Creating the China Patent Dataset

Josh Lerner Namrata Narain Dimitris Papanikolaou
Amit Seru Zunda Winston Xu*

July 5, 2025

Preliminary and Incomplete

Abstract

This paper describes the construction of a new data set of primary Chinese patents. The effort seeks to use the techniques that have been developed over the last 40 years to analyze U.S. patents, while grappling with the unique institutional features of the Chinese intellectual property system. Given the rapid growth of patenting and venture capital in China and the intense interest in innovation in China more generally, it is our hope that this dataset will stimulate work in this area by numerous economists.

*Harvard University; Boston University; Northwestern University; Stanford University; and Stanford University. Lerner, Papanikolaou, and Seru are affiliates of the National Bureau of Economic Research. We thank the Harvard Business School Division of Research, the Hoover Institution, the Private Capital Research Institute, the Smith Richardson Foundation (Award no. 2024-3365), the Stanford Graduate School of Business, the Toulouse Network on Information Technology, and the U.S. National Science Foundation (Award no. 2241237) for financial support. We thank Lee Branstetter, Alex Caracuzzo, Patrick Clapp, Paul di Bello, Erica Fuchs, Hannah Kelsen, Jacques Lawarrée, Kathleen Ryan, Sarah Sewall, Lisa Simon, Allan Song, Andrew Toole, Vesna Vajsbaher, and Ben Zweig for their guidance in the identifying data, financial, and computing resources. Numerous research assistants helped in this effort. Seminar participants at Harvard Business School and a number of practitioners provided helpful comments. Lerner has received compensation for advising limited partners in venture funds, venture capital groups, and governments designing policies relevant to venture capital. All errors and omissions are our own.

1 Introduction

Innovation outside the major industrialized nations has become increasingly important in recent years. This is particularly true in China, which is illustrated in two charts. Figure 1, using data from the World Intellectual Property Organization, reports the annual volume of domestic patent applications in national patent offices between 2001 and 2022. The figure makes clear the substantial disparity in the volume of patenting across nations. It also highlights that China and India largely drove the recent acceleration in aggregate worldwide patenting. The change in annual domestic applications from 2008 to 2022 are 653% and 500% in these nations respectively. Of the remaining eight countries depicted, the largest corresponding increases are 45% (Korea) and 9% (US). The other six nations essentially show no growth or a decline over the period.

Using data from (Lerner et al., 2024), Figure 2 looks at the relationship between the GDP per capita and the aggregate venture investment in each country-year across a number of major nations, over as long a time period as the data permit. In 2001, the US represented the location of 88% of global venture dollars invested, and other developed countries most of the remainder (7%). By 2019, global venture activities became bipolar: while the US continued to lead with 42% of global investment, China had surged to account for 38% of the total (65% of the non-US portion).

Motivated by this rapid rise in innovation in China, we construct a Chinese patent dataset using the methodology pursued with the U.S. patent data since the inception of the first patent project at National Bureau of Economic Research (Griliches, 1984). We emulate many of the key steps undertaken in the literature using the U.S. Patent and Trademark Office (USPTO) data, such as the compilation of citations (e.g., Jaffe and Trajtenberg (2002)) and the identification of their sources (Alcácer and Gittelman, 2006), the calculation of measures of textual novelty (Kalyani et al. (ming); Kelly et al. (2021)), the construction of international patent families (Putnam, 1997), the disambiguation of assignees (e.g., Jaffe and Trajtenberg (2002); Magerman et al. (2006), and the identification of venture-backed patents (Bernstein et al., 2016).

We examine patents from the inception of patent grants in September 1985 through the time that the data were downloaded in September 2023.[1] Following the conventions of the economics of innovation literature (e.g., Jaffe and Trajtenberg (2002)), we focus exclusively on the most important patent class, invention patents. We do not consider design or second-tier patents (confusingly referred to in China as utility patents, the U.S. term for the most important class of awards).

This document describes the process by which we constructed this dataset, the key issues

[1]The modern Chinese patent law was enacted at the 4th Meeting of the Standing Committee of the Sixth National People's Congress on March 12, 1984 (`https://english.cnipa.gov.cn/col/col3068/index.html`). The founders of the People's Republic had envisioned a patent system more than three decades earlier, but the earlier system was quickly abandoned due to the perceived incompatibility of patents and socialism (Cheng, 2023).

encountered, and the choices made. The institutional features of the Chinese patent system, as well as the ways that these data are archived, introduce some unique challenges. Among these are: the absence of a comprehensive, readily accessible official database of Chinese patents, the frequency with which Chinese patent filings are never examined, reflecting the failure of the applicant to pay the fee to trigger the examination, the inclusion in Chinese patent datasets of multiple publications of the same patent, and the tendency of subsequent Chinese patents to cite the first publication of a patent application, rather than the final award.

While there has been a growing literature about innovation in China, efforts to catalog the Chinese patent data along the lines of the various efforts at the NBER and elsewhere have been infrequent.[2] The largest effort to disambiguate Chinese patent assignees that we are aware of, He et al. (2018), matched 191,325 SIPO patents matched to listed firms in China from 1990 to 2010. These efforts are likely to get progressively more difficult as Western scholars face increasingly limited access to major Chinese databases[3] and Chinese scholars must contend with vague but onerous regulations regarding the transfer of data about China.[4]

As part of this project, we are distributing four datasets for public use that include bibliographic information on patents and details on the assignees and inventors linked to each patent. Appendix A provides codebooks for these datasets.

The rest of the document is organized as follows. We describe institutional features of the Chinese patent office in Section 2. In Section 3, we describe how we assemble the universe of Chinese patents by combining information from the China National Intellectual Property Administration (CNIPA), European Patent Office (EPO), and Google Patent databases. Section 4 summarizes the process of cleaning and sorting assignee and inventor data. In Section 5, we describe the construction of measures of patent quality. In Section 7, we present some key facts about the dataset.

[2]Influential work on Chinese innovation include Holmes et al. (2015); Aghion et al. (2015); Fang et al. (2017); Wei et al. (2017); Chen et al. (2021); König et al. (2022); Beraja et al. (2023).

[3]For instance, Western access to the heavily used CNKI database was cut off in March 2023 (`https://www.scmp.com/news/china/article/3214808/portal-china-closing-least-temporarily-and-researchers-are-nervous`).

[4]In particular, China's Data Security Law (中华人民共和国数据安全法; translated into English at `https://digichina.stanford.edu/work/translation-data-security-law-of-the-peoples-republic-of-china/` enacted in 2021 and implemented in large part in 2023, seeks to regulate data processing activities by organizations and individuals within and outside of China. A particular concern of the legislation is limiting the export of "National Care Data," defined as "important[t] in economic and social development," which poses "danger to national security, public interests, or the lawful rights and interests of individuals or organizations" if misused.

2 Institutional Details

Before assembling the patent dataset, it is worth reviewing some of the unique institutional details of the Chinese patent system.

One way that the Chinese data differs from that in the U.S. is due to the mixture of publication types. Chinese patent databases typically combine all patent publications. All Chinese patent applications are published at 18 months, unlike in the U.S. where some subsets of awards (e.g., those only filed in the U.S.) can remain unpublished until issue. These published applications are referred to as A-type publications (公开or 审定公布). From 1993 to April 2000, there was a second publication of examined-but-not-yet issued patents, referred to as B-type publications (审定公告). Publications of issued patents are referred to as C-type publications (through 1993) and B-type ones thereafter (both referred to as 授权公告).[5]

Another institutional difference is that the applications are not automatically reviewed. Rather, CNIPA applicants must request that their patent be reviewed within the first 36 months after publication, or it is deemed to have been withdrawn (Tong et al., 2018). In particular, the fee to have an invention patent application examined, 2500 RMB, is several times that to file the fee for the original application (900 RMB).[6] Moreover, many of the "patent quota" incentive schemes administered by local governments and corporations appear to focus on patent applications rather than awards (see Prud'homme (2012); Fuller (2016)). Perhaps reflecting this, many more patent publications in China are applications that are never issued.

These institutional considerations suggest the desirability of taking a somewhat different approach to constructing a patent dataset in China. While analyses of USPTO data typically focus exclusively on issued patents when computing patent counts, citation scores, and textual analyses, here it is hard to draw such a bright line between applications and awards for the Chinese data. Throughout the analysis below, we will focus on employing all final CNIPA publications of patent applications, even in cases where the patent was never issued.

Each Chinese patent is characterized by the following five pieces of information: application number, publication number, publication type, filing date, and grant date.

A patent's application number is often situated at the top of the patent document and is the first number assigned to it. Patents with the same application number contain the same content. The filing date of the patent is when the patent application is filed, and the application number is generated. When a patent is granted, we observe the date it is granted on, which is its grant date.

Each patent is ultimately identified by a grant ID, which is the combination of its

[5]These schemes are summarized in https://www.epo.org/en/searching-for-patents/helpful-resources/asian/china/numbering/.

[6]`https://english.cnipa.gov.cn/col/col3000/index.html`.

publication number and publication kind. Take the example of a patent identified as CN200910125311A. We call this a patent's "grant ID" and it is made up of three components. The first component is the first two digits, "CN", which is a prefix for patents filed in China. All patents in our dataset have the "CN" prefix. The center numerical portion of this ID is the publication number. The last letter of the ID is the patent's publication kind. The publication kind of a patent signifies whether it is an application or a granted patent. Across all years, patent applications are identified by the publication kind "A." As noted above, granted patents, may end in either "B" or "C" based on their year of issue.

One peculiarity of the Chinese system is that each application may be granted multiple times or even under different publication numbers on the same day. We provide three examples that illustrate the difficulty in constructing this dataset. First, applications and granted documents before April 2010 were published with different numbers.[7] As an example, the patent application CN87101693A was filed on 1987-11-04 and it was granted on 1990-03-21 as CN1007299B. Second, the publication number itself can be shared by multiple patents. For example, the publication number 1102613 corresponds to two patents, one with application number 94104087 and publication kind A and another with application number 98108488 and publication kind C. Finally, one patent application may have two corresponding published documents. For example, the patent application CN200910125311A was filed on 2009-12-31 by the Naval University of Engineering (People's Liberation Army). This patent has two granted documents with the same date, 2012-04-18, under two different grant IDs, CN111903222B and CN114303470B. They have the same title and other content (as seen on patents.google.com.) This same patent will therefore correspond to three different publication numbers in our dataset: 200910125311, 111903222, and 114303470.[8]

In the next section, we grapple with these issues as we combine patents from three sources: the CNIPA, the EPO, and Google.

3 Assembling the Universe of Chinese Patents

The initial task is to identify the universe of Chinese patents. Unlike in the case of the U.S., where PatentsView (an open data platform supported by the Office of the Chief Economist at the U.S. Patent and Trademark Office) presents an authoritative single source of patent data, here the situation is somewhat more complex.

[7] *Ibid.*

[8] Historical changes in Chinese patents dataset, including changes in the formatting of IDs, is discussed at the following URL: `https://www.epo.org/en/searching-for-patents/helpful-resources/asian/china/numbering`.

3.1 Description of the Datasets

We focus our initial efforts on three datasets, the EPO's PATSTAT Global dataset, the CNIPA database, and the Google Patents database.[9]

3.1.1 PATSTAT

PATSTAT is a researcher-accessible version of the DOCDB bibliographic data. DOCDB is the EPO's master bibliographic database, used in patent examinations. PATSTAT has been used extensively by researchers (a search of Google Scholar in February 2024 reveals over 6600 references to the database). Furthermore, unlike the commercial vendors, PATSTAT gives broad rights for researchers to use and redistribute modified versions of their database.[10]

At the same time, PATSTAT has several limitations for our purposes. First, EPO officials acknowledge that they are not certain of the coverage of Chinese patents in PATSTAT provided to EPO by CNIPA. There has been no prior effort to audit the completeness of the Chinese patent data in PATSTAT of which we are aware. Second, while EPO has the full-text versions of the awards, they only allowed us to download the "front-page" information due to stated concerns about bandwidth. Third, there are substantial lags in uploading data to this database: in the version we used (accessed in June 2023), the last Chinese patent award was on December 7, 2021; the last application on November 11, 2021.

We first assemble the EPO dataset. The data from EPO is spread across multiple smaller datasets, which we combine as follows. We start with the dataset containing the universe of Chinese patent IDs. This file contains the ID numbers and dates associated with each patent in its dataset. Second, we use information on the application kind of each patent to select only invention patents. Third, we attach each patent ID to its publication kind. This signifies whether a row in the ID dataset is an application, or a granted patent.

We now briefly describe each of these datasets in turn.

EPO IDs

- Filename: all_cn_ids.csv

[9] We eschewed the use of commercial datasets for this project, both to ensure the ability to redistribute the results and because in many cases, it is difficult to obtain clear answers about database coverage and biases form these firms. For instance, many firms appear to download and resell the European Patent Office data, with no real acknowledgements of its limitations.

[10] The key relevant terms in their document "Terms and conditions for the licensing of EPO databases" (`https://www.epo.org/en/service-support/ordering/raw-data-terms-and-conditions`; as of February 2024) were are in clauses 5.1 ("Upon conclusion of the contract the licensee receives a non-exclusive, non-transferable, worldwide license to use the selected EPO database.") and 5.2 ("The licensee may use the EPO database for his own internal business purposes or to create his own product, this being defined as the licensee's own machine-readable database, publication, service or other product which contains or is made on the basis of data from the EPO database.")

- Variables: APPLN_ID, PAT_PUBLN_ID, PUBLN_NR, PUBLN_AUTH, PUBLN_DATE, APPLN_AUTH, APPLN_NR, APPLN_FILING_DATE, DOCDB_FAMILY_ID.[11]

This file provides the key identifiers needed to construct the database. Note that APPLN_ID and APPLN_NR are two different variables. APPLN_ID is created by the DOCDB, whereas APPLN_NR is created by the patent authority where the application was filed. We use the APPLN_ID variable to match EPO's various datasets when possible and use the APPLN_NR (alongside other information) to match the EPO dataset to the CNIPA and Google datasets that we describe more fully below.

From this file, we also pull each patent's DOCDB family identifier, which identifies the cases where an identical (or very similar) patent was filed in different patent offices. These families are determined in two ways: in some cases, patent applications in different offices may be explicitly linked through an application under the Patent Cooperation Treaty administered by the World Intellectual Property Office; in other cases, they are subjectively determined by the EPO examiner as part of the review process. This identifier linking families has been carried over from the DOCDB database used by the EPO examiners.

In this file, we have 33,318,707 rows. These rows correspond to 28,823,618 unique values of PUBLN_NR and 28,544,853 unique values of APPLN_NR. This dataset has 28,544,872 unique values of the APPLN_ID variable.

PAT_PUBLN_ID uniquely identifies each row. Neither PUBLN_NR nor APPLN_NR uniquely identify patents in this dataset. Rather, we use a combination of ID numbers and dates to identify the same patent. Section 3.C describes our procedure in detail.

Application type.

- Filename (example): appln_kind / cn_1990_09_kind.csv. We have one file per month between September 1985 and December 2021.
- Variables: APPLN_ID, APPLN_KIND.

