Creating the China Patent Dataset

Josh Lerner, Namrata Narain, Dimitris Papanikolaou, and Amit Seru[1]

Preliminary and Incomplete

November 30, 2024

[1] Harvard University; Harvard University; Kellogg 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, Smooth Richardson Foundation, Stanford Graduate School of Business, and the U.S. National Science Foundation (Award no. 2241237) for financial support. Numerous research assistants helped in this effort. Seminar participants at Harvard Business School provided helpful comments, as did a number of practitioners. 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

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 (Lerner et al., 2024) 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.

The modern Chinese patent law was enacted 4th Meeting of the Standing Committee of the Sixth National People's Congress on March 12, 1984.[2] We examine patents from the inception of patent grants in September 1985 through the time that the data were downloaded in September 2023.

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).

In large part, this paper follows a methodology like that 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 (Alcacer and Gittelman, 2006), the calculation of measures of textual novelty (Kalyani et al., 2024; Kelly et al., 2021), the construction of international patent families (Putnam, 1996), 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). The key features of the database are summarized in Appendix A.

But 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. While in some cases these are the publication of an initial application and a single award, in other cases, more complex patterns appear.
- The tendency of subsequent Chinese patents to cite the first publication of a patent application, rather than the final award.

[2] 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).

There has been a growing literature about innovation in China. Influential works 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), and Beraja et al. (2023). But to date, efforts to catalog the Chinese patent data along the lines of the various efforts at the NBER and elsewhere have been infrequent. For instance, 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]

This document describes the process by which we constructed this dataset, the key issues encountered, and the choices made. In the second section, we provide a few stylized facts to motivate the importance of understanding innovation in China. In Section 3, we describe how we assemble the universe of Chinese patents from the China National Intellectual Property Administration (CNIPA), European Patent Office (EPO), and Google Patent databases. The fourth section summarizes the process of cleaning and sorting assignees. Section 5 discusses the linking of these patents to U.S. awards. The next section focuses on creating two measures of patent quality, citations and Kelly et al. (2021) measures of textual novelty. In Section 7, we presents some key facts about the dataset.

2. The Backdrop

Innovation outside the major industrialized nations has become increasingly important in recent years. This is particularly true in China. These trends are illustrated in the two charts discussed below. This section is simply intended to provide some motivation for the construction of the database, and not review the growing literature on Chinese innovation.

Figure 1, using data from the World Intellectual Property Organization (https://www3.wipo.int/ipstats/key-search/indicator), reports the annual volume of domestic patent applications in national patent offices between 2001 and 2022. The countries shown are the top five developed nations (Japan, US, Korea, Germany, and France) and top five emerging nations (China, Russia, India, Iran, and Brazil), both as measured by total domestic patent applications in this period.

[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.

The table makes clear the substantial disparity in the volume of patenting across nations (note the use of the log scale). 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.

Figure 2 looks at aggregate venture capital activity. This, taken from Lerner et al. (2024), 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).

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. We focus our initial efforts on three datasets,[5] the EPO's PATSTAT Global dataset, the CNIPA database, and the Google Patents database.

A. Institutional Details

Before doing so, 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).[6]

[5] 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.

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

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).[7] 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'home, 2012; Fuller, 2016). Perhaps reflecting this, many more patent publications in China are applications that are never issued.

These institutional considerations suggests 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. 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:

1. Application number
2. Publication number
3. Publication type
4. Filing date
5. Grant date

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.

Put another way, a grant ID is publication number x publication type pair.

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."

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

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 (see examples of this problem below). Therefore, publication numbers themselves do not uniquely identify patents:

- First, applications and granted documents before April 2010 were published with different numbers.[8] 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. 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.

Our construction of the dataset will have to grapple with all these disparities.

B. 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. 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.”)

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.

[8] *Ibid.*

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. Fourth, we link the patents to their set of assignees.

We now briefly describe each of these datasets in turn.

EPO IDs

- Filename: all_cn_ids.csv
- Variables: APPLN_ID, PAT_PUBLN_ID, PUBLN_NR, PUBLN_AUTH, PUBLN_DATE, APPLN_AUTH, APPLN_NR, APPLN_FILING_DATE, DOCDB_FAMILY_ID. [9]

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.

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.

[9] 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.

- 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.

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 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 abstracts 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.

Note that there are more rows than either PUBLN_NR or APPLN_NR because each PUBLN_NR x APPLN_NR pair (at least from 2010 and after) can show up multiple times, corresponding to various PUBLN_KIND codes. This means, multiple publications of the same patent show up in the dataset. The dataset is unique at the level of the PUBLN_NR x APPLN_NR x PUBLN_DATE triple.

Having constructed this dataset with the EPO IDs, we now associate them with information on assignees. To do so, we combine two additional sets of EPO data field. The first set contain the map between application IDs and person IDs (that identify assignees.) The second set contains information on the assignees, including their sector (e.g., corporation, university, hospital, …) and country in which they are based. This former variable, PSN_ASSIGNEE, has been determined independently by the ECOOM Centre for Research and Development Monitoring at KU Leuven, and added by the EPO subsequently to the database.

We add:

- Filename 1: patent_data_1985_2000 / cn_1990_09_app_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
    - We use the HAN_ID and HAN_NAME since these variables were created by the Organisaiton for Economic Cooperation and Development (OECD) using the Orbis business register.[10]

Application IDs are often linked to multiple assignees. We only keep rows where INVT_SEQ_NR == 0, which signifies that the assignee is *not* the inventor. In sum, 16,245,435 applications are assigned to at least one assignee. This accounts for 98% of the patents in the sample.

C. *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:

1. 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.
2. We type in "CN" as a required keyword for patent ID.
3. We set the range of issue date for intended years (1985-2023).

In total, we accessed patents with 16,138,878 unique grant IDs and 16,138,655 unique application IDs.

D. *Merging the Two Datasets*

[10] 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.

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

Foremost among these are inaccuracies in the both databases. First, there are typos or incorrect values in the application number variable. Each patent is linked to an application number. While it seem 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. 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. On looking up this patent on patent.google.com, we find that the PDF of this patent has the same application number as in the EPO data (i.e., 96106116). 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.[11] 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.

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.

[11] 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 the 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 9 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 which were published multiple times on the same day. An example is the application number 200910125308, which is 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 4 such cases (corresponding to 8 patents.) I 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.

E. 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 also obtained from Google the full text translated into English for all patents.

Google makes clear that these data are accessible for 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; as of 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 Google Patents Public Data page indicates that identical licensing rules hold.[12]

4. Identifying and Disambiguating Assignees

For each patent, we sought to gather the following information:

1. The name of the entity it is assigned to ("assignee").
2. The sector of the entity: company, university, hospital, non-profit, individual, and various combinations thereof.
3. Whether the patent is assigned to a venture capital-backed firm.
4. Whether the patent is assigned to a state-owned enterprise (SOE).
5. Whether the patent is assigned to the People's Liberation Army and those assigned to one of the "Seven Sons of Defense."
6. The assignee's country.

