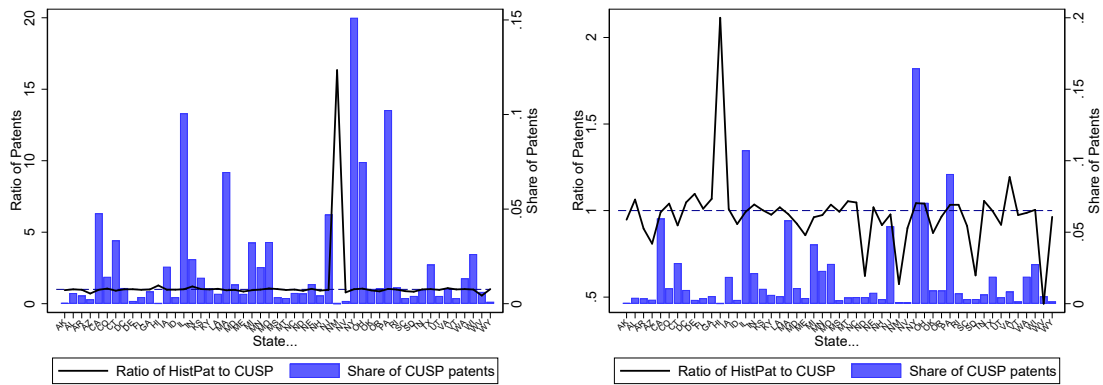
Notes: The black line plots the ratio of the share of patents in each state in SAZ to the share of patents in each state in the HistPat. A ratio of 1 is indicated by the dashed dark gray line. The gray bars plot the distribution of patents in each state in the CUSP. Results are plotted separately for the years 1880, 1900, 1910, 1920, 1930, and 1940.

Figure A11: Distribution of Patents by State in CUSP and HistPat



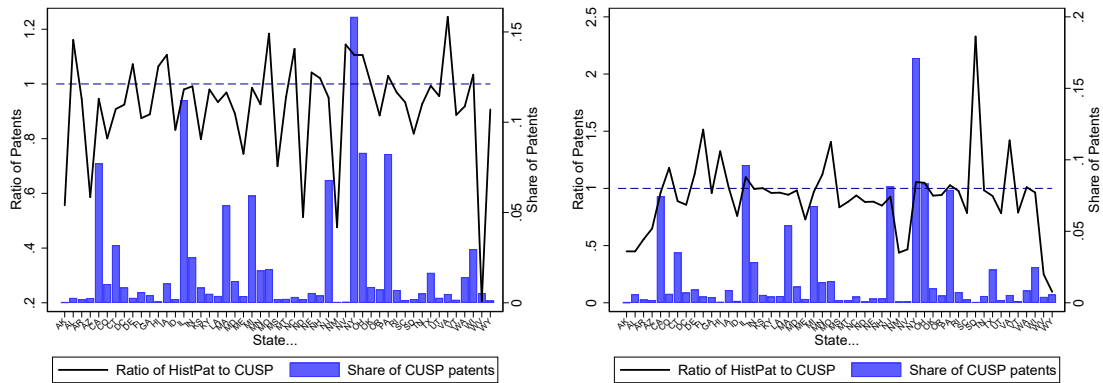
(a) 1880

(b) 1900



(c) 1910

(d) 1920



(e) 1930

(f) 1940

Notes: The black line plots the ratio of the share of patents in each state in HistPat to the share of patents in each state in the CUSP. A ratio of 1 is indicated by the dashed dark blue line. The blue bars plot the distribution of patents in each state in the CUSP. Results are plotted separately for the years 1880, 1900, 1910, 1920, 1930, and 1940.

Table A15: Counts of Patents by State and Dataset: 1880

	CUSP	HistPat	SAZ
AK			
AL	46	54	59
AR	35	36	43
AZ	4	2	2
CA	331	336	331
CO	29	26	34
CT	492	528	559
DC	85	132	99
DE	18	22	26
FL	19	13	16
GA	66	69	249
HI			
IA	261	270	279
ID	5	2	2
IL	844	908	902
IN	300	339	318
KS	76	91	96
KY	132	153	168
LA	54	67	115
MA	1,102	1,213	1,274
MD	209	228	222
ME	82	90	120
MI	439	434	427
MN	96	108	93
MO	281	323	359
MS	28	38	36
MT	3	3	0
NC	25	29	32
ND	1	2	0
NE	28	29	18
NH	73	85	79
NJ	425	430	395
NM	0	1	1
NV	20	22	18
NY	2,192	2,604	2,500
OH	761	866	854
OK	4	0	0
OR	20	20	26
PA	1,061	1,265	1,306
RI	168	195	0
SC	38	37	35
SD	1	5	0
TN	65	74	74
TX	135	136	130
UT	14	9	8
VA	87	90	0
VT	64	72	67
WA	9	20	24
WI	251	249	249
WV	33	21	31
WY	1	2	0