We use the APPLN_ID variable to link the all_cn_ids.csv file to cn_****_**_kind.csv files. Application kind identifies whether a patent is a utility, design, or invention patent. We use the *_kind.csv data to restrict the applications to those whose kind is A, which correspond to Chinese invention patents, as opposed to design and utility model awards. We have 13,736,967 unique application IDs for invention patents.

When a patent is missing its publication kind, we tag the publication kind as "MISS" and determine it in conjunction with data from CNIPA and Google.

[11] All definitions of variables are in the EPO documentation posed at "Data Catalog: PATSTAT Global," Spring 2023, `https://link.epo.org/web/searching-for-patents/business/patstat/data-catalog-patsat-global-spring-en.pdf`, and are not repeated here.

Publication kind.

- Filename (example): cn_abstracts / combined_final / cn_abstract_2012.csv. We have one file for all abstracts for each year.
- Variables: PUBLN_KIND, APPLN_ID, APPLN_AUTH, APPLN_NR, APPLN_KIND, APPLN_FILING_YEAR, APPLN_FILING_DATE, GRANTED, DOCDB_FAMILY_ID, PAT_PUBLN_ID, PUBLN_DATE, PUBLN_NR, PUBLN_AUTH, PUB_YEAR, APPLN_ABSTRACT_LG, APPLN_ABSTRACT, APPLN_ABSTRACT_Translated.

This file contains information on the patent's publication kind (PUBLN_KIND) but is missing this information for several rows in the all_cn_ids.csv file. While the full dataset (design + utility + invention patents) has 28,544,872 application IDs, we only have publication kind information for 22,265,806 patents. We merge this dataset to the data on invention patents if there is a match on APPLN_ID x PUBLN_NR x PUBLN_DATE triple. (These three variables together identify unique rows in the abstract dataset.) Of the 13,736,967 invention patent applications, we have publication kind information for 11,150,901 applications.

The accounting for our set of invention patents is as follows. We have a total of 18,509,330 rows corresponding to 14,026,718 unique values of PUBLN_NR, and 13,736,967 unique values of APPLN_NR.

3.1.2 CNIPA and Google Patents

The CNIPA dataset contains a catalog of seemingly all patents issued by CNIPA. It is only accessible from within China using a national identity number. We restrict the downloads to invention patents through September 2023. Multiple teams of research assistants pulled these records.

We manually searched and downloaded patent data from CNIPA's official website (`https://pss-system.cponline.cnipa.gov.cn`) in the following steps:

- We filter by patent type to retain only the "invention" patents. (We exclude the "utility" and "design" patents. Note that under China's patenting system, the "invention" type is equivalent to the 'utility" type in the US, and the 'utility" and "design" types are equivalent to the "design" type in the U.S.) and select "China" as the country in which the patent is issued.
- We type in "CN" as a required keyword for patent ID.
- We set the range of issue dates for the intended years (1985-2023).

In total, we accessed patents with 16,138,878 unique grant IDs and 16,138,655 unique application IDs. For CNIPA data, we receive application numbers with a decimal digit and a check digit after the decimal. We identify the patent application number as the numerical portion before the decimal point.[12]

3.1.3 Full-Text Data

We obtained the full text of patents in Mandarin and English. The Mandarin ones were from CNIPA, which we have for all but 519,490 patents. Because of the difficulty of downloading the remaining patents from CNIPA (we would have had to do these on a case-by-case basis), in these cases, we scraped their Chinese text from Google Patents instead. We did the following sanity check to check the accuracy of the Chinese text we obtained from Google Patents. We randomly sampled around 1000 out of those 519490 patents, cross-checked those patents' Chinese text from CNIPA with those from Google Patents, and found that all of them are consistent.

For the English text, we obtained from Google the full text translated into English for all patents.

Google makes clear that these data are accessible to researchers and can be redistributed. On its website (`https://console.cloud.google.com/marketplace/product/google_patents_public_datasets/google-patents-research-data?pli=1`; accessed on February 2024), it states "Google Patents Research Data contains the output of much of the data analysis work used in Google Patents (patents.google.com), including machine translations of titles and abstracts from Google Translate, embedding vectors, extracted top terms, similar documents, and forward references.... 'Google Patents Research Data' by Google, based on data provided by IFI CLAIMS Patent Services, is licensed under a Creative Commons Attribution 4.0 International License." The language for the Google Patents Public Data page indicates that identical licensing rules hold.[13]

Going forward, we discuss CNIPA and Google datasets together, and refer to them jointly as the CNIPA dataset.

3.2 Reconciling IDs Across Datasets

We want to combine information from the EPO and CNIPA datasets and keep one version of each patent in the final dataset. In addition to the duplications described in Section 2,

[12]Further information also available at `https://www.lexology.com/library/detail.aspx?g=3a2036a1-9b59-40b8-abf0-551823bb6d65`.

[13]The CC 4.0 license explains 11You are free to: 1. Share — copy and redistribute the material in any medium or format for any purpose, even commercially. 2. Adapt — remix, transform, and build upon the material for any purpose, even commercially. 3. The licensor cannot revoke these freedoms as long as you follow the license terms." Source: `https://creativecommons.org/licenses/by/4.0/deed.en`.

there are a few additional obstacles to achieving this data merger.

Foremost among these are inaccuracies in both databases. First, there are typos or incorrect values in the application number variable. Each patent is linked to an application number. While it seems to be easy to then keep the patent version corresponding to the latest grant date amongst all patents corresponding to an application id, the CNIPA data sometimes has typos or incorrect application numbers.

Consider, for instance, the patent with the grant ID CN1142360A. The application number in the EPO dataset corresponding to this patent is 96106116 (which also corresponds to the application number for this patent on Google.) The application number in the CNIPA dataset is 96102914. Besides the difference in application numbers, all other information between these patents is the same including publication number, filing date, and grant date. Similarly, for the patent number CN1216087A, CNIPA records the application number to be 199898800082, which is incorrect. The correct application number is 98800082 (i.e., the mismatch comes from an erroneous additional of the year of filing at the beginning of the patent.) A third example is the patent CN1665832A. The same patent appears twice in the CNIPA data corresponding to the following different application IDs: CN03815950.3 and CN038159503A. These are the same application IDs, but they were scraped or collected differently.[14] Another example of such a patent is CN1138374A.

This application number problem is not confined to the CNIPA data. For a few dozen patents, the EPO dataset has the incorrect filing date for applications. For example, EPO notes that the patent CN1589302A was filed on 2002-10-16 whereas the date in the CNIPA dataset and on Google is 2002-10-18. Similarly for patent CN1031109A. In total, there are nine such problematic pairs.

Having highlighted these difficulties, we now describe our procedure to keep unique versions of each patent. Intuitively, we want to group all patents with the same content or application together, so that we can then choose the latest granted document for each application. There will be two outputs in this process: one which keeps just one patent for each application (which we will use in the bulk of the analyses), and another that retains the final mapping between all IDs in our dataset and the corresponding "final" version that we kept. This will be needed later when calculating the citations to each patent, since we will need to keep track of citations to all versions of each patent. We are also releasing this "map" dataset publicly to allow researchers with access to other versions of the data to be able to see which version we keep in the final dataset.

Once we have identified all patents that belong to the same group (or application), we ensured that the patent version we keep has the full set of information regarding the application from both the EPO and CNIPA datasets. Each patent was associated with the following: grant id, grant date, publication number, application number, filing date, EPO application ID, and the DOCDC family ID.

[14]We cannot simply extrapolate from this example to eliminate the last two characters of all application numbers since some application numbers show up with an "A" at the end and no extra digit before it.

At a high level, our procedure has three steps. First, we drop exact duplicates, measuring duplicates in a few different ways to allow for typos. Second, we assign correct application numbers when problematic and repeat the procedure, i.e., drop exact duplicates. Finally, we only keep the latest publication corresponding to each application.

To summarize the two datasets:

- EPO dataset
  - Unique grant IDs: 17,822,717 (PUBLN_NR + PUBLICATION_KIND)
  - Unique publication numbers: 14,026,718
  - Unique application numbers: 13,736,967
  - Total rows: 18,509,330 (this is because, as mentioned above, several EPO patents are missing a publication kind, and therefore this dataset is not unique at the grant ID level.)
- CNIPA/Google dataset:
  - Unique grant IDs: 16,138,878
  - Unique publication numbers: 16,138,878
  - Unique application numbers: 16,138,655

We then append both datasets together. The total number of rows from both sources is 35,541,658, corresponding to 24,229,146 unique grant IDs.

We then do a series of efforts to de-duplicate the sample. We:

- Tag patents that are duplicates on publication number x application number x publication kind x grant date quadruples:
  - We drop 10,507,143 patents that have an exact duplicate.
- Tag patents that are duplicates in publication number x application number x grant date triples:
  - We drop 2,533,698 patents.
  - Note that if two patents were exact duplicates on a publication number x application number x grant date triple, then one came from EPO and another from CNIPA. We drop the EPO version of the patent.
- Tag patents that are duplicates in a publication number x grant date x filing date triple:

  - We first ensure that patents that have duplicate information in a publication number x grant date x filing date triple share the same application number. When patents that share a publication number x grant date x filing date triple differ in their application number, we assign them the "correct" application number. In the case of such duplicates, when one patent belongs to CNIPA and the other to EPO, we associate them both with the application number from the EPO. We then drop exact duplicates on these variables.
  - We drop 220,859 patents.

- Tag patents that are duplicates in a publication number x grant date x publication kind triple:
  - We ensure these applications have the same application number. These rows were not caught in the previous step because their filing date was incorrect (in the EPO data). As noted above, we drop nine patents.

- Tag patents that have the same application number but one of the grant ids is missing publication kind.
  - Drop 11 patents.

The eliminations so far focused on excising duplicate publications, either between the EPO and CNIPA data or else within each dataset. The final task to do is to keep only the last patent (by grant date) associated with each application ID: in other words, to make sure that we do not have a published application and grant of the same award. To do so, we drop all but the patents with the most recent associated date. This leads to the dropping of 5,839,889 awards, leaving 16,440,069 publications, each with a unique grant ID. Of these, 10,607,862 have a single version per application. Of the others, the breakdown is as follows:

- 2 versions per application: 5,824,482
- 3 versions per application: 7,702
- 4 versions per application: 5
- 5 versions per application: 1 (this is the application number 201080010943, which was re-issued multiple times).

Two final notes should be made. First, there are patents that were published multiple times on the same day. An example is the application number 200910125308, which was published as both CN114303471B and CN111903224B on the same day (2013-03-13). We keep just the smallest publication number arbitrarily. Similarly for the application number 200910125311. Both CN111903222B and CN114303470B were published on the same date, 2012-04-18. There are four such cases (corresponding to 8 patents), and we arbitrarily keep the "smaller" of the two patents based on their publication number.

Second, our procedure in two cases leads to the wrong number being recorded. First, patent CN1767414A shows up in the EPO dataset with the incorrect application number (200410052026) and in the CNIPA dataset with the correct application number (200510099179). Due to the protocol followed above (which prioritizes the EPO's application number), we keep the CNIPA application number for this patent. Second, patent CN1589302A shows up in the final dataset with the wrong application number (an extra digit.) This is because we update the application number in the CNIPA data only when rows are duplicated in a publication number x filing date x grant date triple. For this patent, the corresponding EPO patent has the wrong filing date. It is noted as 2002-10-18, whereas the date in the CNIPA dataset and in Google is 2002-10-16.

The final number of patents in the dataset is 16,440,045.

4 Identifying and Disambiguating Assignees and Inventors

For each patent, we sought to gather the following information: (a) The cleaned and disambiguated name of the entity to which it is assigned ("assignee"), (b) The sector of the assignee: company, university, hospital, non-profit, individual, and various combinations thereof, (c) the assignee's country, (d) Other characteristics of the assignee: whether they are venture-capital-backed firms, state-owned enterprises (SOE), People's Liberation Army, and "Seven Sons of Defense," (e) The name of the patent's inventors.

The rest of this section provides details on how we compile this information.

4.1 Reconciling Assignee Information Across Datasets

We received data on assignees and inventors from both EPO and CNIPA datasets. As a first step to constructing the patent-assignee-inventor-level data, we reconcile information on assignees and inventors between the EPO and the CNIPA datasets. We first describe each dataset and then the procedure of how we combine them.

Assignee and inventor information from the EPO is constructed using three sets of files. The first set contains the map between patent IDs (PAT_PUBLN_ID) and person IDs (that identify assignees.) The second set contains information on the assignees themselves, including their name, sector (e.g., corporation, university, hospital, ...), and the country in which they are based. The determination of an assignee's sector is made independently by the ECOOM Centre for Research and Development Monitoring at KU Leuven and added by the EPO subsequently to the database.

We put together the following datasets:

- Filename 1: patent_data_1985_2000 / cn_1990_09_pub_per.csv [one file per year-month]
  - Variables: APPLN_ID, PERSON_ID, APPLT_SEQ_NR, INVT_SEQ_NR
- Filename 2: patent_data_1985_2000 / cn_1990_09_assignees.csv [one file per year-month]
  - Variables: PERSON_ID, PERSON_NAME, DOC_STD_NAME, DOC_STD_NAME_ID, PERSON_CTRY_CODE, HAN_NAME, HAN_ID
- Filename 3: patent_data_1985_2000 / cn_1990_09_inventors.csv [one file per year-month]
  - Variables: PERSON_ID, PERSON_NAME, DOC_STD_NAME, DOC_STD_NAME_ID, PERSON_CTRY_CODE, HAN_NAME, HAN_ID

The _pub_per datasets provide crosswalks between patent IDs and person IDs and include one row for each patent's inventor as well as assignee (where both inventors and assignees are allocated a unique PERSON_ID.) We link the _inventor and _assignee datasets to _per_per datasets using PERSON_ID. To link the _assignee dataset to the _pub_per crosswalk, we only keep rows where APPLT_SEQ_NR > 0, limiting the dataset to rows with information on assignees themselves. To link the _inventor dataset, we only keep rows where INVT_SEQ_NR > 0, limiting the crosswalk to rows linking each PAT_PUBLN_ID to its set of inventors. We then merge the _inventor and _assignee datasets to each other on the PAT_PUBLN_ID variable, which leads to all possible pairs of assignees and inventors linked to each PAT_PUBLN_ID. The final merged dataset is unique at the PAT_PUBLN_ID X PERSON_ID_INVENTOR X PERSON_ID_ASSIGNEE level.

The EPO inventor-assignee dataset has three versions of names for both assignees and inventors: PERSON_NAME, DOC_STD_NAME, and HAN_NAME. We use HAN_NAME, where HAN stands for Harmonized Applicant Name, and is generated after disambiguation effects by the OECD HAN project of the OECD.[15] The final EPO dataset has data on inventors and assignees for 14,785,273 grantids (corresponding to 13,336,365 final grantids).