A. Disambiguating Assignees and Completing Characteristics

We received data on assignees from both EPO and CNIPA datasets. As a first step to constructing a patent-assignee-level dataset then, we reconcile information on assignees between the EPO and the CNIPA datasets. There were a few things to note:

1. The CNIPA dataset contained only one assignee per patent. This was reflected in the "current assignee" field on patents.google.com.
2. The EPO dataset contains multiple assignees per patent, corresponding to all assignees that an application was ever assigned to.

[12] The CC 4.0 license explains "You 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" (https://creativecommons.org/licenses/by/4.0/deed.en).

3. The sector of assignee is a variable constructed by the EPO and is as such missing in the CNIPA dataset.

While assignees in the EPO data were coded, assignees in the CNIPA data did not have an associated identifying number. This lack of assignment was 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. Appendix B provides an illustrative lists of the various assignees associated with Lenovo.

We first created an assignee ID for the CNIPA dataset by first ensuring that all assignees with the same name have the same country code. Then, we tagged each unique assignee x assignee country pair with a new ID. All assignees corresponding to one application number were collected. We then merged the EPO and CNIPA assignee datasets as described above. For applications where both CNIPA and EPO have assignee information, we kept the EPO information. This is mostly because 99.5% of the EPO assignees names are in English whereas more than three-quarters of the CNIPA assignee names are in Chinese. We then translate the assignees that are in Chinese using the Google Cloud Translate API.

Next, we cleaned assignee names. We undertook the following steps. First, we cleaned the assignee names: making all assignee names upper case, removing punctuation from all names, converting all characters to ASCII, stemming common words using the mapping provided by the NBER,[13] to ensure there is only one space between each word, and removing leading and trailing spaces from the names.

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

More specifically, for the assignees *not missing* sector assignment, we trained 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. For the assignees missing a sector assignment, we predicted their sector assignment using the trained classifier.

We proceeded similarly for the 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

[13] Some examples of this mapping are as following: TECHNOLOGY: TECH, CORPORATION: CORP, BROEDERNA: BRDR, CENTRO: CENT, CENTRAL: CENT, INTERNATIONAL: INT, etc.

assignees in six nations: China, Hong Kong, Korea, Singapore, Japan and the United States. (Of 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.) 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 impute values for assignees missing country values. In the final dataset, the other major countries of assignees are Germany, France, and Great Britain.

We then did a second round of cleaning assignee names. Note that we cleaned assignee names once before training the XGBoost algorithm to learn sector and country assignments and then once after. This was for two reasons. First, in the final cleanup, we wanted to strip away names of Chinese cities and provinces from entities that are labeled as "COMPANY." This ensured that assignee names such as Lenovo Beijing were cleaned up to be simply Lenovo. But to do this, we first needed to predict the sector for assignees that were missing this information. The XGBoost algorithm to predict the sector worked much better when the assignee names still had organizational identifiers such as Ltd, Corp, LLC, etc. The XGBoost algorithm was also able to predict the assignee's country easier if the organizational tags are present.

For the second round of cleaning, 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.

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
    - 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

B. 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 (Select Committee, 2024). Thus, we wished to identify venture-backed firms.

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 was to define venture-backed patents as those that were applied for within five years of the firm's first round of venture financing.

To do so, we coded patents as VC-backed if they were filed in the first year or the following five years of a company's first venture capital funding. We used PitchBook, which we have argued elsewhere (Lerner, et al., 2024) appear 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 receive at least one deal from an investor whose type is Venture Capital.

We merged the patent assignee dataset with the Pitchbook dataset based on the names of companies. We undertook 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 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 is 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 the letter-based matching, we also ensured the procedure is quite fast, since calculating Levenshtein distance is a computationally slow process. Appendix C shows some illustrative results from this matching process.

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 amongst these matches, so we choose the tianyancha SOE name that is the first when sorted alphabetically.

We examined the top assignees (all those with over one hundred patents) and seek to understand whether they were indeed venture backed firms. The primary sources that we use 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 D. Because of the disambiguation process described in Section 4.A, this has the effect of recoding subsidiaries of these entities as well. We also determine that assignees whose type is HOSPITAL, UNIVERSITY, or COMPANY HOSPITAL are unlikely to be VC-backed themselves. For such assignees, we change the coding of patents to be not VC-backed.

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.)

C. *Identifying State-Owned Enterprises*

We seek to identify patents owned by state-owned enterprises, who have been documented to play a large (and indeed increasing) role in the innovation landscape (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[14]). We cleaned these names with a combination of steps described in Section 4.A, and then merge with the patents dataset using the two-step procedure described in Section 4.B.

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

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

  - INDIVIDUAL: 44
  - COMPANY GOV NON-PROFIT: 30
  - HOSPITAL: 16
  - COMPANY HOSPITAL: 10
  - COMPANY UNIVERSITY: 2
  - COMPANY GOV NON-PROFIT UNIVERSITY: 1

D. *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.

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 Appendix E.

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
- COMPANY UNIVERSITY: 13
- GOV NON-PROFIT UNIVERSITY: 103

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. 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."[15]

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

[15] In May 2020, The Trump Administration cancelled 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. *Identifying Chinese Patents Also Filed in the U.S. and Cross-Citations*

A. U.S. Cross-Filings

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.[16])

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. We identify all Chinese patents with a corresponding patent award in the U.S.

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

B. 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.

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

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.

6. *Measuring Patent Importance*

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 varied across technologies and over time, we normalized the citations by the mean number received by other patents in that four-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 2023.

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 (1989) and others to impede price discovery and efficiency, we do not consider a Kogan et al. style analysis.

A. *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 (identified as described in Section 3.D), 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.

B. Kelly 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).

We proceeded in two parallel efforts to create Kelly scores, based on the English translations and the original Mandarin.

English Kelly scores

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 Kelly scores, please see Kogan, Papanikolaou, Seru, and Stoffman (2017).

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 Kelly 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 Kelly 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.

7. *Assigning a Primary Patent Class*

The USPTO clearly delineates for each patent a primary technology class. These originally used the U.S. Patent Classification scheme, which had approximately 140 thousand subclassifications. The bulk of the world's patent offices, including CNIPA, used the World Intellectual Property Organization's International Patent Classification scheme, which had about seventy subclasses.

Between 2011 and 2015, the USPTO transitioned to the Combined Patent Classification (CPC) scheme, which was jointly developed by EPO and USPTO. With approximately two thousand subclasses, it has a high level of detail. But the scheme mirrors exactly the IPC for and CPC codes at the first four digits (apart from class Y02). Much of the recent research using U.S. patents has focused on using the primary CPC class (which PatentsView has backfilled for earlier patents).