Table A16: Counts of Patents by State and Dataset: 1900

	CUSP	HistPat	SAZ
AK	4	4	0
AL	89	88	91
AR	86	75	88
AZ	21	16	17
CA	727	689	661
CO	270	245	247
CT	712	657	696
DC	235	235	227
DE	29	27	37
FL	72	70	79
GA	162	137	1,188
HI	7	6	8
IA	530	478	511
ID	35	29	31
IL	2,250	2,219	2,280
IN	551	544	597
KS	251	227	242
KY	217	191	227
LA	132	127	185
MA	1,687	1,541	1,492
MD	303	278	286
ME	129	123	293
MI	725	669	671
MN	407	378	357
MO	667	670	729
MS	67	58	69
MT	71	65	0
NC	73	63	73
ND	45	41	47
NE	184	177	182
NH	104	96	85
NJ	972	877	927
NM	2	3	9
NV	12	13	12
NY	3,438	3,312	3,502
OH	1,489	1,483	1,494
OK	54	45	59
OR	112	84	84
PA	2,284	2,249	2,330
RI	225	213	0
SC	38	39	45
SD	52	36	46
TN	188	179	194
TX	349	326	330
UT	65	56	56
VA	191	176	1
VT	55	53	65
WA	154	133	162
WI	482	420	442
WV	106	83	35
WY	15	13	15

Table A17: Counts of Patents by State and Dataset: 1910

	CUSP	HistPat	SAZ
AK	14	13	1
AL	174	171	159
AR	132	126	101
AZ	71	49	40
CA	1,486	1,416	1,072
CO	442	458	399
CT	1,040	925	542
DC	258	254	217
DE	44	44	86
FL	107	102	106
GA	200	195	1,031
HI	8	10	8
IA	609	584	493
ID	105	101	88
IL	3,130	3,088	2,238
IN	730	862	571
KS	427	428	363
KY	248	243	197
LA	162	166	236
MA	2,164	1,987	1,369
MD	314	297	261
ME	161	134	269
MI	1,004	928	643
MN	599	580	464
MO	1,013	1,057	929
MS	105	105	89
MT	92	86	0
NC	176	170	130
ND	167	148	158
NE	319	326	272
NH	133	119	91
NJ	1,464	1,389	1,114
NM	2	32	38
NV	44	34	33
NY	4,702	4,668	3,269
OH	2,328	2,403	1,485
OK	243	222	188
OR	251	212	177
PA	3,183	3,230	2,332
RI	263	255	0
SC	91	78	76
SD	127	104	110
TN	230	225	180
TX	642	642	540
UT	126	119	99
VA	243	259	0
VT	89	87	70
WA	415	415	353
WI	814	792	589
WV	193	107	41
WY	32	32	34

Table A18: Counts of Patents by State and Dataset: 1920

	CUSP	HistPat	SAZ
AK	17	16	0
AL	133	140	141
AR	118	105	97
AZ	89	72	45
CA	1,919	1,888	1,692
CO	345	356	336
CT	913	830	868
DC	308	321	279
DE	84	92	180
FL	122	122	175
GA	168	178	525
HI	10	21	28
IA	603	604	559
ID	84	77	75
IL	3,459	3,414	3,079
IN	687	707	643
KS	335	333	304
KY	198	192	173
LA	166	168	312
MA	1,878	1,827	1,481
MD	350	321	745
ME	125	107	476
MI	1,334	1,275	1,037
MN	739	715	621
MO	896	921	913
MS	75	74	71
MT	138	145	0
NC	147	153	151
ND	146	90	134
NE	246	249	225
NH	100	91	72
NJ	1,743	1,693	1,579
NM	35	20	47
NV	28	25	23
NY	5,307	5,497	5,026
OH	2,276	2,352	1,893
OK	296	256	258
OR	295	283	273
PA	2,924	3,001	2,526
RI	231	238	0
SC	101	91	86
SD	103	64	95
TN	205	215	217
TX	609	603	523
UT	142	130	105
VA	272	322	1
VT	47	45	34
WA	609	597	569
WI	889	887	707
WV	165	76	86
WY	47	45	36

Table A19: Counts of Patents by State and Dataset: 1930

	CUSP	HistPat	SAZ
AK	11	6	0
AL	104	118	90
AR	74	69	49
AZ	99	57	30
CA	2,977	2,761	1,930
CO	399	313	196
CT	1,227	1,093	962
DC	326	295	217
DE	104	109	404
FL	227	195	222
GA	161	141	1,375
HI	23	24	19
IA	412	447	338
ID	76	62	43
IL	4,324	4,155	3,531
IN	969	942	650
KS	323	253	188
KY	191	184	169
LA	145	133	304
MA	2,079	1,975	1,332
MD	462	404	455
ME	142	104	353
MI	2,286	2,210	1,406
MN	690	627	414
MO	717	832	577
MS	75	52	38
MT	83	77	0
NC	124	137	112
ND	76	38	34
NE	208	213	155
NH	154	154	87
NJ	2,620	2,441	1,852
NM	15	7	20
NV	21	23	23
NY	6,104	6,621	5,329
OH	3,200	3,469	2,350
OK	336	328	229
OR	281	244	173
PA	3,174	3,203	2,271
RI	272	258	0
SC	56	51	38
SD	79	63	52
TN	195	178	134
TX	639	622	437
UT	106	99	84
VA	186	227	0
VT	64	56	37
WA	545	491	399
WI	1,144	1,160	727
WV	203	41	89
WY	45	40	29