We construct a corresponding assignee-inventor level dataset for CNIPA as follows. For several patents, the inventor or assignee field had several values separated by a semi-colon. We create a patent-inventor-assignee level dataset by splitting each row into semi-colons, which leads to each row of the dataset corresponding to a (patent ID, inventor, and assignee) tuple. There is only one raw version of all names provided by CNIPA, which is what we use here. The final CNIPA dataset has data on 16,811,460 grantids (corresponding to 16,413,491 final grantids)

Having collected data on assignees and inventors from both EPO and CNIPA/Google, we combine them. We use the map constructed in the previous section that links all IDs in the EPO and CNIPA/Google datasets to the ID that is included in the final dataset. This allows us to see the inventors/ assignees linked to all versions of the same patent. This final

[15]More information on HAN IDs can be found on the WIPO website: https://www.wipo.int/edocs/mdocs/cws/en/cws_wk_ge_16/cws_wk_ge_16_oecd.pdf.

dataset has inventor and assignee information for 16,439,535 patents (99.9% of the patents in the final dataset).[16]

Note that while the EPO dataset contains unique IDs for each assignee and inventor (PERSON_ID), there is no such identifying information in the CNIPA/Google data. This lack of assignment is an issue since companies frequently patented—to an extent even greater than that in the U.S., where this has been a long-standing issue (Jaffe and Trajtenberg, 2002)—using the names of subsidiaries. Table B.1 provides an illustrative list of the various assignees associated with Lenovo. We create IDs for assignees missing one after we clean up their names. We describe this procedure next.

The final dataset has assignee and inventor names in both English and other languages. We translate the non-English names using the Google Cloud Translate API and provide both the original Chinese language names (when available) and the English translations.

4.2 Cleaning and Imputing Assignee Sector and Country

We then turned to completing two fields that were in some cases missing, sector and country (see the accounting at the end). In each case, we use machine learning techniques to fill in the missing data.

We prepare the training data by cleaning assignee names as follows: we make all assignee names upper case, remove punctuation, convert all characters to ASCII, ensure there is only one space between each word, and remove leading and trailing spaces from the names.

For the assignees missing a sector assignment, we predicted their sector assignment using the trained classifier. The sector of the assignee is a variable constructed by the EPO and is, as such, missing in the CNIPA dataset. For the assignees not missing sector assignments, we train the XGBoost algorithm on assignee names with the following labels: COMPANY, UNIVERSITY, GOV NON-PROFIT, GOV NON-PROFIT UNIVERSITY, HOSPITAL, COMPANY GOV NON-PROFIT, COMPANY GOV NON-PROFIT UNIVERSITY, COMPANY UNIVERSITY, COMPANY HOSPITAL, INDIVIDUAL, and UNKNOWN. The accuracy of the algorithm, as measured by a hold-out set, was 90.52%. Note that we dropped the single assignee whose sector is GOV NON-PROFIT HOSPITAL since the XGBoost algorithm needs a label to appear at least twice for training.

We proceeded similarly for missing country assignments. For the assignee names not missing country assignment, we trained the XGBoost algorithm on assignee names with two-digit country names, and then for the assignees missing country assignment, we predicted their country assignment using the trained classifier. To impute missing country values, we restrict the training set to assignees in six nations that are responsible for the largest share of patents in China: China, Hong Kong, Korea, Singapore, Japan and the United States. (Of

[16] Just like we produce a map for IDs, allowing researchers to trace which patents we dropped, we also provide a map for the assignee-inventor data. Further information in Appendix A.

the 1,194,800 assignees—here counts are based on IDs rather than their names—1,087,990 (91%) belong to one of these six countries.)[17]

4.3 Disambiguating Assignee Names

We implement three further steps to identify unique firms, i.e., "disambiguate" assignees.

First, for the assignees that were labeled as "COMPANY," we remove names of Chinese cities and provinces. We use the china-cities (version 0.0.4) Python package for this list. We do not make this transformation for other assignees since universities, hospitals, and laboratories are often identified by the name of their location. We also remove all leading numbers in assignee names if the numbers have at least five digits. This affects a range of assignees who were recorded as "23456 Ontorio Corp." or the like. We also stripped away terms indicating organization type (Ltd, Corp, GMBH, etc.) We use the python library cleanco to do this and apply it three times since a single application often removes only partial extensions, such as removing just Ltd from a company that ends in Co. Ltd.

Second, we stem common words using the mapping provided by the NBER. Some examples of this mapping are as following: TECHNOLOGY: TECH, CORPORATION: CORP, BROEDERNA: BRDR, CENTRO: CENT, CENTRAL: CENT, INTERNATIONAL: INT, etc.

As noted above, EPO provides PERSON_ID for each assignee and the OECD provides a HAN_ID corresponding to each disambiguated HAN_NAME. We construct an ID for assignees missing an ID, i.e., those that were taken from CNIPA. We do this by assigning a unique number to each unique tuple of (cleaned assignee name, assignee sector, assignee country).

The following summarizes the resulting mixture of assignees.

- Of the 16,440,045 grant IDs, 16,243,616 (99%) are linked to at least 1 assignee.
- In the raw assignee data, there are 2,691,649 unique assignee IDs and 2,414,183 unique assignee names. Of these assignees, 911 (0.03% of 2,691,645) are missing information on sector. 911 (0.03%) are missing information on their country of location.
- After translation, cleaning, and imputation (as described above), the description of the final sample of assignees is as follows:
  - The number of unique assignee names: 1,923,497

[17]Note that this procedure does not drop assignees that do not belong to one of these six countries. Instead, we only train the algorithm on assignees that belong to one of these six countries, and then only impute values for assignees missing country values. In the final dataset, the other major countries of assignees are Germany, France, and Great Britain.

- The distribution of patents to sectors:
  - COMPANY: 12,417,008
  - UNIVERSITY: 2,470,851
  - INDIVIDUAL: 1,550,531
  - GOV NON-PROFIT: 494,054
  - GOV NON-PROFIT UNIVERSITY: 259,115
  - UNKNOWN: 137,074
  - HOSPITAL: 63,206
  - COMPANY GOV NON-PROFIT: 38,214
  - COMPANY UNIVERSITY: 6,350
  - COMPANY GOV NON-PROFIT UNIVERSITY: 2,848
  - COMPANY HOSPITAL: 2,765
- The distribution of assignees to sectors:
  - COMPANY: 1,430,365
  - INDIVIDUAL: 318,051
  - UNKNOWN: 63,156
  - GOV NON-PROFIT: 56,537
  - UNIVERSITY: 28,754
  - COMPANY GOV NON-PROFIT: 11,441
  - HOSPITAL: 6,901
  - GOV NON-PROFIT UNIVERSITY: 4,414
  - COMPANY UNIVERSITY: 2,554
  - COMPANY HOSPITAL: 1,049
  - COMPANY GOV NON-PROFIT UNIVERSITY: 275
- The distribution of assignees to countries:
  - CN: 1,727,927
  - US: 77,038
  - JP: 37,099
  - KR: 13,686
  - DE: 11,476
  - GB: 5,496

4.4 Identifying Venture Backed Firms

Venture capital has long been understood to be associated with greater and more consequential innovations (Kortum and Lerner, 2000; Bernstein et al., 2016). In the context of China, venture investments have also been seen a key mechanism for advancing the transfer

and development of national security-related and dual use technologies (on the Strategic Competition Between the United States and The Chinese Communist Party, 2024). Thus, we wished to identify venture-backed firms.

We tag patents as being backed by venture capital if they were filed within five years of the assignee's first round of venture financing. We wanted to restrict attention to the patents that should reasonably be coded as venture-backed. As examples of what we would want to exclude, many patents in our sample are awarded to firms like Alibaba, which had received venture financing decades before the grant date of many of the patents in our sample. One approach is to code patents as venture-backed if the application year falls between the first and last years of venture financing. However, this approach has two potentially undesirable features. First, it artificially deflates the number of measured venture-backed patents in time periods when firms went public very quickly, and artificially inflates the number of venture-backed patents during time periods where firms remained private for much longer periods. Second, a number of firms in the sample received their last round of financing many years after their initial round. In most cases, these instances are buyouts or other "take private" deals, private investments in public entities (PIPEs), or venture financings of companies that had been taken private after an initial public offering, none of which correspond to the traditional definition of venture activity. While it was feasible to purge a number of these financings, it is not possible to do so in all cases. Reflecting these concerns, our chosen approach to define venture-backed patents as those that were applied for within five years of the firm's first round of venture financing helps mitigate these concerns.

We used PitchBook, which we have argued elsewhere (Lerner et al., 2024) appears to be the best commercial dataset for documenting global venture capital activity. We extracted the universe of companies headquartered in China, restricting the sample to firms ever financed by investors whose Primary Investor Type is listed as "Venture Capital." We impose no restrictions on the investors themselves, whose headquarters can be in any country. We use the following four files from PitchBook to merge company identifiers with investor types: Deal.dat, DealInvestorRelation.dat, Investor.dat, and Company.dat. Our data was pulled from the quarterly PitchBook data release dated January 11, 2023. There were a total of 23,302 Chinese firms that received at least one deal from an investor whose type is Venture Capital.

We merged the assignee dataset with the PitchBook dataset based on the names of companies using a two-step process:

- We first used a vectorizer to learn the vocabulary, i.e., the full set of words in the names of patents' assignees (note that we vectorized only after having cleaned the firm names based on the procedure described above.) Based on this vocabulary, we transformed both the patent assignees and VC-backed firms into vectors based on the TF-IDF (Term Frequency-Inverse Document Frequency) values of the words in the assignee names. We then calculated the pairwise similarity between each assignee name in the patent dataset and the company name in the PitchBook dataset. For each assignee, we kept the VC-backed firm whose name has a cosine similarity value of at least 0.8.

This was a low cut-off value and led to a large number of potentially overly lenient matches. Note that each patent assignee was only allowed to link to one firm from the PitchBook dataset, but each PitchBook firm may be linked to multiple assignees in the patent dataset.

- Second, to cull the matches further, we calculated the Levenshtein distance for each pair of matches. The Levenshtein distance measures the total number of characters that would have to be moved in a pair of strings to make them exactly alike. We used this measure to construct a Levenshtein score by dividing the Levenshtein distance by the sum of the lengths of the matched pair of names. Normalizing the Levenshtein distance by the total length of strings ensured that we did not penalize longer names, which are more likely to have a larger number of discrepancies. We then only kept the matches where the score was in the bottom 25th percentile in the dataset.

This two-step procedure led to a much better matching performance than either of the methods on its own. By utilizing word-based matching before letter-based matching, we also ensured the procedure was quite fast since calculating Levenshtein distance is a computationally slow process. Table B.2 shows some illustrative results from this matching process.

We examined the top assignees (all those with over one hundred patents) and sought to understand whether they were indeed venture-backed firms. The primary sources that we used here were the detailed profiles assembled by PitchBook and CrunchBase (another venture capital database), though Wikipedia articles and news stories in English and Mandarin also proved to be useful information sources. We discovered that some assignees are long-established entities that have corporate venture units (and thus are providers and recipients of venture investments), including in two cases foreign corporations that appear to have set up dedicated corporate venture teams in China. We changed the coding of these patents to not being VC-backed. The assignees removed by this process are listed in Appendix C.

After these deletions, the sectoral distribution of the venture-backed awards is as follows:

- Total number of assignees tagged as VC-backed: 10,962 (count of assignee IDs)
- Total number of patents assigned to VC-backed firms (patents where the filing year is within 5 years of their first deal from a VC firm): 103,445 (corresponding to 9,755 assignee IDs.)

4.5 Identifying State-Owned Enterprises

State-owned enterprises have been documented to play a large (and indeed increasing) role in the innovation landscape in China (Wei et al., 2017; König et al., 2022).

We collect a list of China's state-owned enterprises (SOEs) from TianYanCha, a commercial database that has been used in academic research (e.g., Beraja et al. (2023)) though

access has been recently cut off to Western researchers.[18] We cleaned these names just as we cleaned assignee names and then merged with the patents dataset using the two-step procedure described in Section 4.4.

Note that, like all string-matching techniques, this technique is also imperfect. 3,010 assignees (with unique assignee IDs) are matched to multiple SOEs. For example, CHANGZHOU RAIL TRANSIT DEVELOPMENT CO., LTD. becomes RAIL TRANSIT DEV after the two-step cleaning procedure above and is linked to RAIL TRANSIT, RAIL TRANSIT DEV, and RAIL TRANSIT IND DEV. There is no clear way for us to choose among these matches, so we choose the TianYanCha SOE name that is the first when sorted alphabetically.

To summarize these patents:

- Total number of assignees tagged as SOEs: 28,124
- Total number of patents assigned to SOEs: 497,576
- Sector allocation (count of assignee id in each sector):
  - COMPANY: 27,607
  - GOV NON-PROFIT: 210
  - UNKNOWN: 113
  - UNIVERSITY: 91
  - INDIVIDUAL: 44
  - COMPANY GOV NON-PROFIT: 30
  - HOSPITAL: 16
  - COMPANY HOSPITAL: 10
  - COMPANY UNIVERSITY: 2
  - COMPANY GOV NON-PROFIT UNIVERSITY: 1

4.6 Identifying PLA-Affiliated and Related Patents

After examining the data, we discovered that a substantial number of patents were assigned to entities affiliated with the People's Liberation Army, the armed wing of the Chinese Communist Party and the principal military force of the People's Republic of China. Rather than being assigned to a single, readily identified assignee, these were typically awarded to various sub-units.

[18] Source: `https://www.reuters.com/business/finance/chinas-top-financial-data-provider-restricts-offshore-access-due-new-rules-2023-05-04/`, Accessed in 2023.

We identify an assignee as the PLA if it contains any of: FORCES, PLA, TROOPS, LIBERATION, or ARMY. We tag 5,302 assignees as affiliated with the PLA. Some representative examples of assignees are listed in Table B.3.

Amongst them, the sector allocation was as follows:

- COMPANY: 459
- GOV NON-PROFIT: 2,081
- UNKNOWN: 465
- UNIVERSITY: 1,179
- INDIVIDUAL: 118
- COMPANY GOV NON-PROFIT: 52
- HOSPITAL: 810
- COMPANY HOSPITAL: 22

While as in U.S., numerous universities may undertake national security-relevant research, special attention has the been given to the "Seven Sons of National Defense" in China. These public universities have affiliations with the Ministry of Industry and Information Technology of China and are believed to have close ties with the PLA. Stoff and Tiffert (2020), for instance, document that these schools as "directly support the country's defense research and industrial base and that operate as prime pathways for harvesting US research and diverting it to military applications."'[19]

In total, there are 176,629 patents affiliated with these seven universities.

4.7 Cleaning Inventor Names

The vast majority of inventor names are not in English. We use the Google Translate API to translate them to English and produce both the original names (when available) and translated versions.

[19]In May 2020, The Trump Administration canceled the visas of Chinese graduate students and researchers who have direct ties to these seven universities (`https://www.insidehighered.com/news/2020/05/29/us-plans-cancel-visas-students-ties-universities-connected-chinese-military`).