In the Chinese patents, one or more patent classes (using the IPC scheme) is indicated at the top of the patent, with the first one often indicated in bold. In some cases, multiple IPC classes are indicated in bold. There is no documentation that we are aware of that suggests the CNIPA makes a designation akin to the USPTO's primary class.

In all our datasets (EPO, CNIPA, and Google), we observe multiple IPC codes for each patent without any primary code being tagged. To be consistent in assigning a primary IPC code to each patent, we take the modal four-digit IPC class as the primary class. When there are multiple modal classes, we choose one at random.

8. *A First Look*

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, 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.[17]

Figure 6 shows the distribution of venture-backed patentees, by year of their first venture investment: the growth of the Chinese sector, and the subsequent slowdown in the late 2010s, are evident from the data.

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 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).

Figures 8, 9, and 10 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: the older patents have in average 2.2 awards, as opposed to 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, as reported in Table 7. 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).

Fill in Figure 10

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 et al., 2004), the representation of defense-linked assignees, and the greater youth (a mean publication date of

[17] 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.

December 2017 for the Chinese-assigned awards as oppose dot June 2017 for the total sample) of the awards to local entities.

9. 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. (2024) to facilitating merging with other datasets. It is our hope that this will be a resource to scholars worldwide.

References

Aghion, Philippe, Jing Cai, Mathias Dewatripont, Luosha Du, Ann Harrison, and Patrick Legros. 2015. “Industrial Policy and Competition.” *American Economic Journal: Macroeconomics* 7, 1–32.

Alcacer, Juan, and Michelle Gittelman. 2006. “Patent Citations as a Measure of Knowledge Flows: The Influence of Examiner Citations. *Review of Economics and Statistics* 88, 774–779.

Bernstein, Shai, Xavier Giroud, and Richard R. Townsend. 2016. “The Impact of Venture Capital Monitoring.” *Journal of Finance* 71, 1591–1622.

Beraja, Martin, David Y. Yang, and Noam Yuchtman. 2023. “Data-Intensive Innovation and the State: Evidence from AI firms in China.” *Review of Economic Studies* 90, 1701–1723.

Aakash Kalyani, Bloom, Nicholas, Marcela Cavalho, Tarek A. Hassan, Josh Lerner, and Ahmed Tahoun. 2024. “The Diffusion of New Technologie” *Quarterly Journal of Economics*, forthcoming.

Branstetter, Lee G., and Guangwei Li. 2022. “Does ‘Made in China 2025’ Work for China? Evidence from Chinese Listed Firms.” National Bureau of Economic Research Working Paper Series No. 30676, https://www.nber.org/papers/w30676.

Chen, Zhao, Zhikuo Liu, Juan Carlos Suárez Serrato, and Daniel Y. Xu. 2021. "Notching R&D Investment with Corporate Income Tax Cuts in China.” *American Economic Review* 111, 2065–2100.

Cheng, Wenting. 2023. *China in Global Governance of Intellectual Property: Implications for Global Distributive Justice*. London, Palgrave Macmillan.

Fama, Eugene. 1989. “Perspectives on October 1987, or What Did We Learn from the Crash.” In Robert W. Kamphuis, Jr., Roger C. Kormendi, and J.W. Henry Watson, editors, *Black Monday and the Future of Financial Markets*. Homewood, IL: Irwin, pp. 71-82.

Fang, Lily H., Josh Lerner, and Chaopeng Wu. 2017. “Intellectual Property Rights Protection, Ownership, and Innovation: Evidence from China.” *Review of Financial Studies* 30, 2446–77.

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

Griliches, Zvi (editor). 1984. *R&D, Patents, and Productivity*. Chicago: University of Chicago Press for National Bureau of Economic Research.

He, Zi-Lin, Tony W. Tong, Yuchen Zhang, and Wenlong He. 2018. “Constructing a Chinese Patent Database of Listed Firms in China: Descriptions, Lessons, and Insights.” *Journal of Economics and Management Strategy* 27, 579–606.

Hegde, Deepak, Kyle Herkenhoff, and Chenqi Zhu. 2023. “Patent Publication and Innovation.” *Journal of Political Economy* 131, 1845-1903.

Holmes, Thomas J., Ellen R. McGrattan, and Edward C. Prescott. 2015. “Quid Pro Quo: Technology Capital Transfers for Market Access in China. *Review of Economic Studies* 82, 1154-1193.

Jaffe, Adam B., and Manuel Trajtenberg. 2002. *Patents, Citations, and Innovations: A Window on the Knowledge Economy*, Cambridge, MIT Press.

Kelly, Bryan, Dimitris Papanikolaou, Amit Seru, and Matthew Taddy. 2021. “Measuring Innovation over the Long Run.” *American Economic Review: Insights* 3, 303-20.

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

Kőnig, Michael, Kjetil Storesletten, Zheng Song, and Fabrizio Zilibotti. 2022. “From Imitation to Innovation: Where Is All That Chinese R&D Going?” *Econometrica* 90, 1615-54.

Kortum, Samuel, and Josh Lerner. 2000. “Assessing the Impact of Venture Capital on Innovation.” *RAND Journal of Economics* 31, 674–692.

Lamoreaux, Naomi R., and Kenneth L. Sokoloff. 2005. “The Decline of the Independent Inventor: A Schumpeterian Story?,” National Bureau of Economic Research working paper no. 11654.

Lei, Zhen, Zhen Sun, and Brian Wright. 2012. “Are Chinese Patent Applications Politically Driven? Evidence from China’s Domestic Patent Applications.” Unpublished working paper, Organisation for Economic Cooperation and Development.

Lerner, Josh, Junxi Liu, Jacob Moscona, and David Yang. 2024. “Appropriate Entrepreneurship? The Rise of Chinese Venture Capital and the Developing World.” National Bureau of Economic Research working paper no. 32193.

Lerner, Josh, Amit Seru, Nick Short, and Yuan Sun. 2023. “Financial Innovation in the Twenty-First Century: Evidence from U.S. Patents.” *Journal of Political Economy*, forthcoming.

Li, Guan-Cheng, Ronald Lai, Alexander D’Amour, David M. Doolin, Ye Sun, Vetle I. Torvik, Amy Z. Yu, and Lee Fleming. 2014. “Disambiguation and Co-Authorship Networks of the U.S. Patent Inventor Database (1975–2010).” *Research Policy* 43, 941-955.

Magerman, Tom, Bart Van Looy, and Xiaoyan Song. 2006. “Data Production Methods for Harmonized Patent Statistics: Patentee Name Harmonization.” KUL Working Paper No. MSI 0605, https://ssrn.com/abstract=944470.