Table A20: Counts of Patents by State and Dataset: 1940

	CUSP	HistPat	SAZ
AK	6	3	0
AL	196	93	90
AR	70	41	33
AZ	59	40	42
CA	2,455	2,493	2,181
CO	206	256	181
CT	1,160	1,085	967
DC	232	209	216
DE	303	359	684
FL	140	223	210
GA	126	128	1,931
HI	22	30	26
IA	290	308	262
ID	43	34	37
IL	3,178	3,676	3,541
IN	935	980	761
KS	181	191	167
KY	141	143	140
LA	154	156	282
MA	1,787	1,777	1,227
MD	370	383	346
ME	86	66	205
MI	2,231	2,281	1,809
MN	473	559	464
MO	496	734	666
MS	59	52	42
MT	57	53	0
NC	141	139	135
ND	28	26	25
NE	95	88	73
NH	97	87	72
NJ	2,678	2,617	2,299
NM	26	12	29
NV	31	15	19
NY	5,648	6,278	5,798
OH	2,748	3,038	2,429
OK	332	328	288
OR	165	164	163
PA	2,607	2,822	2,630
RI	245	253	0
SC	75	62	65
SD	16	38	36
TN	156	161	135
TX	763	750	679
UT	58	48	51
VA	178	267	0
VT	39	32	28
WA	281	299	264
WI	822	835	679
WV	136	36	83
WY	188	19	19

G Patents by U.S. and Foreign Inventors

The results in the previous section use only those patents for whom an inventor is residing in the U.S. But the CUSP, HistPat, and Jim Shaw data also contain inventors that file patents with the USPTO while residing in foreign countries. In Figures A12 and A13, I plot the number and share of total patents for which an inventor resides in the U.S. or a foreign country, respectively. The USPTO aggregate patenting statistics list separate counts of the number of patents issued to U.S. inventors in each year. I consider an inventor to be based in the U.S. if the inventor's country is listed as the U.S. or if the inventor's state of residence is one of the U.S. states, the District of Columbia, or if the state is missing but the country is listed as "United States".⁵ Because the HPDF and KPSS datasets do not record inventor information, I omit them from this analysis. Because the SAZ dataset is only designed to pull out names of U.S.-based inventors, I omit it from Figure A13

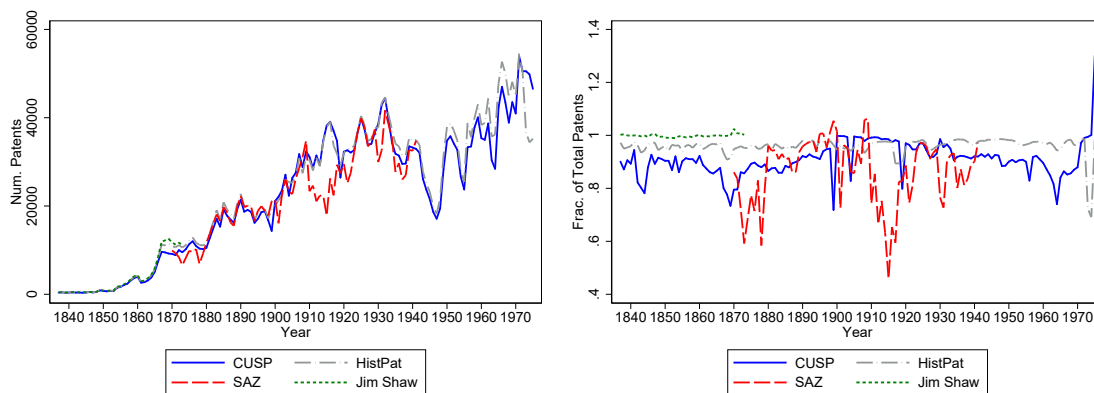
In general, the results in Figure A12 are similar to those in Figure 5, with the exception that when examining only the U.S.-based patents, the SAZ data perform better relative to CUSP and HistPat. This is to be expected, as both the HistPat and SAZ contain data only on U.S. inventors; SAZ was constructed to match inventor names to the U.S. decennial population censuses. In some years, however, the SAZ performs too well: in 1895, 1896, 1898, 1899, 1900, 1908, and 1909, SAZ contains more than 100% of patents to U.S.-resident inventors. This likely occurs for two reasons. First and most important, as noted above, patents from Germany are erroneously recorded as being from the state of Georgia. Second, the SAZ data may mis-classify utility patents as other types of patents (e.g., design or plant

⁵The number of patents issued to inventors living in Puerto Rico or other U.S. territories, especially during historical periods, are trivial. For inventors in the military, the Jim Shaw data lists the location as "United States Army" or "United States Navy." I record these inventors as U.S.-based as well.

patents) as utility patents.

Figure A13 shows that the CUSP data contains a larger number of foreign-based inventors than either the HistPat or Jim Shaw data in very year. To calculate the aggregate number of patents with a foreign inventor from the USPTO aggregate statistics, I subtract the number of patents with a U.S.-based inventor from the total number of issued patents each year. This undercounts the number of patents with a foreign inventor if foreign inventors collaborate with U.S. inventors on the same patent; I have noticed a number of these cases while manually inspecting patents. It is thus not surprising that the CUSP reports more than 100% of the aggregate number of foreign patents in panel (b). the Jim Shaw and HistPat data report close to 100%; recall that especially in early years these datasets usually contain the location for only one inventor, and thus are likely to give a number of foreign patents closer to how I constructed the USPTO's aggregate count of foreign patents.