5 Measures of Patent Quality

There have been three widely used measures of patent importance. These three measures, while positively correlated (Kelly et al., 2021), differ in both their methodologies and points of focus, and thus identify different patents and firms as the most impactful:

- The first of these was the subsequent patent citations that the patent garnered. This metric measures the scientific value of a patent based on how many follow-on innovations build on that patent (Jaffe and Trajtenberg, 2002). Because the propensity to cite patents varies across technologies and over time, we normalize the citations by the mean number received by other patents in that one-digit Combined Patent Classification (CPC) class and awarded in the same quarter.
- The second impact measure was the Kogan et al. (2017) estimate of patent value, based on market reactions to the award grants. This measure could only be calculated for publicly traded firms. Unlike the other two measures, this metric only captures private, rather than private and social, returns.
- The final measure was the metric of patent novelty developed by Kelly et al. (2021), based on a comparison of the patent text with prior and subsequent patents. Because this measure requires a substantial corpus of subsequent patents, it was only calculated for patents awarded through the end of 2022.

We will focus in this paper on the first and third methodologies. Given the extent of price limits in Chinese stock markets (Zhang et al. (2022) is a recent discussion), which have been widely understood since the work of Fama (1988) and others to impede price discovery and efficiency, we do not consider a Kogan et al. (2017) style analysis.

5.1 Citations

When constructing citation counts, we must ensure two things. First, we do not want to double-count multiple citations made by different publications associated with the same patent. Second, we want to sum all citations made to a patent, even if they were made to its original application rather than the final granted publication.

To identify the citations, we employ Google BigQuery to extract citation data from Google Patents. These data are combined in a citation dataset. Each row in the citation dataset corresponds to a pair of patents where the grant ID refers to the patent that is cited and fw_cite_id refers to the patent that cites grant ID.

We look only at citations from the last version of the related Chinese patent document to other Chinese invention patent publications. We wanted to make sure that we did not

double-count citations. In particular, we only used the citations in the latest publication of each patent, rather than those from earlier or duplicative publications of the same award. We tabulate separately all citations, and those that are made by the patent examiner. We do not examine citations to patent awards in other nations and scientific publications, both of which appear to be much rarer than among U.S. awards in any case.

Given the proclivity of Chinese patents to cite the original publication rather than the final award, we aggregate all citations made to the various publications (i.e., A, B, or/and C) associated with a given award. Specifically, we use the map created when constructing the id dataset that maps all versions of a patent to the same application number. We merged the patents to this map and sum all citations to the application number. We then merge this back to the main dataset, so we have a count of forward citations, which we associate with the last published document for each application.

For example, the citation dataset contains citations to both CN1048867A and CN1015467B. Both are the same patent, corresponding to the application number 90102904. We compiled these together as part of the forward citation count for this patent.

5.2 KPST Scores

Kelly et al. (2021) introduced an alternative measure of patent importance, which focuses on the text of patent documents. The authors used advances in textual analysis to create links between each new invention and the set of existing and subsequent patents. Specifically, they constructed measures of textual similarity to quantify commonality in the topical content of each pair of patents. To identify significant (high quality) patents, the focus on those whose content is distinct from prior patents (is novel) but is like future patents (is impactful).

In this section, we describe the practical implementation of the score using data on the full text of Chinese patents. For a detailed, mathematical treatment of the construction of KPST scores, please see Kelly et al. (2021).

The starting point of this exercise is 39 files with full text of all patents granted in China between 1985 and 2023. Each row in these files has five columns corresponding to a patent's grant id, title, abstract, description, and claims. Files are organized by year of publication of a patent. The English translation of Chinese patents was extracted from Google.

Step 1.

We import each file, combine all the patent's text into one large vector, clean and standardize this text, and then vectorize it.

Step 2.

Based on the vectorized files, we create a "dictionary" that assigns term IDs to terms. This is a crucial step, since it makes the size of the follow-on files much smaller, thereby

speeding up the later computation. This file first aggregates terms from all the years of patent data, drops those that are part of the nltk "stopwords" directory (common terms used in English), and drops those that show up fewer than 20 times across all years. In the final dictionary, there are 2,234,383 terms with IDs assigned alphabetically from 0 to 2,234,382.

Step 3.

We re-organize the files by year of filing of patents. In practice, we import each vectorized filed from Step 1 (where files were organized by year of publication), merge the file to the dictionary so that all terms are assigned their IDs, and then collect patents filed in the same year together. The KPST score is calculated based on all patents filed in the same year rather than all patents that are published in the same year. Having assigned term IDs, we construct normalized Term Frequency (TF-norm) scores for each term in each patent. The final output is 39 files (one for each year between 1985 and 2023), where each row contains a patent-term ID pair, and a TF-norm number, which is equal to the ratio of the number of times that the term appears in the document, divided by the total number of terms in the document.

Step 4.

We create Backwards Inverse Document Frequency (BIDF) scores. These scores are best explained with a concrete example. When constructing pairwise similarity for patents belonging to two years, say 1990 and 2000, we need to normalize the term frequency (TF-norm) of each term in the patents by the popularity of that term in a shared base year. For each pair of years, the base year is equal to one year before the smallest year. For 1990 and 2000, therefore, BIDF calculates the prevalence of a term in patents filed in all years before 1989 (inclusive), where 1989 is the base year.

Step 5.

We construct pairwise similarity between each pair of patents in the data. In practice, we only construct pairwise similarity for year-pairs that are eventually used as inputs in calculating aggregate KPST scores. For example, there is no need to calculate pairwise similarities between patents filed in 1986 and 2021 since 2021 is more than twenty years after 1986. Our maximum forward similarity calculation is limited to 20 years, and maximum backward similarity is limited to 5 years.

Step 6.

We create patent-level Kelly scores. As and when available, we want to construct the following 5 scores for each patent: backward similarity for 5 years (bw5), forward similarity for 1 year (fw01), forward similarity for 2-5 years (fw25), forward similarity for 6-10 years (fw610), and forward similarity for 11-20 years (fw1120). For a 10-year Kelly score, for example, we calculate the following ratio: (fw01 + fw25 + fw610)/bw5.

5.2.1 Cross-National Citations

We also sought to gather patent citations from US patents to Chinese ones and vice versa, and vice versa. For Chinese cites of U.S. patents, we used Python scripts from the 'google.cloud.bigquery' library to extract data from 'google_citation.csv', which contains all the citations of patents in 'allids_map_finalid.csv,' as well as the 'allids_map_finalid.csv,' which contains the comprehensive set of Chinese patent IDs, both of which are part of Google Patents. Citations where the cited patent ID starts with 'US' were kept, resulting in 7,801,694 citations, including both applications and granted patents. After excluding citations to patent applications (which followed a format like US2019188295A1), the final count of citations from CN to US patents was 3,395,578.

For US citations of Chinese patents, we initially explored the use of PatentsView, which had 2,922,482 such citations. However, these citations were simply of the Chinese patent number and not the trailing letter, which (as discussed above) inconsistencies in the assignment of patents in some cases. The citations that could be extracted from Google Patents had the trailing letter and were used to ensure consistency. Thus, we again used Python scripts utilizing the 'google.cloud.bigquery' library to extract the needed data using the 'patents-public-data.patents.publications' file. All US patents citing CN patents were initially pulled, resulting in 2,914,338 citations, including citations from both applications and granted patents. After excluding citations from US patent applications (again, those with a format like US2019188295A1), the final count of citations from US to CN patents was 2,812,935.

5.2.2 Identifying Chinese Patents Also Filed in the U.S. and Cross-Citations

One empirical avenue we wished to explore was whether Chinese inventors were filing consistently in the U.S., or whether there were systematic differences across which patents were or were not being filed in the U.S. We followed two separate procedures to tag patents that were cross-filed in the U.S., reflecting the differences across the two databases.

For patents with data in PATSTAT, the procedure was straightforward. Using the DOCDB identifier, we were able to determine whether there was an associated U.S. patent by using the PATSTAT file for patents filed in the U.S. (For these patents, the publication authority is listed as the U.S.) We merged the Chinese patents data with the U.S. patents data using DOCDB_FAMILY_ID variable. All Chinese patents whose DOCDB_FAMILY_ID could be found in the U.S. patents data were tagged as having been cross-filed in the U.S. These include patents originally filed in China that were subsequently applied for in the U.S. (our key focus), as well as cases where an investor in the U.S. filed subsequently filed in China or where an investor in a third nation later applied in both nations. (Typically, inventors will file domestically first and then decide whether to pursue foreign filings, as this procedure can allow them to defer the costs of a foreign filing for between one and three

years, depending on the procedure used.)[20]

The CNIPA-only patents, however, did not have an associated DOCDB family number or any other indicator that there were additional filings. Thus, for the CNIPA-only patents, we find whether they were cross-filed in the U.S. by using U.S. patent data. We download the file g_foreign_priority.tsv from PatentsView. This file provides information about all U.S. patents for which an earlier patent filing elsewhere (i.e., in a foreign country) was a priority (original) filing. For each patent with a foreign priority, this dataset also provides the corresponding foreign application number. We identify all patents where the variable foreign_country_filed records CN; we find that 136,586 Chinese patents were filed in the U.S. using this method.[21]

In total, 2,142,197 (13%) of the Chinese patents are tagged as being cross-filed in the U.S.

6 Publicly Available Data

We are releasing the following four files publicly:

1. cn_pats.csv
2. cn_assignees_inventors.csv
3. map_cn_pats.csv
4. map_cn_assignees_inventors.csv

cn_pats.csv contains all the 16,440,045 grant IDs in the final dataset as well as their primary bibliographic information: application numbers, dates of filing and publication, dummies for the countries of origin of their assignees, citation counts, and KPST scores.

cn_assignees_inventors.csv contains data on the assignees and inventors associated with each patent. For each assignee, the dataset contains several other pieces of information, such as its sector, country, and type (SOE/ VC/ PLA/ Seven Sons of Defense).

map_cn_pats.csv contains data on all the patent IDs received from the EPO, CNIPA, and Google datasets, alongside the final grant ID that shows up in cn_pats.csv.

[20]For a helpful overview of the international patent filing process, see https://www.mewburn.com/law-practice-library/international-pct-patent-applications-the-basics.

[21]There are approximately 55,000 patent application numbers corresponding to Chinese patents in the g_foreign_priority data that do not exist in any of our datasets. We spot-checked several of them using information on patent title, assignee names, and grant dates. Similar patents (i.e. those filed by the same assignees on nearby dates and on similar topics) do exist in our data. We believe part of the reason that these application numbers are missing from the data is that they are incorrect.

map_cn_assignees_inventors.csv links each patent in the map (map_cn_pats.csv) to information about its assignees and inventors.

Appendix A contains the codebook for each of these files with corresponding variable names and descriptions.

7 Summary of the Data

We now present some summary statistics on the sample. This gives an initial sense of the nature of the Chinese awards.

Figure 3 presents the breakdown of Chinese patent publications by nationality. Panel A shows how the Chinese share of the awards, while below one-half for much of the 1990s and early 2000s, began increasing sharply about 2004. In recent years, about 90% of the awards have been to Chinese assignees. Panel B looks at the breakdown of other countries. The U.S. remains the most important other patenting nation, though Korea and other major industrialized nations are also represented.

Figure 4 looks at the breakdown of the patent publications. Recall that we use in each case we use the latest publication data. We see how C patents replace B patents during the period when CNIPA had a tri-partite publication process. About 60% of the patents are, however, are A publications, consistent with accounts that many applicants are content to leave their applications unexamined. The share trends up in recent years, presumably reflecting some awards that will ultimately be examined and published.

Figure 5 examines the subset of patent publications associated with Chinese entities. Here, we break down the nature of the assignee that filed the publication. If there are multiple assignees, we assign the patent pro rata to the different entities. We see the sharp decline of individual inventors and the rise of corporate assignees. (These patterns mirrored what happened in the United States, as documented by Lamoreaux and Sokoloff (2005).) We also see how the share of patenting by universities and non-profits among Chinese entities seems much greater than in the U.S.[22]

Figure 6 shows the evolution of Chinese patents (associated with domestic entities only) backed by Chinese or U.S. VC firms over time. It's evident that there is a sharp increase in Chinese patents backed by Chinese VC firms after 2008, outpacing the growth of those backed by U.S. VC firms.

Figure 7 looks at the mixture of technologies in these patents. Again, we focus on Chinese awards to Chinese entities and USPTO ones to U.S. entities. We compare the CPC section (one-digit patent class) for the U.S. awards with the corresponding IPC section for

[22]For instance, looking at U.S. assignees of USPTO patents applied for between 2000 and 2018 and awarded by 2019, Lerner et al. (2013) find that 2.55% of awards were assigned to universities.

the Chinese awards. The two panels reveal the much greater representation of sections G and H (Physics and Electricity, which include the bulk of the awards in information and communications technologies) in the U.S., though also the rapid Chinese catch-up in the share of section G, largely at the expense of sections A (Human Necessities, which includes pharmaceuticals) and C (Chemistry; Metallurgy).

Figure 8, Figure 9, Figure 10 and look at our two measures of patent quality. Figure 8 looks at the mean number of citations in Chinese awards for patents that are ten or more years old; those between five and ten years; and those under five years of age. The citation density is substantially less for the Chinese awards: patents ten or more years of age have on average 4.5 citations, while those granted between 2013 and 2017 have 1.3 citations. This is much less than in the U.S.: for instance, (Hegde et al., 2023) find that U.S. patents have on average 13.2 citations in the ten years after their initial publication, whether at issue or a pre-grant disclosure.

Figure 9 looks at the temporal pattern of citations. Here we confine the analysis to all Chinese patents that were at least 10 years old as of September 2023 and had at least one subsequent Chinese patent citation. As a result, the mean number of citations to these patents is greater than the total population. We find that the annual rate of citations peaks five years after final publication, and then tails off after. There remains a long tail of much older citations. These patterns are consistent with the U.S. patterns documented by Jaffe and Trajtenberg (2002).

Figure 10 looks at the second measure of patent quality - KPST scores. Specifically, the panels present the distribution of Chinese patents' KPST scores over different horizons (i.e., using 1-year, 5-year, 10-year, and 20-year forward similarity measures, respectively). Since the distribution of raw KPST scores is highly skewed, for a more straightforward interpretation, we first transform the original calculated KPST scores in log terms. In addition, as documented in Kelly et al. (2021), the value of original KPST measures could be mechanically affected by factors including language shifts and the rapid expansion in patent filings over time. We adjust the log-transformed KPST scores by removing patent cohort filing year fixed effects. After the adjustment, a patent with a positive and higher adjusted KPST value suggests that the patent has a higher quality (i.e., high impact measured by the forward similarity and high novelty measured by the backward similarity) relative to patents filed in the same year. Compared the distribution of the same quality measure using US patents data granted between 1985 and 2022 as shown in Figure D.1, we find the distribution of Chinese patents' adjusted KPST values has a much thicker left tail relative to that in the US, implying that the quality distribution of Chinese patents is much more skewed (with a higher proportion of patents with relative low qualities.)

Table 1 presents some statistics for the entire sample. Among the key patterns are the relatively small number of venture-backed patents relative to elsewhere in the world (Lerner and Gompers, 2002), the representation of defense-linked assignees, and the greater youth (a mean publication date of December 2017 for the Chinese-assigned awards as opposed to June 2017 for the total sample) of the awards to local entities.