Myers, Kyle R., and Lauren Lanahan. 2022. "Estimating Spillovers from Publicly Funded R&D: Evidence from the US Department of Energy." *American Economic Review* 112, 2393-23.

Prud'homme, Dan. 2012. *Dulling the Cutting Edge: How Patent-Related Policies and Practices Hamper Innovation in China.* Beijing, European Union Chamber of Commerce in China.

Putnam, Jonathan D. 1996. *The Value of International Patent Rights*. Unpublished Ph.D. dissertation, Yale University.

The Select Committee on the Strategic Competition Between the United States and The Chinese Communist Party, U.S. House of Representative. 2024. *The CCP's Investors: How American Venture Capital Fuels the PRC Military and Human Rights Abuses*. https://selectcommitteeontheccp.house.gov/sites/evo-subsites/selectcommitteeontheccp.house.gov/files/evo-media-document/2024-02-08%20-%20VC%20Report%20-%20FINAL.pdf.

Stoff, Jeffrey, and Glenn Tiffert. 2020. "Under the Radar: National Security Risk in US-China Scientific Collaboration." In Glenn Tiffert, ed., *Global Engagement: Rethinking Risk in the Research Enterprise*. Stanford, Hoover Institution, 2020, 19–104.

Tong, Tony W., Kun Zhang, Zi-Lin He, and Yuchen 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, Shang-Jin, Zhuan Xie, and Xiaobo Zhang. 2017. "From 'Made in China' to 'Innovated in China': Necessity, Prospect, and Challenges." *Journal of Economic Perspectives* 31, 49–70.

Yin, Deyun, Kazuyuki Motohashi, and Jianwei Dang. 2020. "Large-scale name disambiguation of Chinese patent inventors (1985–2016)." *Scientometrics* 122, 765-790.

Zhang, Xiaotao, Ziqiao Wang, Jing Hao, and Feng 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.

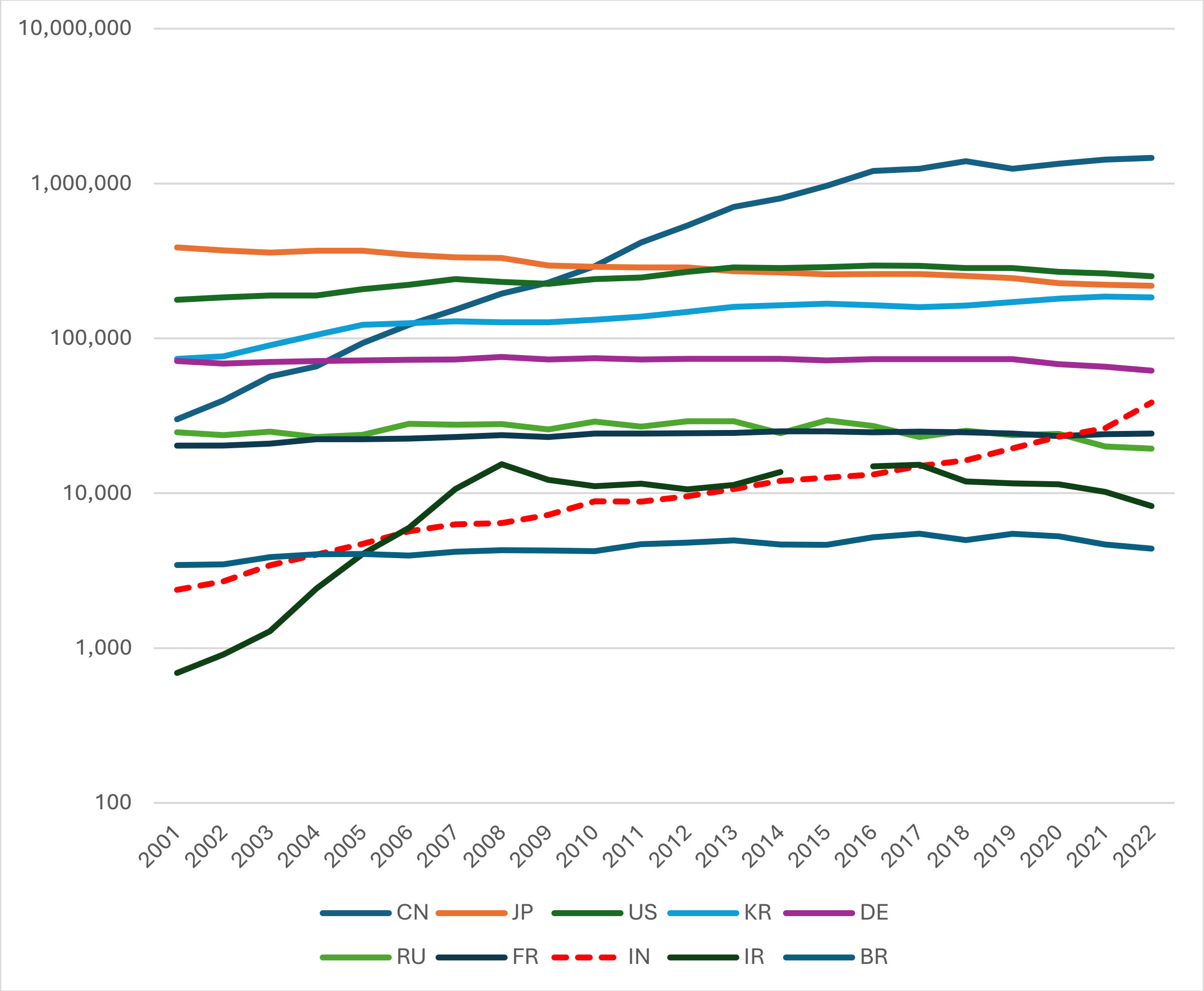


Figure 1. Domestic patent applications in top five developed (Japan, U.S., Korea, Germany, and France) and emerging (China, Russia, India, Iran, and Brazil) economies, based on cumulative patenting in the period, 2001-22. The source is the World Intellectual Property Organization.