Figure A12: U.S.-Based Patents

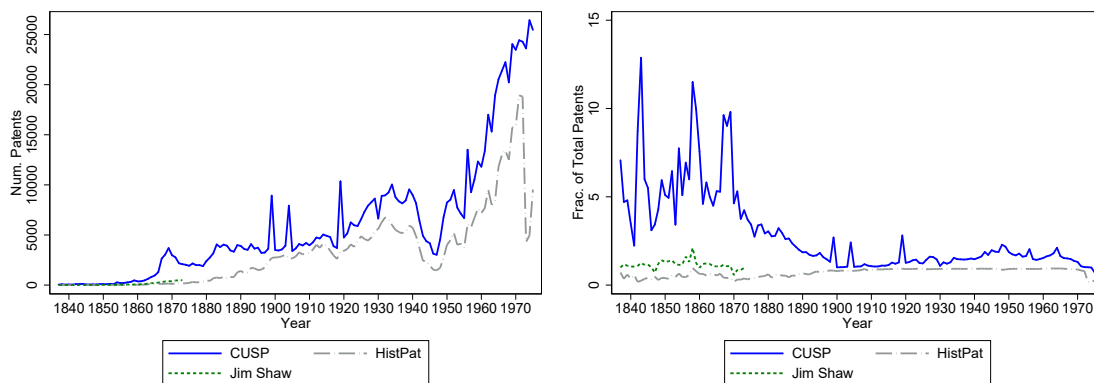


(a) Num. Patents to U.S. Inventors

(b) Frac. of All USPTO Patents to U.S. Inventors

Notes: Panel (a) plots the number of patents to an inventor residing in the U.S. Panel (b) plots the fraction of the total U.S.-based inventors from the USPTO aggregate patents that were successfully linked to a U.S. state each year using each patent dataset.

Figure A13: Foreign-Based Patents



(a) Num. Patents to Non-U.S. Inventors

(b) Frac. of All USPTO Patents to Non-U.S. Inventors

Notes: Panel (a) plots the number of patents to inventors not residing in the U.S. each year using each patent dataset. Panel (b) plots the fraction of the total foreign-based patents from the USPTO aggregate patents each year using each patent dataset.

H Additional Differences Between Datasets

Figure A14 replicates panel (a) of Figure 10 except I show when the issue years differ between the HPDF dataset and each of the CUSP, HistPat, and Jim Shaw datasets.

Digging into differences in the patent issue dates in more detail, I investigate whether it is possible to get a sense of which dataset is more likely to be correct when two datasets record different issue dates. One attempt to do this is see if one dataset is more likely to record patents “out of order.” The PTO assigns patent numbers sequentially by date of issuance. Thus, if a dataset contains a patent number with a later issue date than the patent that succeeds it, this indicates that at least one of those patents is out of order. I display these results in Table A21. The first row displays the number of instances in which a patent number has a later issue year than the patent that succeeds it. These are extremely rare, with only 26 cases in all of the CUSP and HistPat, three cases in the HPDF, and 29 cases in the Jim Shaw dataset. The second row displays the number of instances in which a patent

number has a later issue date than the succeeding patent, where the issue date uses the day, month, and year instead of just the year. Mistakes here are more common, even in the HPDF, although they are still quite rare relative to the overall number of patents in these datasets.

To ensure that the results in Figure 11 (showing the number of patents in different states in different datasets) are not driven by different ways of ordering inventors across datasets, in Appendix H, I plot similar results but looking for differences in the state of residence of the inventor residing in the first alphabetical state in each dataset. In this case, I find even more patents that list different states in different datasets, likely because such a way of counting accentuates failing to parse all inventors, as indicated in Figure 6.

In the body of the paper, I use states of residence as the “baseline” plots to show when different datasets because it is a relatively fine geographic area (at least, fine relative to, say, country of residence), yet state names are easier to clean than are county names. I want to ensure that differences in locations are not driven by choices in cleaning or harmonizing location names.⁶ Nevertheless, in many cases researchers will want to use patents aggregated to finer levels than states, and so it is informative to see how often the CUSP and HistPat datasets report different inventors’ counties of residence. I report these results in Figure A16, using the county of residence of the first listed inventor; results using the county of residence

⁶An additional challenge using county names is that county boundaries change over time, counties merge or split, etc. The CUSP data places patents in their current county (Berkes, 2018) (the HistPat data is less explicit on this point, but I believe they also assign patents to the patent’s current county). By “current county,” I mean the patent’s county at the time the CUSP dataset was built, circa 2017. This is in contrast to the SAZ and Jim Shaw data, which locate patents in their counties at the time the patent issued. Using the SAZ data, I have experimented with these issues and conclude that different ways of handling county boundaries affect a relatively small number of patents; these results are available upon request. Most patents are from large cities, which typically don’t change their counties (although there may be issues with large cities annexing neighboring areas). For more issues in determining a patent’s county, see Appendix E above.