Final Thoughts

This essay has described the creation of a database on Chinese patents. The critical ways in which the available data and institutional features of the Chinese system differ are highlights, as well as our responses to these challenges.

We acknowledge that there are several natural next steps to develop these data further. Two potential avenues are the disambiguation of assignee names following the approach of Li et al. (2014)—an initial effort to do so with the Chinese data is Yin et al. (2020)—and the development of bigrams a la Kalyani et al. (ming) to facilitating merging with other datasets. It is our hope that this will be a resource to scholars worldwide.

References

Aghion, P., J. Cai, M. Dewatripont, L. Du, A. Harrison, and P. Legros (2015). Industrial Policy and Competition. *American Economic Journal: Macroeconomics 7*(4), 1–32.

Alcácer, J. and M. Gittelman (2006). Patent Citations as a Measure of Knowledge Flows: The Influence of Examiner Citations. *The Review of Economics and Statistics 88*(4), 774–779.

Beraja, M., D. Y. Yang, and N. Yuchtman (2023). Data-intensive Innovation and the State: Evidence from AI Firms in China. *The Review of Economic Studies 90*(4), 1701–1723.

Bernstein, S., X. Giroud, and R. R. Townsend (2016). The Impact of Venture Capital Monitoring. *The Journal of Finance 71*(4), 1591–1622.

Chen, Z., Z. Liu, J. C. Suárez Serrato, and D. Y. Xu (2021). Notching R&D Investment with Corporate Income Tax Cuts in China. *American Economic Review 111*(7), 2065–2100.

Cheng, W. (2023). *China in Global Governance of Intellectual Property: Implications for Global Distributive Justice* (1st ed.). Palgrave Macmillan.

Fama, E. F. (1988). *Perspectives on October 1987, Or, What Did We Learn from the Crash?* Center for Research in Security Prices, Graduate School of Business, University of Chicago.

Fang, L. H., J. Lerner, and C. Wu (2017). Intellectual Property Rights Protection, Ownership, and Innovation: Evidence from China. *The Review of Financial Studies 30*(7), 2446–2477.

Fuller, D. B. (2016). *Paper Tigers, Hidden Dragons: Firms and the Political Economy of China's Technological Development*. Oxford University Press.

Griliches, Z. (1984). R&D, Patents, and Productivity. Backup Publisher: National Bureau of Economic Research.

He, Z.-L., T. W. Tong, Y. Zhang, and W. He (2018). Constructing a Chinese Patent Database of listed firms in China: Descriptions, lessons, and insights. *Journal of Economics & Management Strategy 27*(3), 579–606.

Hegde, D., K. Herkenhoff, and C. Zhu (2023). Patent Publication and Innovation. *Journal of Political Economy*.

Holmes, T. J., E. R. McGrattan, and E. C. Prescott (2015). Quid Pro Quo: Technology Capital Transfers for Market Access in China. *The Review of Economic Studies 82*(3), 1154–1193.

Jaffe, A. B. and M. Trajtenberg (2002). *Patents, Citations, and Innovations: A Window on the Knowledge Economy*. MIT Press.

Kalyani, A., N. Bloom, M. Carvalho, T. A. Hassan, J. Lerner, and A. Tahoun (Forthcoming). Diffusion of New Technologies. *Quarterly Journal of Economics*.

Kelly, B., D. Papanikolaou, A. Seru, and M. Taddy (2021). Measuring Technological Innovation over the Long Run. *American Economic Review: Insights 3*(3), 303–320.

Kogan, L., D. Papanikolaou, A. Seru, and N. Stoffman (2017). Technological Innovation, Resource Allocation, and Growth. *The Quarterly Journal of Economics 132*(2), 665–712.

König, M., K. Storesletten, Z. Song, and F. Zilibotti (2022). From Imitation to Innovation: Where Is All That Chinese R&D Going? *Econometrica 90*(4), 1615–1654.

Kortum, S. and J. Lerner (2000). Assessing the Contribution of Venture Capital to Innovation. *The RAND Journal of Economics 31*(4), 674–692.

Lamoreaux, N. R. and K. L. Sokoloff (2005). The Decline of the Independent Inventor: A Schumpterian Story? Working Paper 11654.

Lerner, J. and P. Gompers (2002). *The Venture Capital Cycle*.

Lerner, J., J. Liu, J. Moscona, and D. Yang (2024). Appropriate Entrepreneurship? The Rise of China and the Developing World. Technical report, National Bureau of Economic Research.

Li, G.-C., R. Lai, D. Alexander, D. M. Doolin, Y. Sun, V. I. Torvik, A. Z. Yu, and L. Fleming (2014). Disambiguation and co-authorship networks of the U.S. patent inventor database (1975–2010). *Research Policy 43*(6), 941–955.

Magerman, T., B. Van Looy, and X. Song (2006). Data Production Methods for Harmonized Patent Statistics: Patentee Name Harmonization. SSRN Scholarly Paper 944470.

on the Strategic Competition Between the United States, T. S. C. and U. H. o. R. The Chinese Communist Party (2024). The CCP's Investors: How American Venture Capital

Fuels the PRC Military and Human Rights Abuses. Accessed: 2025-01-20.

Prud'homme, D. (2012). *Dulling the Cutting Edge: How Patent-Related Policies and Practices Hamper Innovation in China*. European Chamber.

Putnam, J. (1997). The Value of International Patent Rights. Technical report, Yale University.

Stoff, J. and G. Tiffert (2020). Under the Radar: National Security Risk in US-China Scientific Collaboration. In G. Tiffert (Ed.), *Global Engagement: Rethinking Risk in the Research Enterprise*, pp. 19–104. Stanford, Hoover Institution.

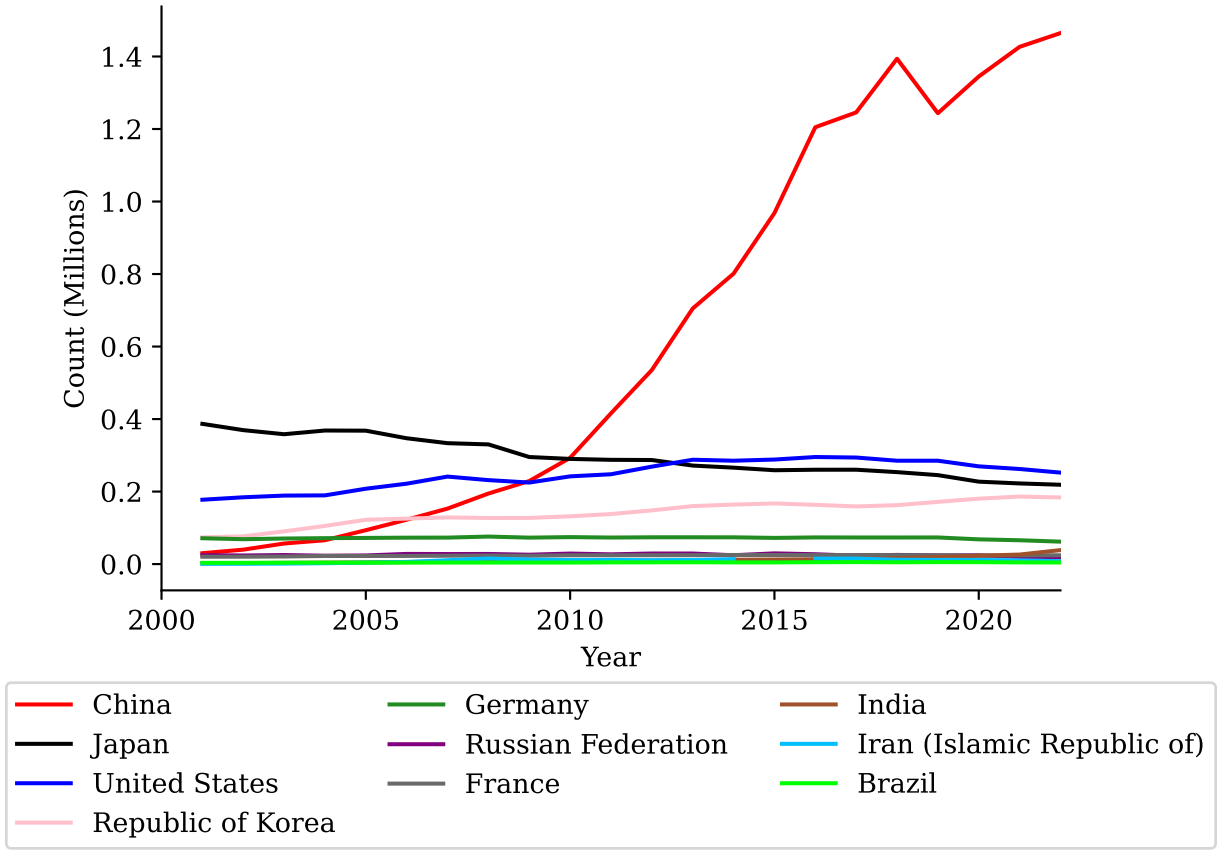
Tong, T. W., K. Zhang, Z.-L. He, and Y. Zhang (2018). What Determines the Duration of Patent Examination in China? An Outcome-Specific Duration Analysis of Invention Patent Applications at SIPO. *Research Policy 47*, 583–591.

Wei, S.-J., Z. Xie, and X. Zhang (2017). From 'Made in China' to 'Innovated in China': Necessity, Prospect, and Challenges. *Journal of Economic Perspectives 31*, 49–70.

Yin, D., K. Motohashi, and J. Dang (2020). Large-scale name disambiguation of Chinese patent inventors (1985–2016). *Scientometrics 122*, 765–790.

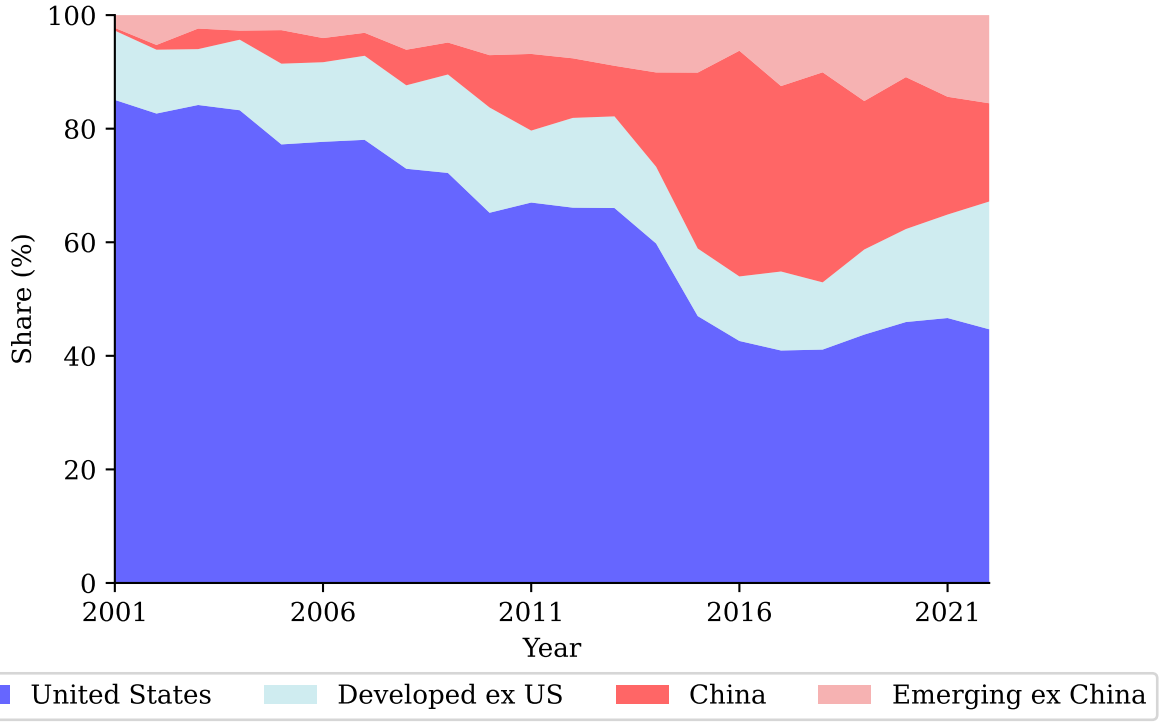
Zhang, X., Z. Wang, J. Hao, and F. He (2022). Price limit and stock market quality: Evidence from a quasi-natural experiment in the Chinese stock market. *Pacific-Basin Finance Journal 74*, 101778.

Figure 1: Domestic Patent Applications in Top Five Developed and Emerging Economies, 2001-2022



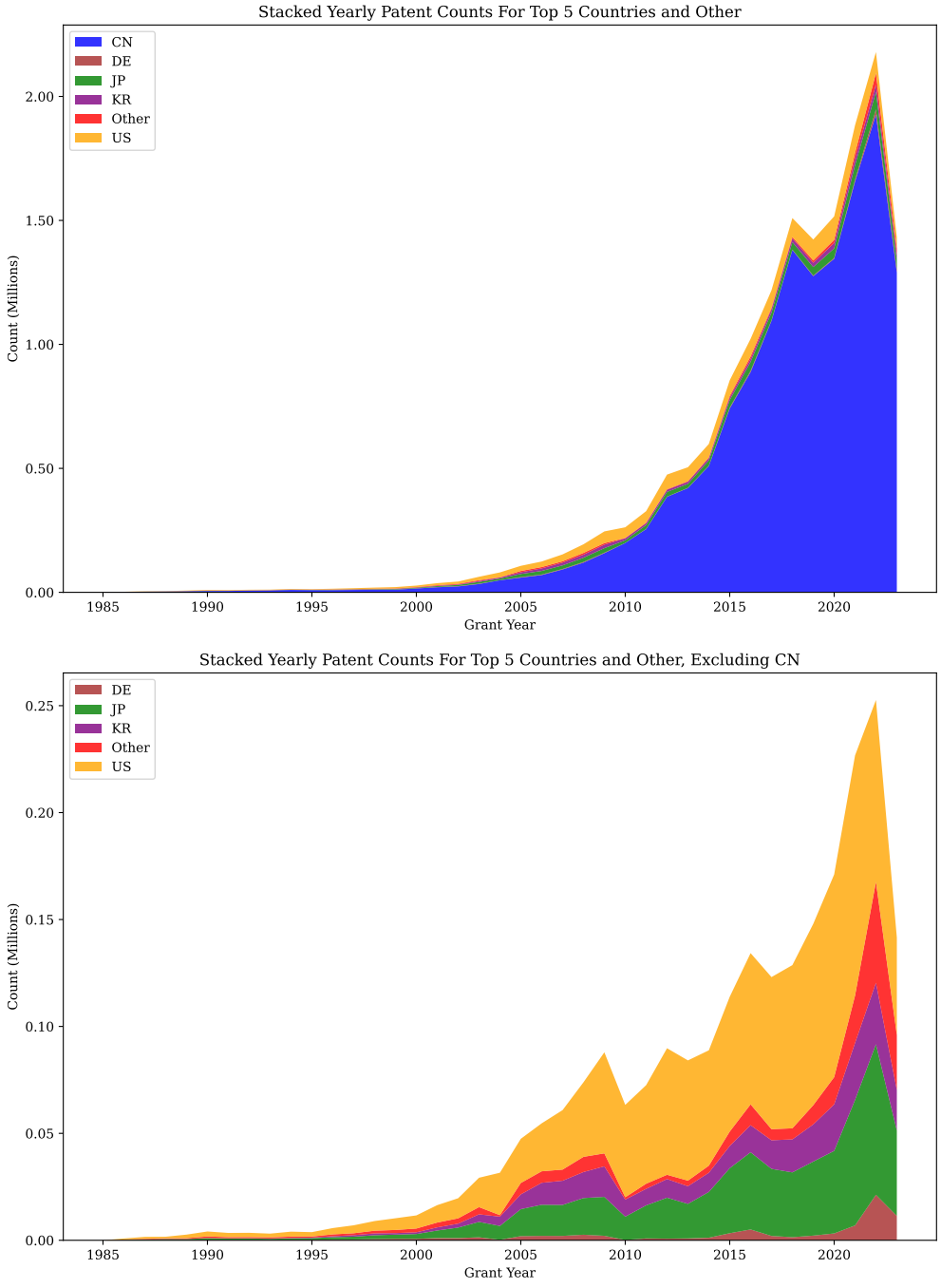
Notes: Domestic patent applications in the top five developed (Japan, U.S., Korea, Germany, and France) and emerging (China, Russia, India, Iran, and Brazil) economies, based on cumulative patenting in the period, 2001-22. The source is the World Intellectual Property Organization, data accessed from (`https://www3.wipo.int/ipstats/key-search/indicator`).