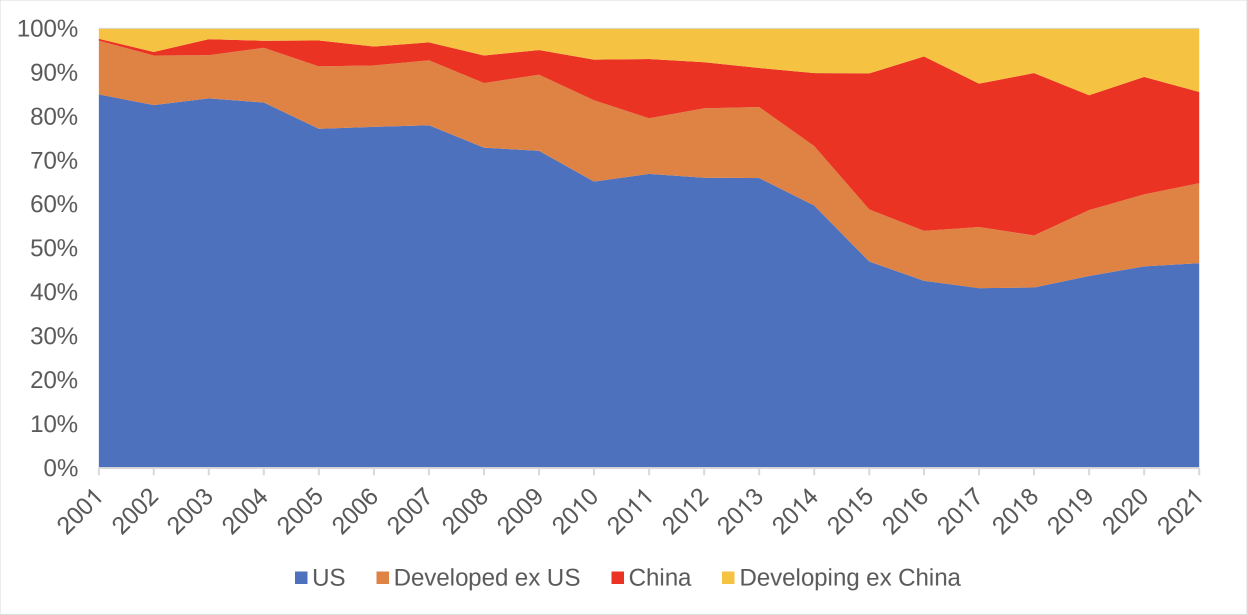


Figure 2. Relative share of global venture capital funding, 2001-21, from Lerner et al. (2024).

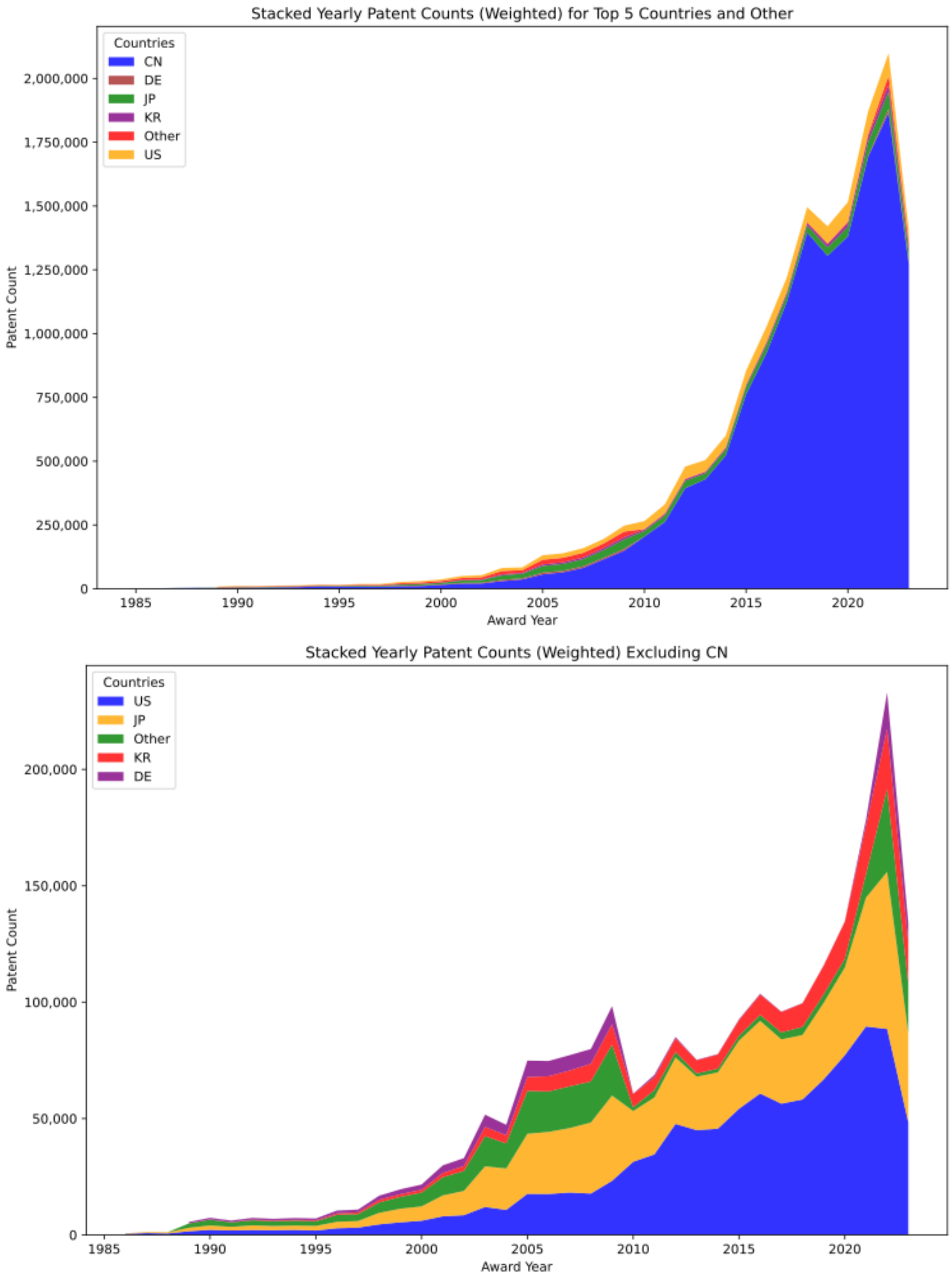


Figure 3: Chinese patent publications, by nation and year. The top panel shows all awards; the lower, all awards excluding China. Patents with multiple assignees are assigned proportionately to their nationality.