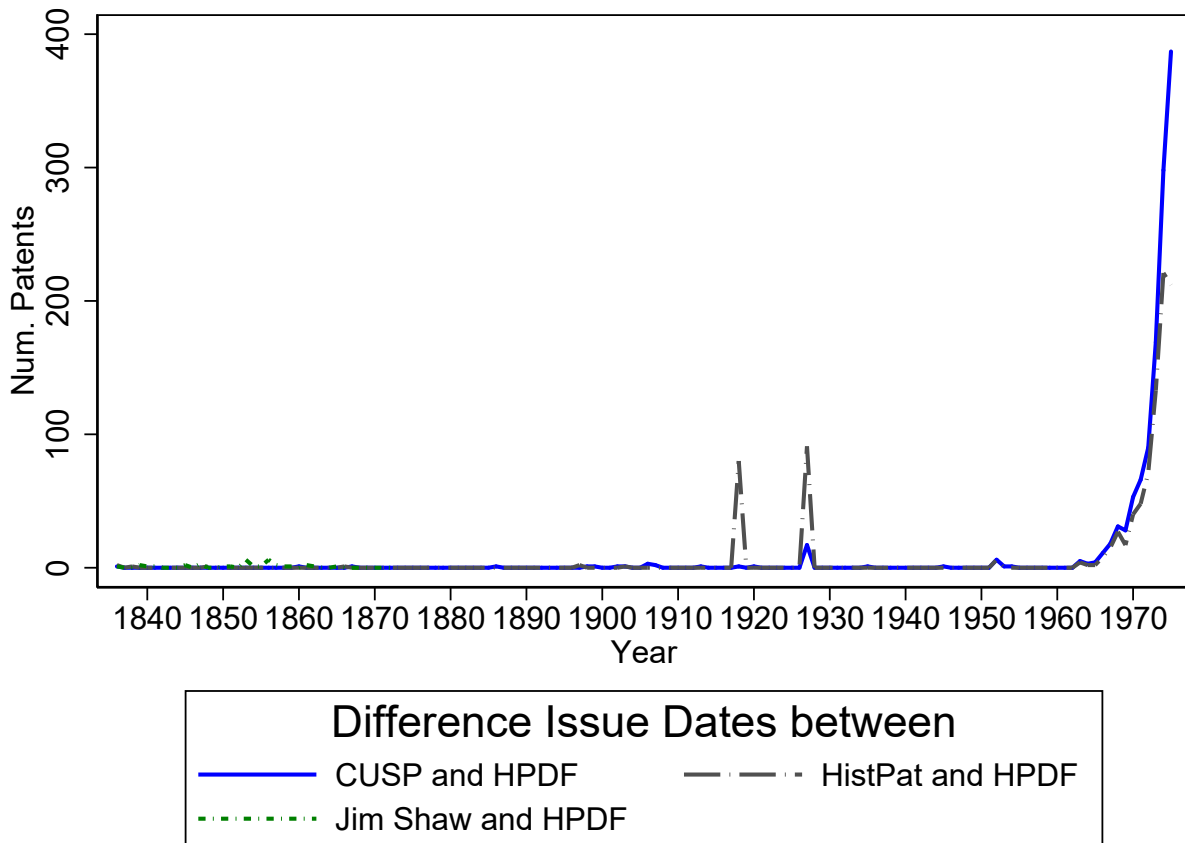
of the first alphabetic county are available upon request. Not surprisingly, counties of residence differ across datasets more often than do states.

Figure A17 conducts a similar exercise but examining a more aggregated measure of geography: country of residence of the first listed inventor. I plot the difference between the CUSP and HistPat, between the CUSP and Jim Shaw, and between the HistPat and Jim Shaw, since all three of these datasets contain patents from both U.S.- and foreign-based inventors. Prior to 1970, fewer than 5% of patents in the CUSP report a different country between the CUSP and HistPat for all but a couple of years (differences between the CUSP and Jim Shaw and HistPat and Jim Shaw are even smaller for all years in common). The 1970s correspond to a period when international patenting skyrockets, as well as a period when coverage in the HistPat becomes less accurate; either could explain the observed spike. For most years, country of residence is easier to identify than is state or county. The fact that 3-5% of patents still record different countries of residence for most years is yet additional suggestive evidence that mis-recording patent numbers may be driving discrepancies.

Figure A18 plots the number and fraction of patents numbers for which a different number of inventors is recorded between the CUSP and HistPat, between the CUSP and Jim Shaw, and between the HistPat and Jim Shaw. Here, the share of patent numbers with different information is higher than any of the geographic measures. This should not be surprising; recall from Section 3.3 that the HistPat, and perhaps to a lesser extent the Jim Shaw, data do not appear to do a good job of capturing inventors beyond the first inventor for many years. The fraction of patent numbers that disagree between the CUSP and HistPat increases after 1930; this is the same time that team size on patents begins to increase dramatically, leading to many more opportunities for error in correctly counting the number of inventors.

Finally, Figure A19 plots the number of fraction of patent numbers for which a different patent class or subclass is recorded between the CUSP and HPDF. For this exercise, I use USPC classifications. Reassuringly, the 3-digit USPC classes are almost never different between the CUSP and HPDF. USPC sub-classes, which correspond to a 6-digit patent class, disagree far more often, in 10-20% of CUSP patents in most years. I verify that these differences are not driven by cases in which a patent subclass is missing in either the CUSP or HPDF. This is surprising, since patent classification information is not reported on patent documents; instead the CUSP obtains this information from the USPTO (Berkes, 2018).