Figure 2: Relative Share of Global Venture Capital Funding, 2001-2021



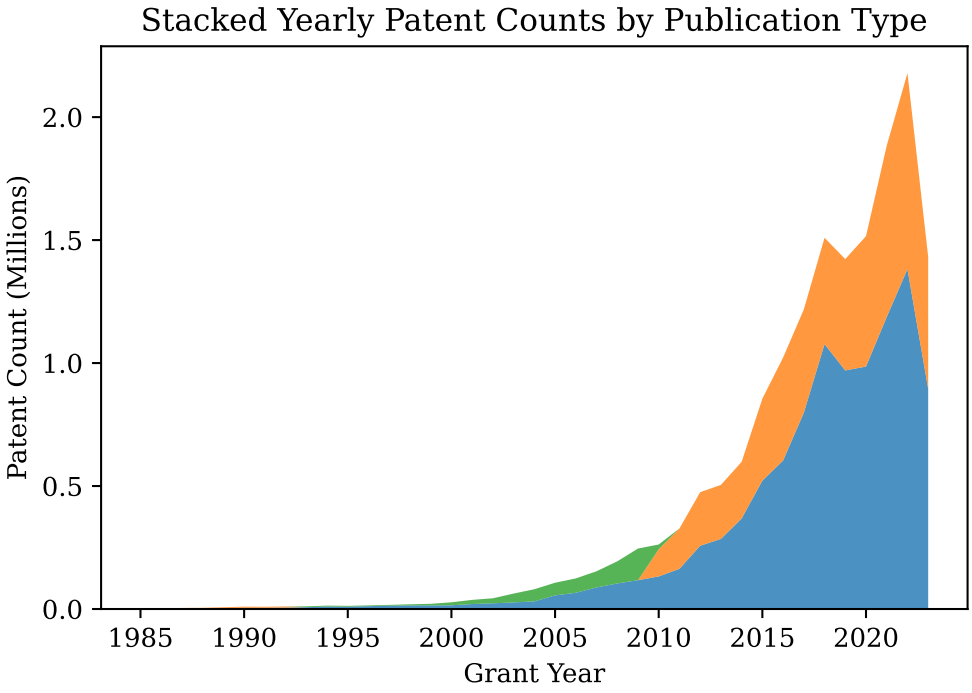
Notes: This figure plots the share of global venture capital investments going to different parts of the world each year. Data from (Lerner et al., 2024).

Figure 3: Chinese Patenting By Nationality and Grant Year



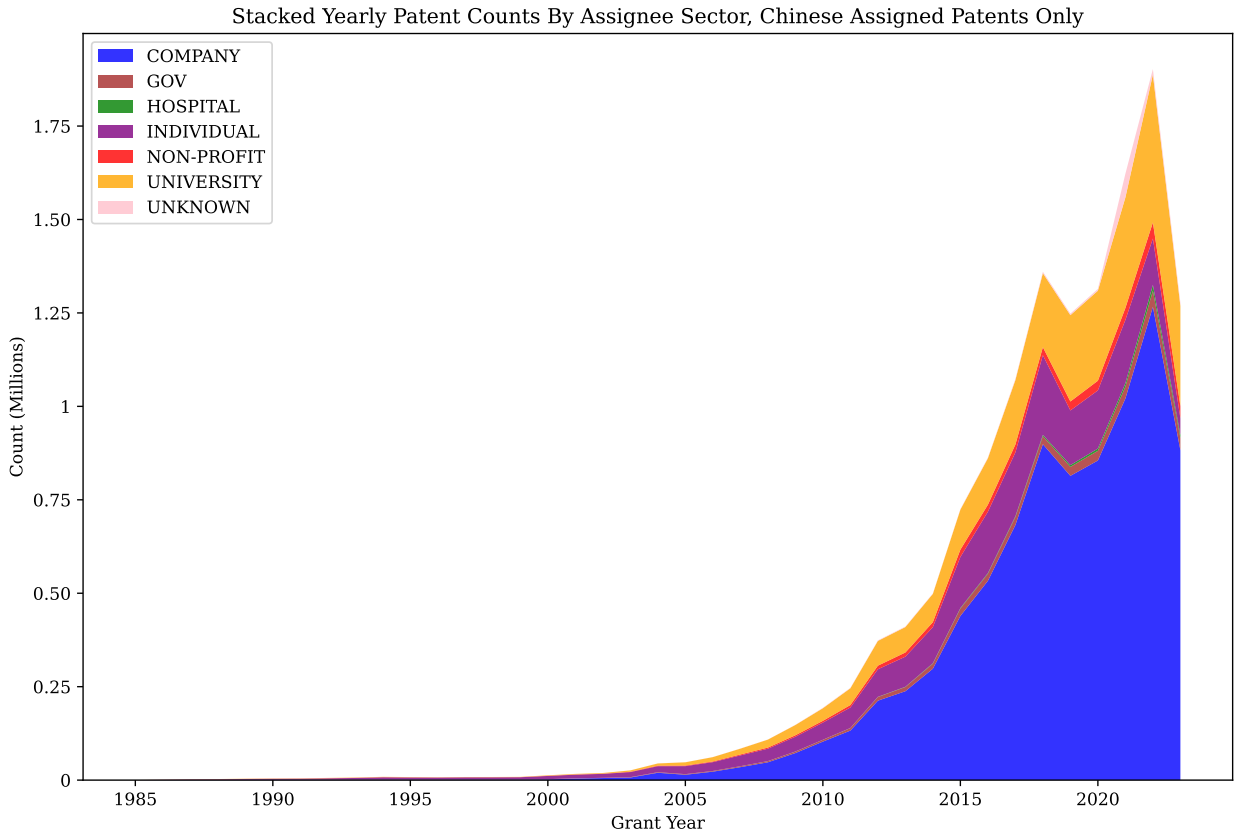
Notes: Composition of Chinese patent by assignee nationality between September 1985 and August 2023. The top panel presents the decomposition of the top five assignee nationalities of Chinese patents, with all remaining countries grouped under the category labeled as "Other". The lower panel displays the decomposition of the top assignee nationalities, excluding patents granted to assignees in China. For patents with assignees from multiple countries, we assign the patent pro rata to different nationalities.

Figure 4: Chinese Patent Publications by Patent Type and Grant Year



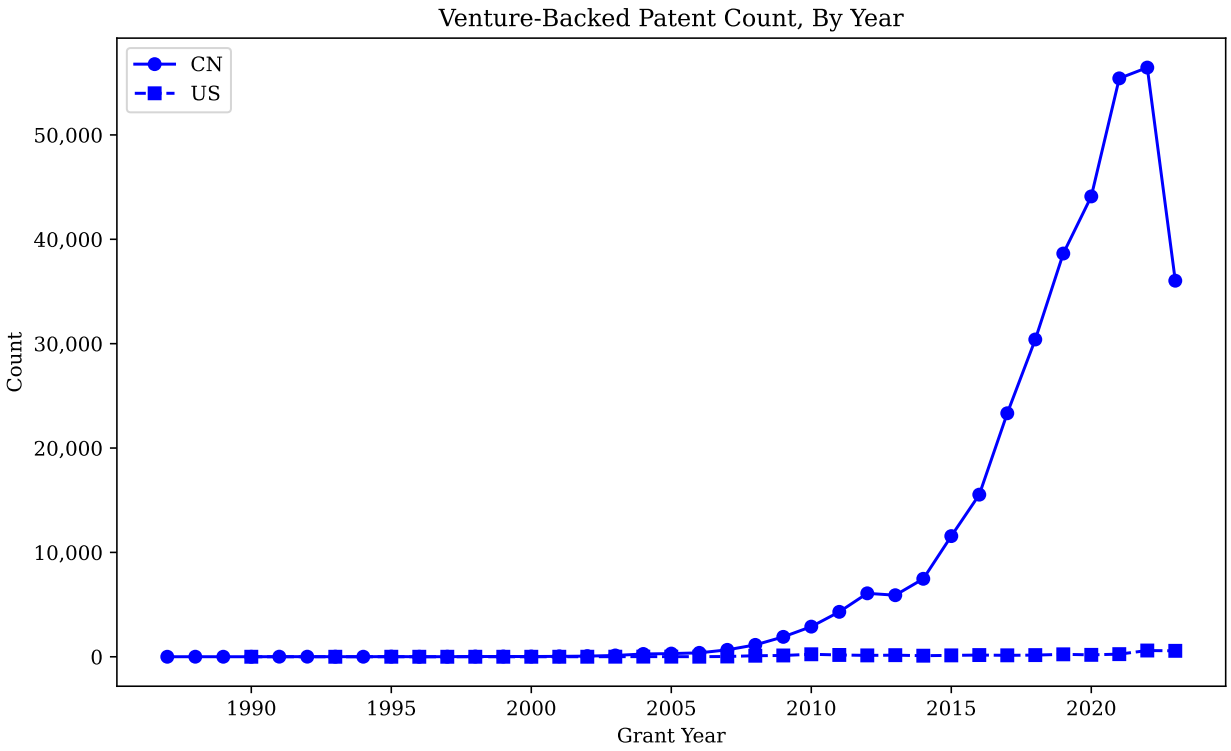
Notes: This figure exhibits the different publication types present in the final Chinese patents data.

Figure 5: Domestic Chinese Patenting By Assignee Sector and Grant Year



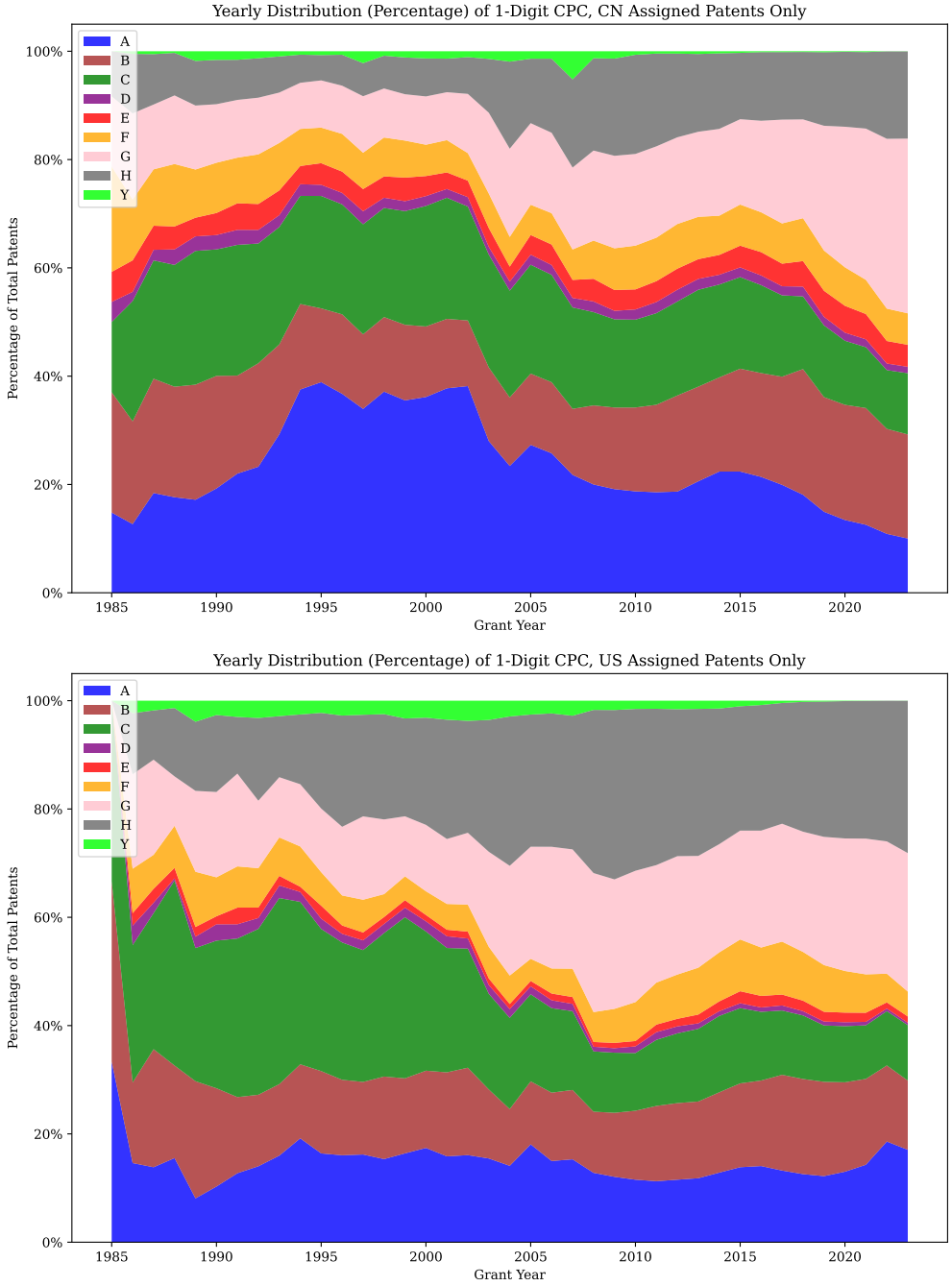
Notes: Composition of Chinese patent associated with domestic entities by sector between September 1985 and August 2023. We categorize patents' assignees into the following sectors: company, government, hospital, university, non-profit, and individual, with all remaining assignees grouped under the category labeled as "Unknown". For patents with assignees from multiple sectors, we assign the patent pro rata to different sectors.