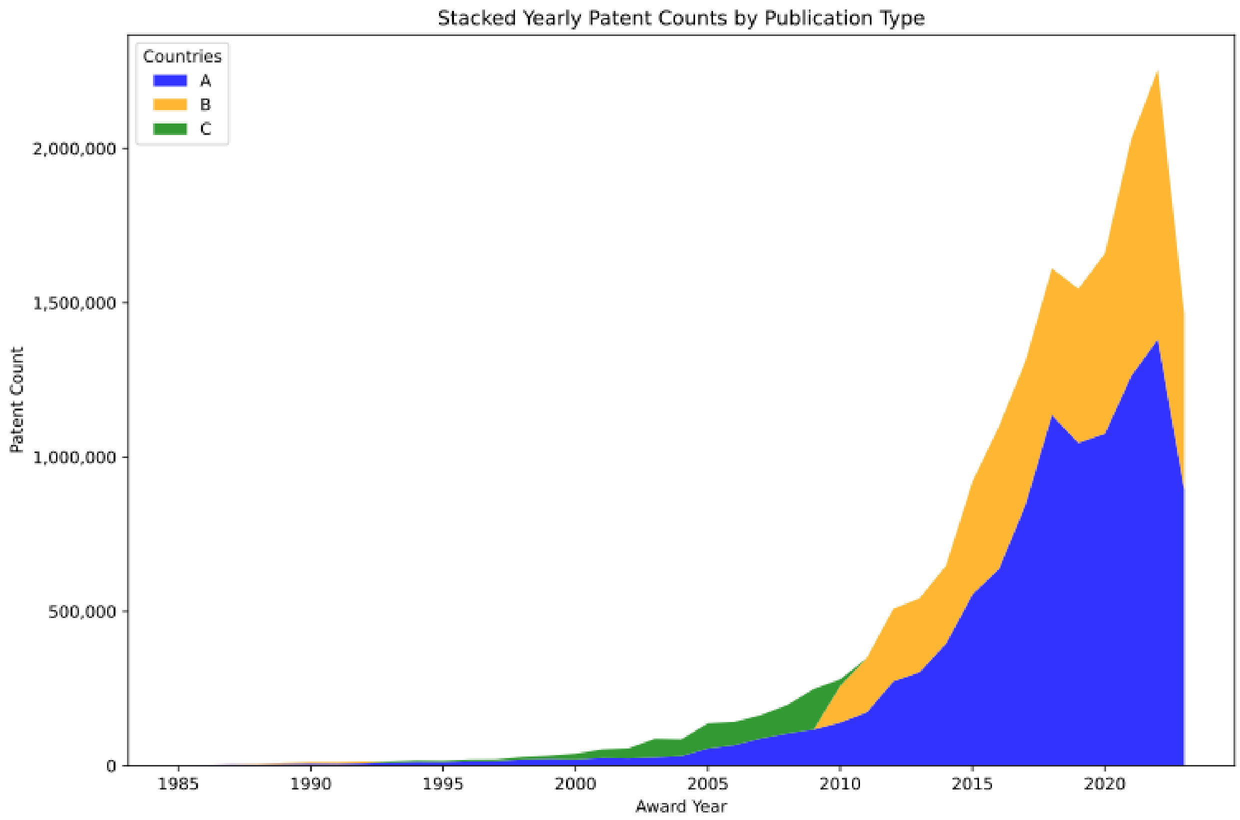


Figure 4: Chinese patent publications by patent type and year.

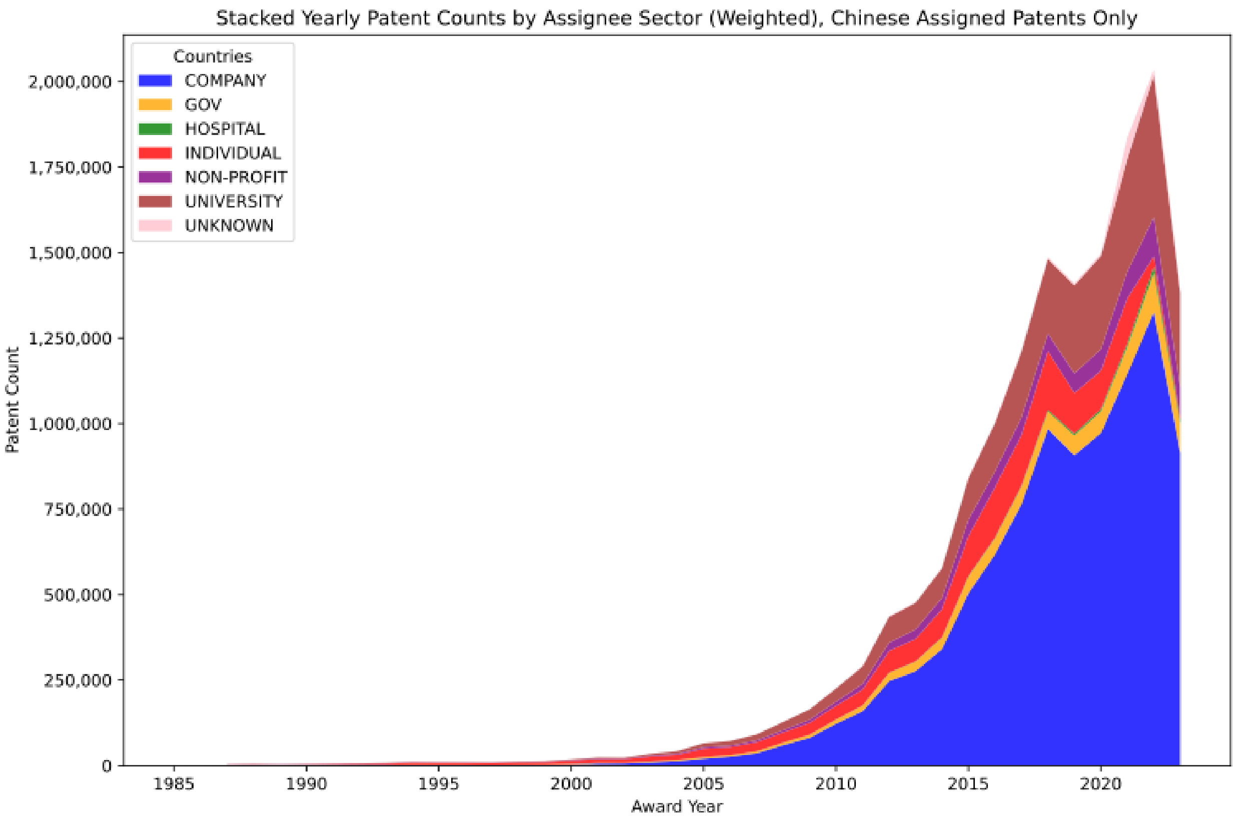


Figure 5: Chinese-assigned Chinese patent publications, by assignee sector and year. Patents with multiple assignees are assigned proportionately to their sector.