Figure A14: Patents that Have Different Issue Dates in HPDF and Other Datasets

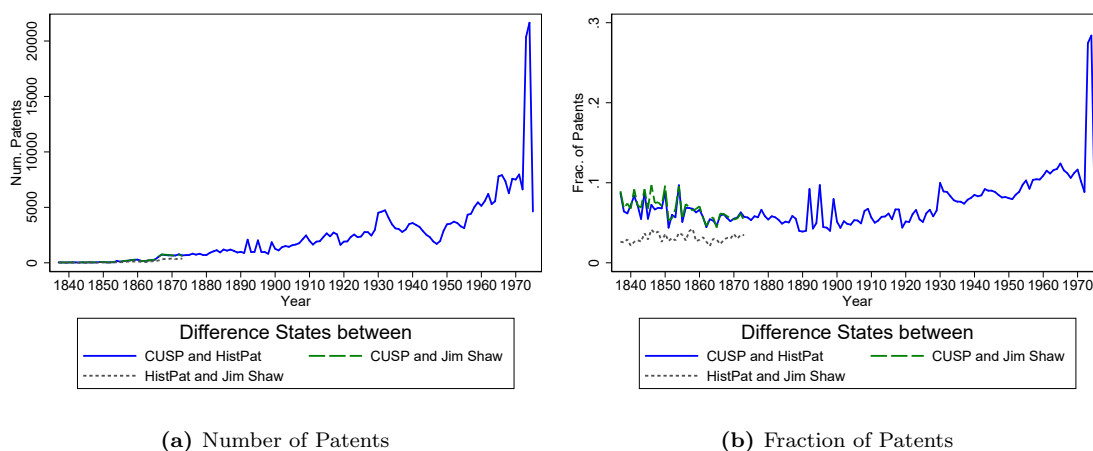


Notes: The number of patents that record a different issue year in HPDF and one other dataset.

Table A21: Out-of-Order Patents

	CUSP	HistPat	HPDF	Jim Shaw
Year	26	26	3	29
Full Date	1,905		1,735	1,119

Notes: The number of patents in each dataset that are “out-of-order.” A patent with a given patent number is out-of-order if the issue date for that patent number is later than the issue date of the next patent number. The first row uses the issue year to determine patent issue dates. The second row uses the issue day, month, and year to determine issue dates.

Figure A15: Patents that Have Different States of Residence (by First Alphabetical State)

(a) Number of Patents

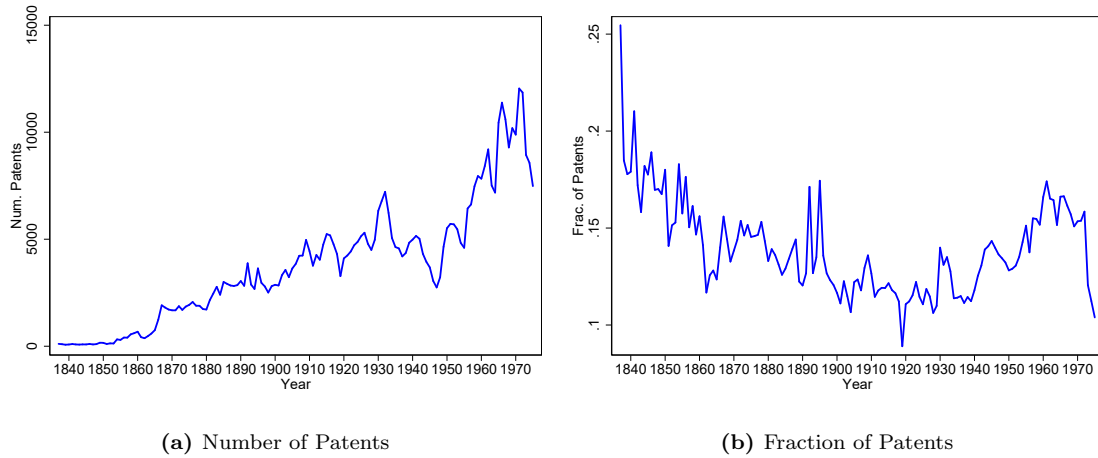
(b) Fraction of Patents

Notes: The number (panel a) and fraction (panel b) of patents for which the state of residence of the first listed inventor is different in one dataset compared to another. State of residence is calculated for the first inventor alphabetically by state name.

I Design and Plant Patents

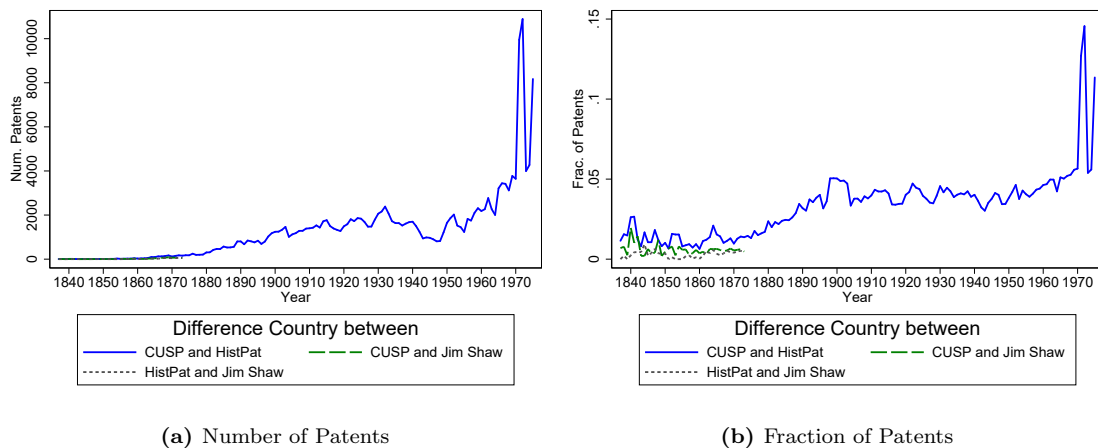
In the Annual Reports, patents of different types are recorded in different sections of the reports. To record patents of different types, Sarada et al. (2019) therefore search the Annual Reports for section headings or other mentions of distinct patent types (e.g., “Design Patents”, “Plant Patents”, “Reissues”, etc.). Years in which zero design patents are recorded typically mean that design patents were reported in a separate volume that was not OCREd