Figure 6: Venture-Backed Domestic Chinese Patenting by Grant Year



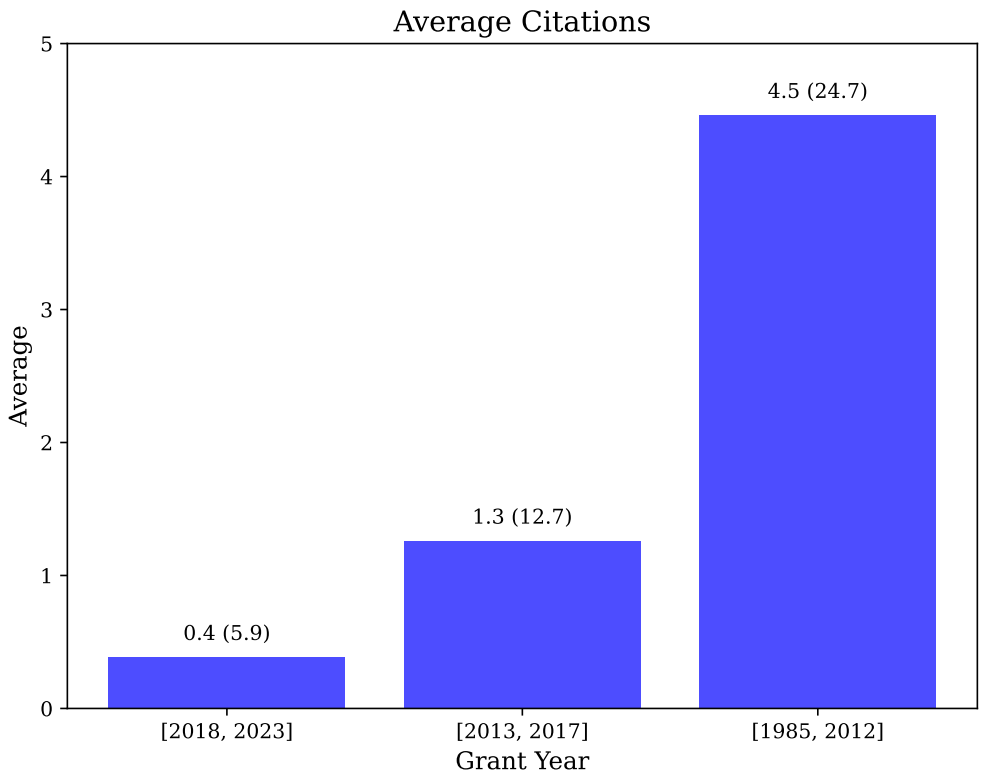
Notes: Number of Chinese patents associated with domestic entities and backed by Chinese or U.S. venture capital (VC) firms between September 1985 and August 2023. A patent is classified as China (U.S.) venture-backed if at least one assignee received China (U.S.) VC funding within five years prior to patent issuance. The solid line represents the total number of Chinese patents backed by Chinese VC firms across the grant year, while the dashed line represents those backed by U.S. VC firms.

Figure 7: Distribution of One-Digit Primary IPC/CPC Classes, By Year



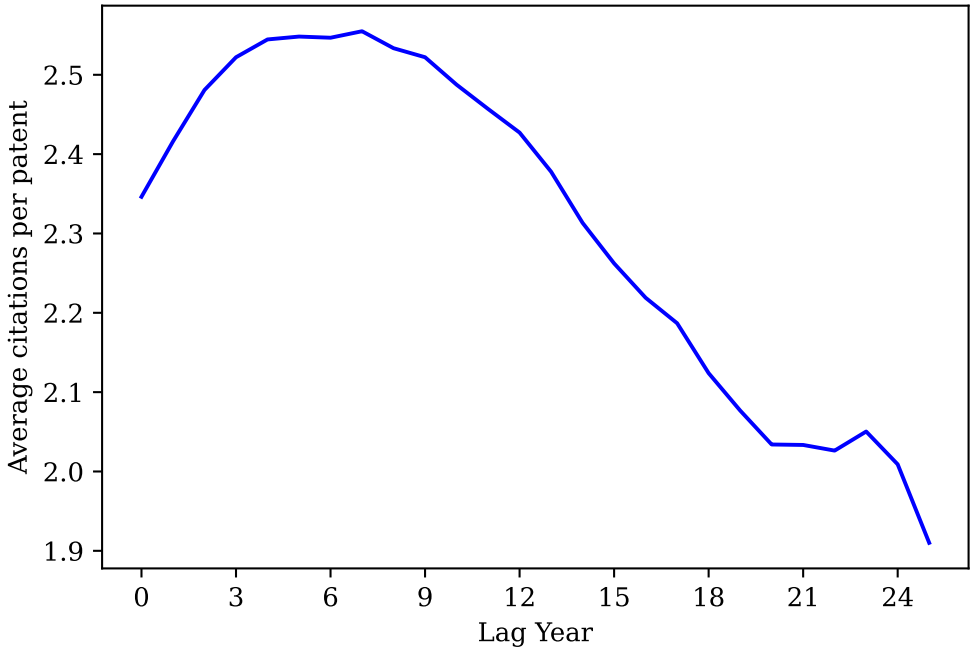
Notes: Composition of Chinese-assigned and U.S.-assigned patents by technology class from September 1985 to August 2023. The top panel shows the decomposition of Chinese-assigned patents, while the lower panel presents the breakdown of U.S.-assigned patents, categorized by their primary 1-digit CPC (or IPC) class.

Figure 8: Citations



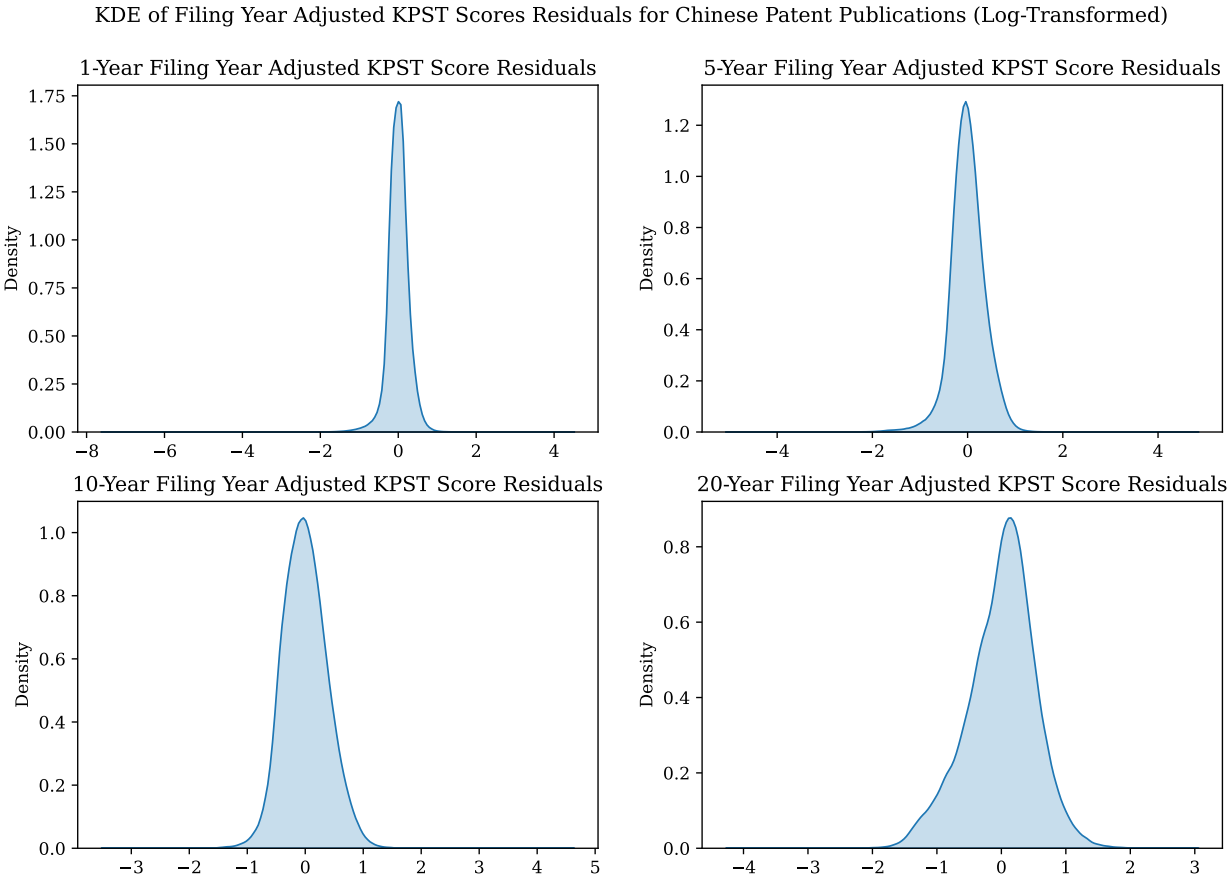
Notes: Citations per patent for Chinese patent publications through September 2023, by grant year. The numbers above each bar represent the mean and the standard deviation (in parentheses) of total citations in each group.

Figure 9: Time Path of Citations



Notes: Citation lags for Chinese patent publications, relative to last publication data (for patents 10+ years old with at least one citation).

Figure 10: Distribution of Adjusted KPST Scores for Chinese Patents



Notes: Distribution of adjusted log-transformed KPST scores for Chinese patents. To account for factors such as language shifts and the rapid expansion in patent filings over time, which could mechanically influence raw KPST scores, we adjust the log-transformed KPST scores by removing patent cohort filing year fixed effects following Kelly et al. (2021). The panels display the distributions of adjusted log-transformed KPST scores over different horizons (i.e., 1-year, 5-year, 10-year, and 20-year for forward similarity measures.

Table 1: Summary Statistics for Chinese Patent Data

Panel A: Patents by Assignee Country		
Assignee Country	**Count**	**Percentage**
CN	14,106,864	85.81
Cross-Filed in the US	483,704	3.51
DE	79,361	0.48
JP	566,388	3.45
KR	277,468	1.69
Other	203,822	1.24
US	1,205,632	7.33
Panel B: Patents by Publication Type		
Publication Type	**Count**	**Percentage**
A	10,813,010	61.31
B	6,106,188	34.62
C	718,142	4.07
Panel C: Top Ten Assignees of Chinese Patents		
Assignee Name	**Count**	**Percentage**
HUAWEI TECH	98,070	0.60
ZTE	55,139	0.34
ZHEJIANG UNIV	50,350	0.31
SAMSUNG ELECTRONICS	47,712	0.29
STATE GRID CORP OF CHINA	42,222	0.26
TSINGHUA UNIV	38,540	0.23
TENCENT TECH	38,069	0.23
SOUTHEAST UNIV	32,775	0.20
TIANJIN UNIV	31,930	0.19
SOUTH CHINA UNIV OF TECH	31,365	0.19
Panel D: Patents by Assignee Sector, China-Assigned Patents Only		
Assignee Sector	**Count**	**Percentage**
COMPANY	8,618,815	62.61
CN VC-Backed	343,158	2.49
US VC-Backed	3,549	0.03
State-Owned Enterprise	349,066	2.54
Other	7,923,042	57.55
GOV	305,461	2.22
Army	60,239	0.44

Table 1, Continued

Assignee Country	**Count**	**Percentage**
other	245,222	1.78
HOSPITAL	58,020	0.42
INDIVIDUAL	1,932,829	14.04
NON-PROFIT	305,461	2.22
UNIVERSITY	2,428,463	17.64
UNKNOWN	117,244	0.85

Panel E: Continuous Variables (All Patents)

Variable	**Mean**	**Median**	**Standard Deviation**
Grant Year	2017.49	2019.00	5.05
File Year	2015.90	2017.00	5.63
Total Cites	1.18	0.00	12.22

Panel F: Continuous Variables (China-Assigned Patents Only)

Variable	**Mean**	**Median**	**Standard Deviation**
Grant Year	2017.97	2019.00	4.49
File Year	2016.70	2018.00	4.88
Total Cites	0.24	0.00	4.22

Notes: The table reports the key summary statistics for Chinese patent data. Panel A reports the summary of patent counts by assignee countries. Among those patents granted to assignees associated with Chinese entities, we also report the number (and percentage) of patents cross-filed in the US. Panel B reports the summary of patent counts by publication type. Panel C reports the patenting activity of the top ten assignees in our data. Panel D reports the decomposition of patent counts by assignee sectors, in which we only focus on the patents with Chinese assignees. Panels E and F report the summary statistics of the other key variables, including grant year, file year, and total cites, for all Chinese patents and the sample, including patents assigned to Chinese entities only.

A Database Documentation

This section contains a codebook (a description of the included variables) for each of the four publicly released datasets.

A.1 Dataset 1: cn_pats.csv

Variable	Description
grantid	Unique identifier for a patent
grantdate	Patent award date
filingdate	Patent filing date
appnr	Application number
appid	Application ID (EPO identifier)
docdb_fam_id	DOCDB family ID
pat_in_us	Dummy equal to 1 if filed in the U.S.
citations_examiner	Citations attributed to examiners
citations_non_examiner	Citations not attributed to examiners
citations_total	Sum of examiner and non-examiner citations
kpst_fw_01	1-year forward similarity score
kpst_fw_25	2-5-year forward similarity score
kpst_fw_610	6-10-year forward similarity score
kpst_fw_1120	11-20-year forward similarity score
kpst_bw_5	5-year backward similarity score
kpst_score_1	1-year KPST score
kpst_score_5	5-year KPST score
kpst_score_10	10-year KPST score
kpst_score_20	20-year KPST score
cpc_ipc	Four-digit technological classification

A.2 Dataset 2: cn_pats_assignees_inventors.csv

Variable	Description
grantid	Unique identifier for a patent
grantdate	Patent award date
filingdate	Patent filing date
appnr	Application number
appid	Application ID (EPO identifier)
assignee_name	Name of assignee (in English); "Missing" if missing
inventor_name	Name of inventor (in English); "Missing" if missing
assignee_name_orig	Original assignee name
inventor_name_orig	Original inventor name
assignee_id	Constructed unique identifier of assignees
inventor_id	Constructed unique identifier of inventors
source	Source of assignee and inventor data (EPO or CNIPA/Google)
assignee_sector	Sector of assignee
assignee_country	Country of assignee
assignee_sector_impute_flag	Dummy equal to 1 if assignee sector imputed
assignee_country_impute_flag	Dummy equal to 1 if assignee country imputed
assignee_pla	Dummy equal to 1 if assignee tagged as People's Liberation Army
assignee_soe	Dummy equal to 1 if assignee tagged as SOE
assignee_soe_match	Matched SOE name if assignee_soe equal to 1
assignee_vc	Dummy equal to 1 if assignee ever backed by VC
assignee_vc_first_deal	Year of first VC deal received by assignee

A.3 Dataset 3: map_cn_pats.csv

Variable	Description
grantid	Identifier for a patent
grantdate	Patent award date
filingdate	Patent filing date
appnr	Application number
grantidfinal	Corresponding ID in the final dataset (cn_pats.csv)
grantdatefinal	Grant date of corresponding final ID
filingdatefinal	Filing date of corresponding final ID