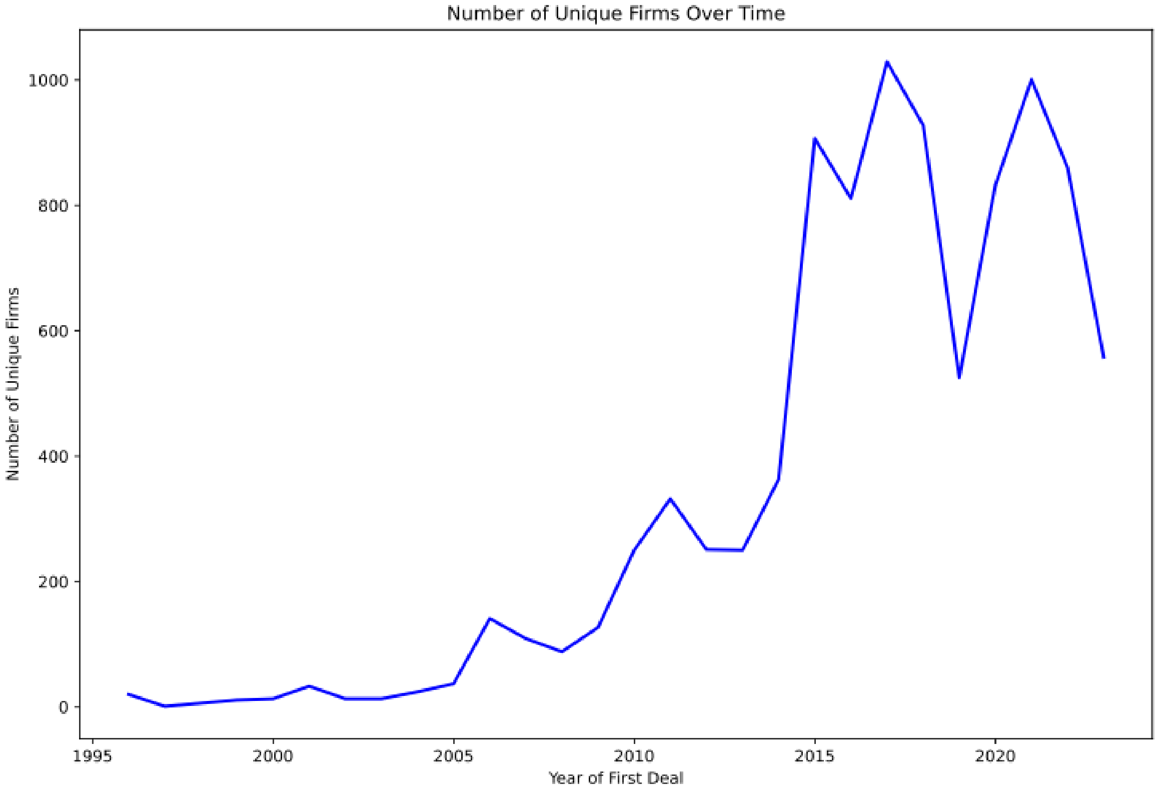


Figure 6: Venture-backed Chinese assignees in Chinese patent publications, by year of first venture capital investment.

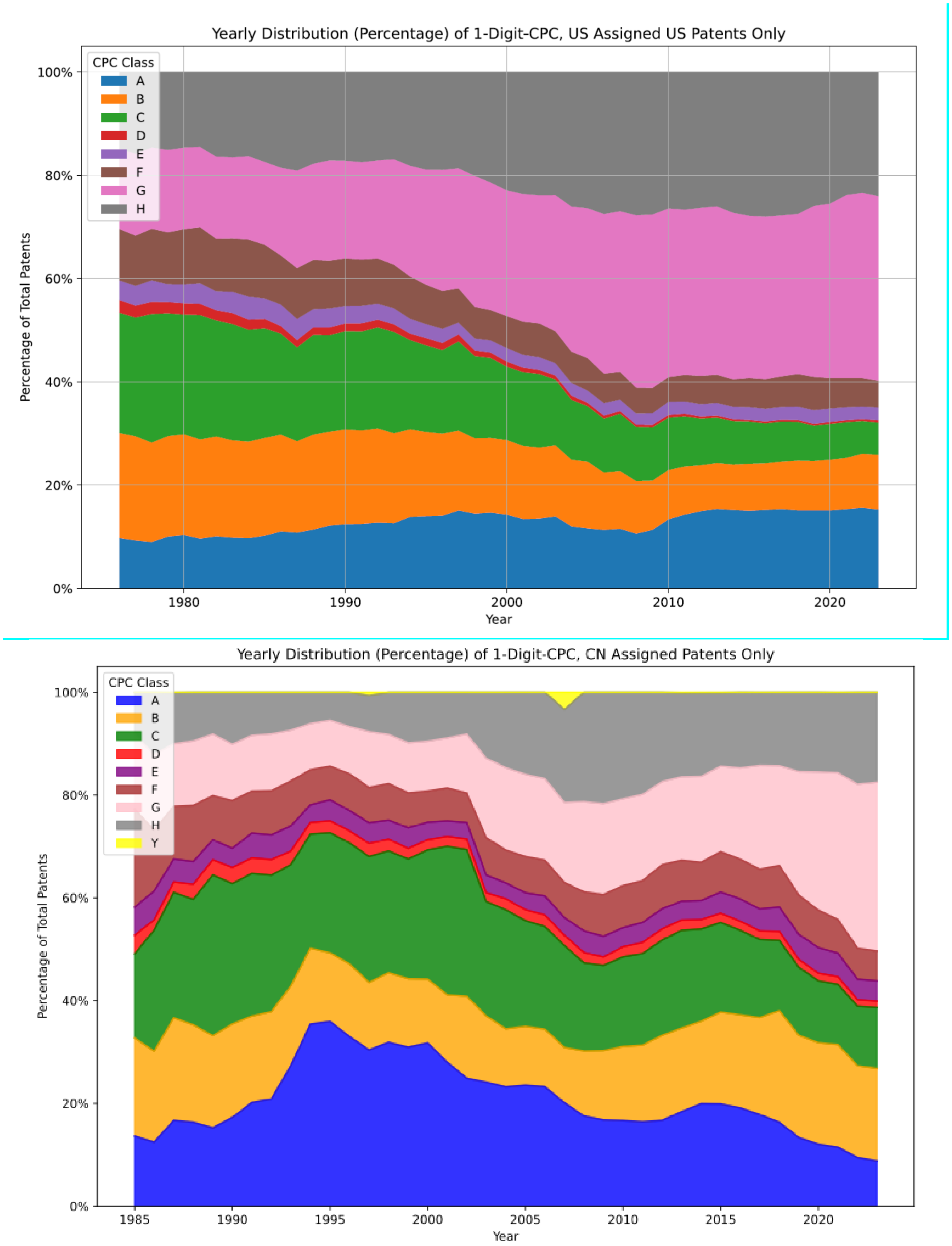


Figure 7: Distribution of one-digit primary IPC/CPC classes, by year. The top panel is the breakdown of Chinese-assigned CNIPA patents; the lower panel, U.S.-assigned USPTO patents.