Figure A16: Patents that Have Different Counties of Residence



Notes: The number (panel a) and fraction (panel b) of patents for which the county of residence of the first listed inventor is different in one dataset compared to another.

Figure A17: Patents that Have Different Countries of Residence

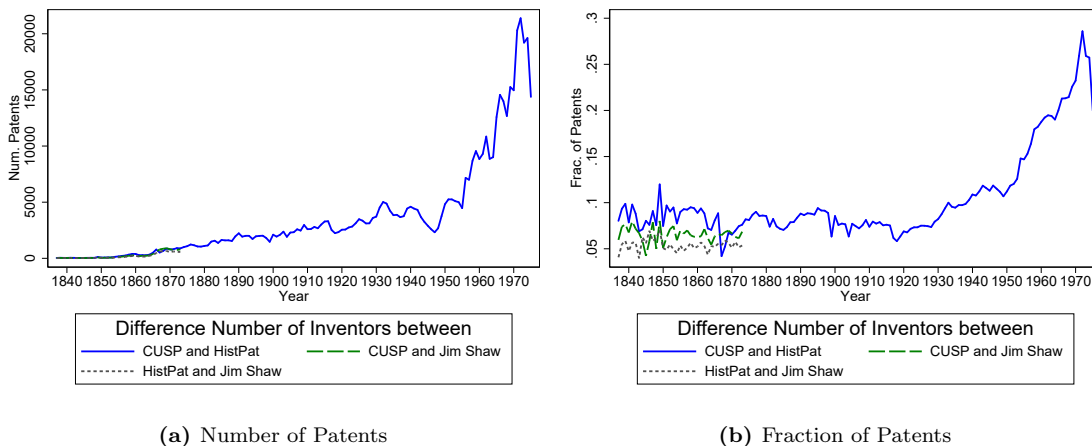


Notes: The number (panel a) and fraction (panel b) of patents for which the country of residence of the first listed inventor is different in one dataset compared to another.

and parsed. In the SAZ data, the type of invention to which a given record belongs is coded in the “invention_type” variable.

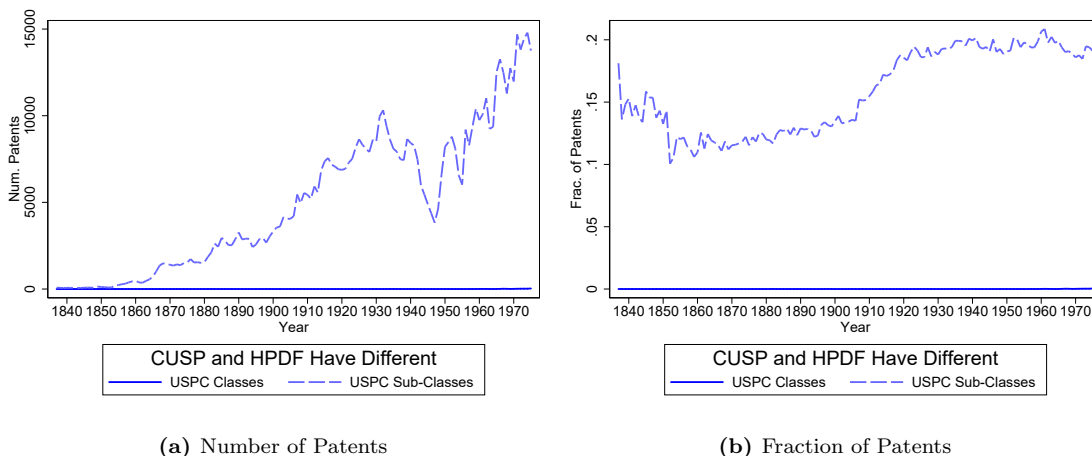
Surprisingly, the Jim Shaw data lists more design patents than are recorded in the USPTO aggregate data in 1854. This is likely due to either an error in the published USPTO aggregate data or, more likely a mis-coding of the issue year for some of the design patents

Figure A18: Patents that Have Different Numbers of Recorded Inventors Across Datasets



Notes: The number (panel a) and fraction (panel b) of patents for which the number of inventors listed is different in one dataset compared to another.

Figure A19: Patents that Have Different Classes and Sub-Classes Across Datasets



Notes: The number (panel a) and fraction (panel b) of patents for which the USPC class or sub-class is different in the CUSP relative to the HPDF, by issue year in the CUSP.

that actually issued in 1853. There are relatively few design patents in the early years, so a relatively small number of mis-coded issue years can lead to large swings in the fraction of issued design patents. In 1854, the Jim Shaw data record 11 more design patents than are listed in the USPTO aggregate data, while in 1853 they record 11 fewer.