A.4 Dataset 4: map_cn_pats_assignees_inventors.csv

Variable	Description
grantid	Unique identifier for a patent
grantdate	Patent award date
filingdate	Patent filing date
appnr	Application number
appid	Application ID (EPO identifier)
grantidfinal	Corresponding ID in the final dataset (cn_pats.csv)
grantdatefinal	Grant date of corresponding final ID
filingdatefinal	Filing date of corresponding final ID
assignee_name_orig	Original assignee name
inventor_name_orig	Original inventor name
source	Source of assignee and inventor data (EPO or CNIPA/Google)

B Exhibits

Table B.1: Representative Patenting Entities Associated with Lenovo

Assignee	Assignee Name Cleaned
LENOVO (BEIJING) LTD.	LENOVO
LENOVO (BEIJING) CO., LTD.	LENOVO
LENOVO (BEIJING) LIMITED	LENOVO
LENOVO(BEIJING)CO.,LTD.	LENOVO
LENOVO (BEIJING) LIMITED.	LENOVO
LENOVO BEIJING LTD	LENOVO
LENOVO BEIJING LTD.	LENOVO
LENOVO BEIJING CO., LTD.	LENOVO
LENOVO BEIJING CO.,LTD.	LENOVO
LENOVO (BEIJING) CO.,LTD.	LENOVO
LENOVO BEIJING CO LTD	LENOVO
LENOVO CO LTD	LENOVO
LENOVO PRIVATE CO LTD	LENOVO
LENOVO	LENOVO
LENOVO (BEIJING) CO., LTD	LENOVO
LENOVO INC.	LENOVO
LENOVO(BEIJING) CO., LTD.	LENOVO
LENOVO PRIVATE CO., LTD.	LENOVO
LENOVO (SHANGHAI) CO., LTD.	LENOVO
SHANGHAI LENOVO CO., LTD.	LENOVO
LENOVO SHANGHAI CO LTD	LENOVO
LENOVO (SHANGHAI) CO., LTD	LENOVO
LENOVO (BEIJING) CO.,LTD	LENOVO

Table B.2: Illustration of PitchBook Matching

Assignee	Assignee Name Cleaned	Pitchbook Name
AIKANG MEDTECH CO., LTD.	AIKANG MEDTECH	AIKANG MEDTECH
AIKE CORP.	AIKE	AIKE
AILEKE NEW MATERIALS (FOSHAN) CO., LTD.	AILEKE NEW MATERIALS	AILEKE TECHNOLOGY
AIMAN TECHNOLOGY (BEIJING) CO., LTD.	AIMAN TECH	AIMAN DATA
AIMER CO LTD	AIMER	AIMER
AIMER CO., LTD	AIMER	AIMER
AIMER CO., LTD.	AIMER	AIMER
AIMER CORP	AIMER	AIMER
AIMING TECHNOLOGY CO., LTD.	AIMING TECH	AIMING MED
AIMORE ACOUSTICS INC.	AIMORE ACOUSTICS	AIMORE ACOUSTICS
AINEMO INC.	AINEMO	AINEMO
AIR MASTER (GUANGZHOU) LTD.	AIR MASTER	FMEA MASTER
AIRDOC LLC	AIRDOC	AIRDOC
AIROBOTICS LTD	AIROBOTICS	AIROBOTICS
AIROBOTICS LTD.	AIROBOTICS	AIROBOTICS
AIRUISI CO., LTD.	AIRUISI	AIRUISI
AITEN CORP	AITEN	AITEN ROBOT
AITOS AS	AITOS	AITOS
AIVD BIOTECH INC.	AIVD BIOTECH	AIVD BIOTECH
AIWAYS AUTO (SHANGHAI) CO., LTD.	AIWAYS AUTO	AIWAYS
AIWAYS AUTO CO., LTD.	AIWAYS AUTO	AIWAYS
AIWAYS MOTOR CO., LTD.	AIWAYS MOTOR	AIWAYS
AKESO BIOPHARMA INC	AKESO BIOPHARMA	AKESO BIOPHARMA
AKESO BIOPHARMA INC.	AKESO BIOPHARMA	AKESO BIOPHARMA
AKESO BIOPHARMA, INC.	AKESO BIOPHARMA	AKESO BIOPHARMA
AKSO HEALTH TECHNOLOGY (GUANGZHOU) CO., LTD.	AKSO HEALTH TECH	AKSO BIOTECH
ALAUDA, INC	ALAUDA	ALAUDA
ALIBABA (CHINA) CO., LTD.	ALIBABA CHINA	ALIBABA GROUP
ALIBABA CHINA CO LTD	ALIBABA CHINA	ALIBABA GROUP
ALIBABA GROUP CO LTD	ALIBABA GROUP	ALIBABA GROUP

Table B.3: Sample People's Liberation Army-Linked Assignees

	Assignee	Sector
0	63653 FORCES, PLA	GOV NON-PROFIT
1	91049 UNIT OF THE PLA	GOV NON-PROFIT
2	AIR FORCE ENGINEERING UNIV. OF PLA	UNIVERSITY
3	CHINESE PLA GENERAL HOSPITAL	HOSPITAL
4	CHINESE PLA JINAN MILITARY REGION JOINT LOGISTICS DEPARTMENT DRUG & EQUIPMENT TESTING LABORATORY	UNKNOWN
5	INFORMATIZATION GUANGZHOU REGION MILITARY DELEGATE OFFICE OF CHINA PLA GENERAL POLITICAL DEPARTMENT	UNKNOWN
6	INSTITUTE OF PHARMACOLOGY AND TOXICOLOGY ACADEMY OF MILITARY MEDICAL SCIENCES P.L.A. CHINA	GOV NON-PROFIT
7	LOGISTICAL ENGINEERING UNIVERSITY OF PLA	UNIVERSITY
8	MILITARY TRANSPORTATION UNIVERSITY, PLA	UNIVERSITY
9	NATIONAL UNIVERSITY OF DEFENSE TECHNOLOGY OF PLA	UNIVERSITY
10	NATIONAL UNIVERSITY OF DEFENSE TECHNOLOGY, PLA	UNIVERSITY
11	NAVAL UNIVERSITY OF ENGINEERING, PLA	UNIVERSITY
12	NAVY MEDICINE RESEARCH INSTITUTE OF PLA	GOV NON-PROFIT
13	NO.63971 TROOPS, PLA	GOV NON-PROFIT
14	PEOPLE'S LIBERATION ARMY, ORDNANCE ENGINEERING COLLEGE	UNIVERSITY
15	PLA DEFENSE INFORMATION SCHOOL	UNIVERSITY
16	PLA UNIVERSITY OF SCIENCE AND TECHNOLOGY	UNIVERSITY
17	SECOND MILITARY MEDICAL UNIVERSITY, PLA	UNIVERSITY
18	THE ENGINEERING ACADEMY OF ARMORED FORCES OF THE PEOPLE'S LIBERATION ARMY	UNIVERSITY
19	THE FIRST AFFILIATED HOSPITAL OF THIRD MILITARY MEDICAL UNIVERSITY OF PLA	UNIVERSITY
20	THE FOURTH MILITARY MEDICAL UNIVERSITY OF THE CHINESE PEOPLE'S LIBERATION ARMY	UNIVERSITY

C Patenting Entities Incorrectly Identified as Venture-Backed by PitchBook

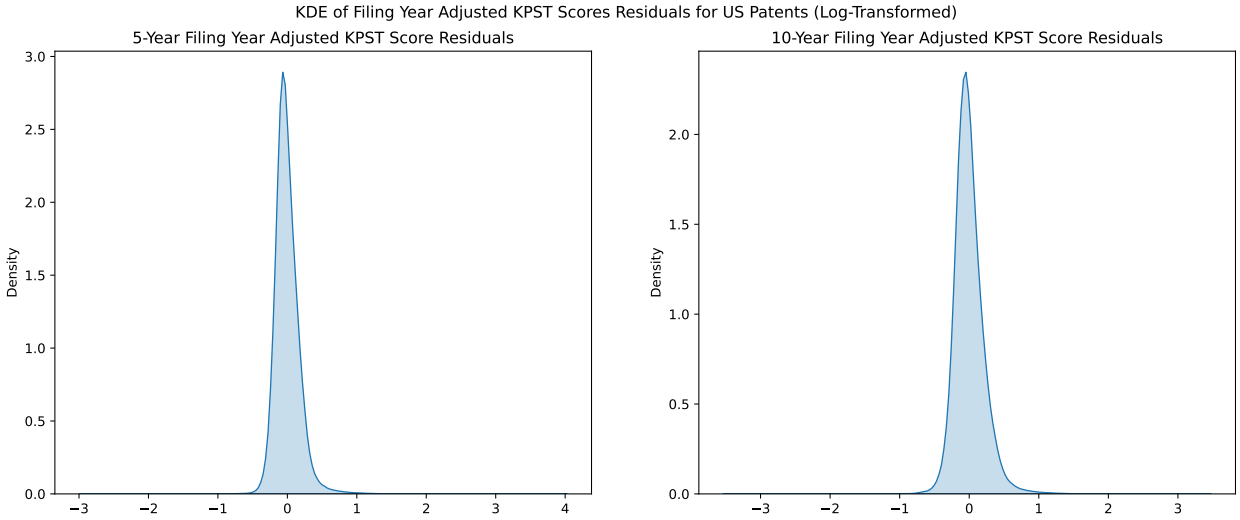
AIR FORCE ENG UNIV OF PLA
AVIC BEIJING AERONAUTICAL MFG TECH RES INST
BAI JIANMIN
BEIJING BOE OPTOELECTRONICS TECH
BEIJING CHEM UNIV
BEIJING EASTWELL SMART TECH
BEIJING FRIENDSHIP HOSPITAL AFFILIATED TO CAPITAL MED UNIV
BEIJING FRIENDSHIP HOSPITAL CAPITAL MED UNIV
BEIJING INFORMATION SCI & TECH UNIV
BEIJING INST OF TECH
BEIJING UNIV OF SCI & TECH INFORMATION
BENQ P
BOE TECH
BRIDGESTONE P
CHANGCHUN INST OF APPLIED CHEM CHINESE ACAD OF SCI
CHANGCHUN INST OF APPLIED CHEM CHINESEACADEMY OF SCI
CHANGCHUN INST OF APPLIED CHEMISTRYCHINESE ACAD OF SCI
CHANGSHA ZOOMLION HEAVY IND SCI & TECH DEV
CHANGZHOU UNIV
CHINA TOBACHUNAN IND
CHONGQING RONGHAI NAT ENG RES CENT OF ULTRASONIC MEDICINE
DAYE NONFERROUS DESIGN & RES INST
EAST CHINA UNIV OF SCI & TECH
ENN R&D
ENN TECH DEV
ESSILOR INT PAGNIE GEN DOPTIQUE
FUKU PRECISION PONENTS SHENZHEN
HISUN PHARM NANTONG
HU SHAOJING
HUAZHONG UNIV OF SCI & TECH
HUNAN CHINA TOBACINDUSTRY
INNOLUX DISPLAY
INST OF ACOUSTICS CHINESE ACAD OF SCI
INST OF METAL RES CHINESE ACAD OF SCI
INST OF PROCESS ENG CHINESE ACADEMYOF SCI
INST OF TELECOM TRANSMISSION MIN OF INFORMATION TECH & IND
INST OF TELECOM TRANSMISSION MIN OF INFORMATION TECH & TELECOM IND
JIANGSU UNIV
JIANGSU YINCHUNBIYA TEA INST
JILIN UNIV

JINAN UNIV
JUSHI
KANG XINSHAN
LENOVO BEIJING
LI FEIYU
LI GUANGWU
LI JIALIN
LI XULIANG
LI YAN
LIU QUANWU
LO PHAM
LU JIWEN
LUO MENGMING
NANCHANG UNIV
NINGBO UNIV
NINGBO UNIV OF TECH
PETROCHINA PANY
PING AN TECH SHENZHEN
PU TING
QUNKANG TECH SHENZHEN
SANFORD BURNHAM MED RES INST
SANGDIYA MEDICINE TECH SHANGHA
SHANDONG NHU PHARM
SHANDONG UNIV OF TECH
SHANGHAI INST MATERIA MEDICA CAS
SHANGHAI INST OF MATERIA MEDICA CHINESE ACAD OF SCI
SHANGHAI INST OF MATERIA MEDICACHINESE ACAD OF SCI
SHANGHAI INST OF MICROSYSTEM & INFORMATION TECH CHINESE ACAD OF SCI
SHANGHAI UNIV
SHANGHAI YAOGU BIOMEDICAL INNOVATION RES INST
SHANGYU NHU BIOCHEMICAL IND
SHAO YIMING
SHENZHEN BGI RES INST
SHENZHEN GOODIX TECH HLDGS
SHENZHEN HUADA LIFE SCI RES INST
SHENZHEN TCL NEW TECH
SHENZHEN UNIV
SHENZHEN YOUBTECH TECH
SICHUAN UNIV
SOUTHEAST UNIV
TANG CHUANBIN
TANG XIAOOU
TCL
THE FIRST AFFILIATED HOSPITAL OF THE THIRD MILITARY MED UNIV OF THE CHINESE
THE FIRST AFFILIATED HOSPITAL OF THIRD MILITARY MED UNIV OF PLA

THE UNIV OF ARIZONA
TIAN LI
TIANJIN INT JOINT RES INST OF BIOMEDICINE
TSINGHUA TONGFANG NUCTECH ; TSINGHUA UNIV
TSINGHUA UNIV
TSINGHUA UNIV ; TSINGHUA TONGFANG WEISHI TECH
TYELECTRONICS
TYELECTRONICS DONGGUAN
TYELECTRONICS P
TYELECTRONICS PORATION
TYELECTRONICS SHANGHAI
TYELECTRONICS SHENZHEN
UNIV OF ELECTRONIC SCI & TECH
UNIV OF ELECTRONIC SCI & TECH OF CHINA
UNIV XIAMEN
XIA NAN
XIAMEN UNIV
YU LUOJIA
ZENG MIN FRANK
ZHANG XIANGMIN
ZHEJIANG A & F UNIV
ZHEJIANG AGRIC & FORESTRY UNIV
ZHEJIANG HISUN PHARM
ZHEJIANG NHU
ZHEJIANG NHU PORATION
ZHEJIANG UNIV
ZHEJIANG UNIV OF TECH
ZHEN DING TECH HLDGS
ZHOU JIA
ZHOU XING
ZOOMLION HEAVY IND SCI & TCHNOLOGY DEV
ZOOMLION HEAVY IND SCI & TECH DEV

D Additional Figures

Figure D.1: Distribution of Adjusted KPST Scores for US Patents



Notes: Distribution of adjusted log-transformed KPST scores for US patents between 1985 and 2022. To account for factors such as language shifts and the rapid expansion in patent filings over time, which could mechanically influence raw KPST scores, we adjust the log-transformed KPST scores by removing patent cohort filing year fixed effects following Kelly et al. (2021). The panels display the distributions of adjusted log-transformed KPST scores over different horizons (i.e., 5-year, 10-year for forward similarity measures).