Figure A20: Share of Design and Plant Patents in SAZ and Jim Shaw Data

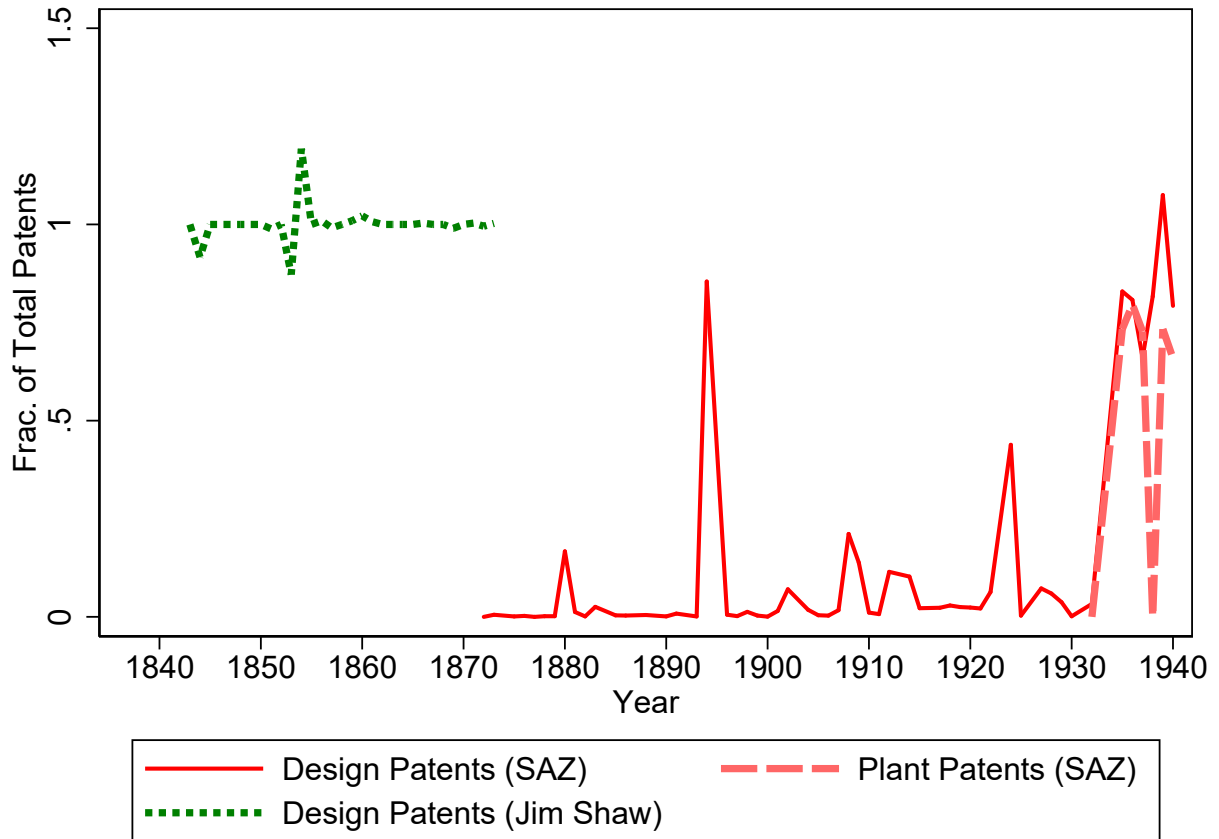


Figure A21: $\log(\text{Num. Patents})$

Notes: The share of design all design patents issued by the USPTO in each year in Jim Shaw data (green dotted line) and the SAZ data (red solid line) and the share of all plant patents issued by the USPTO in each year in the SAZ data (red dashed line). Data on aggregate numbers of design and plant patents granted each year by the USPTO are from https://www.uspto.gov/web/offices/ac/ido/oeip/taf/h_counts.htm.

References

- Andrews, M. J. (2020). *How do institutions of higher education affect local invention? Evidence from the establishment of U.S. colleges.* (Unpublished, NBER)
- Berkes, E. (2018). *Comprehensive universe of U.S. patents (CUSP): data and facts.* (Unpublished, Ohio State University)
- Billington, S. D., & Hanna, A. J. (2020). *That's classified! Inventing a new patent taxonomy.* (Unpublished, Ulster University)
- European Patent Office. (2015). *Data catalog: PATSTAT.* Ann Arbor, MI: Inter-university Consortium for Political and Social Research. (2015 spring edition, Version 5.03)
- Google Patents Team, T. (2016). *Your Google Patents question.* (electronic correspondence, February 16, 2016)
- Kang, B., & Tarasconi, G. (2016, September). PATSTAT revisited: suggestions for better usage. *World Patent Information*, 46, 56-63.
- Manson, S., Schroeder, J., Riper, D. V., & Ruggles, S. (2019). *IPUMS national historical geographic information system: version 13.0.* Minneapolis, MN: University of Minnesota. (<http://doi.org/10.18128/D050.V14.0>)
- Raider, R. (2016). *Patent data.* (electronic correspondence, October 21, 2016)
- Sarada, Andrews, M. J., & Ziebarth, N. L. (2019, October). Changes in the demographics of American inventors, 1870-1940. *Explorations in Economic History*, 74.