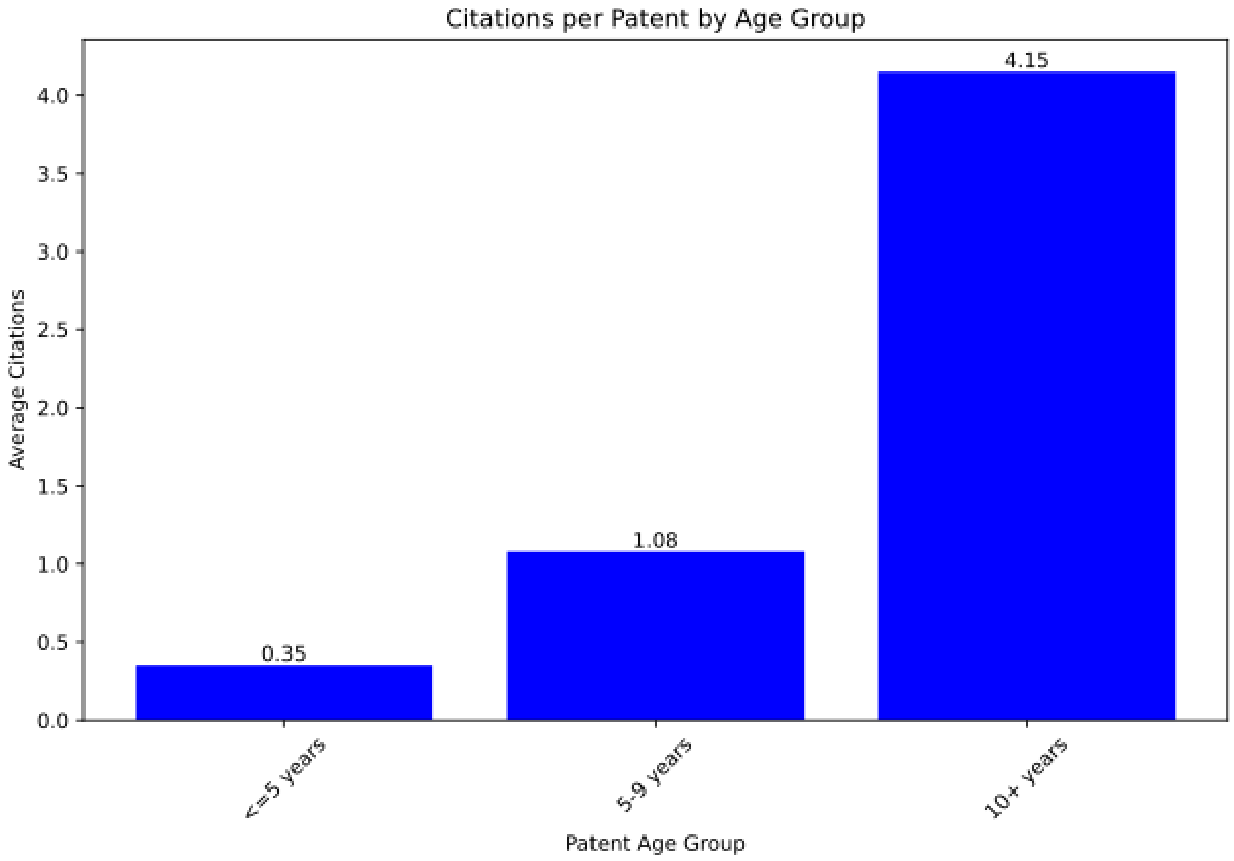


Figure 8: Citations per patent for Chinese patent publications through September 2023, by age of award in September 2023.

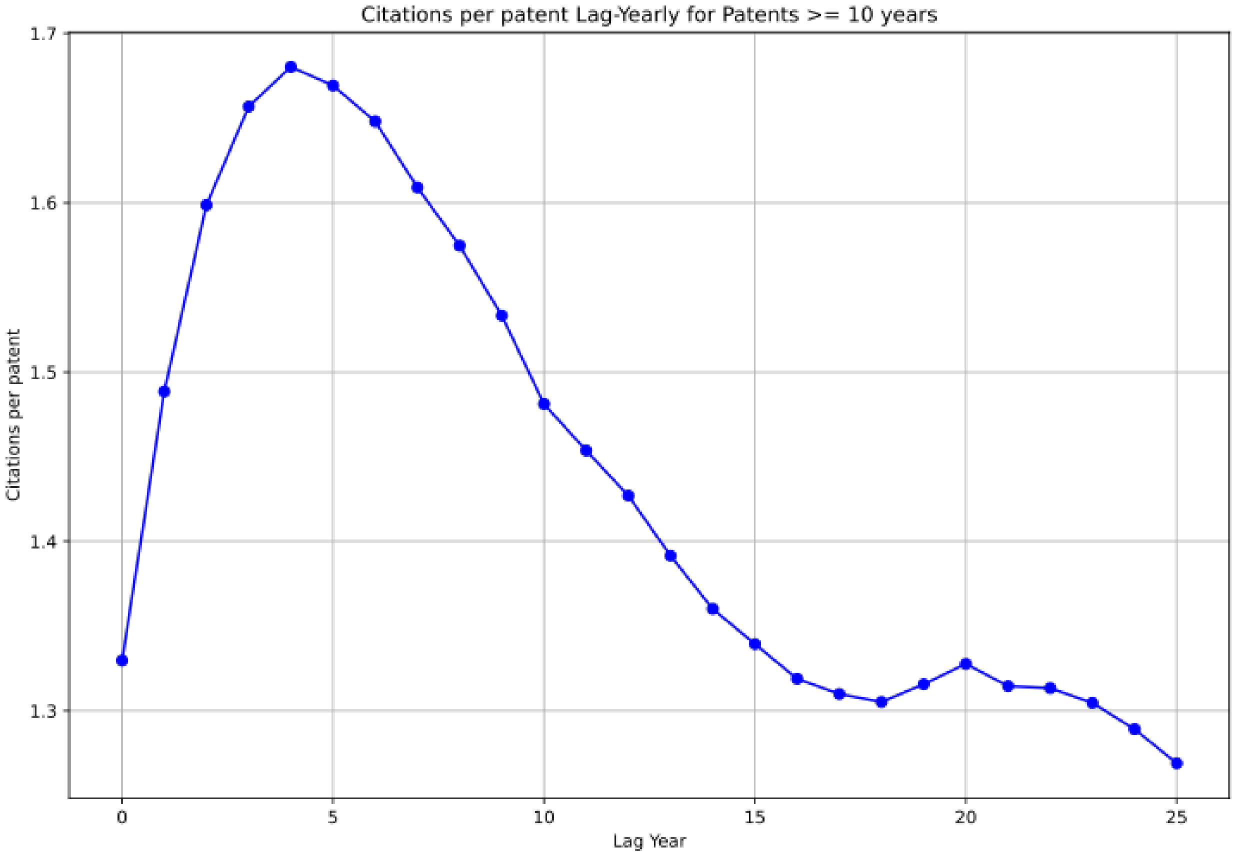


Figure 9: Citation lags for Chinese patent publications, relative to last publication data (for patents 10+ years old with at least one citation).

Figure 10: Distribution of Kelly scores for Chinese patent publications.

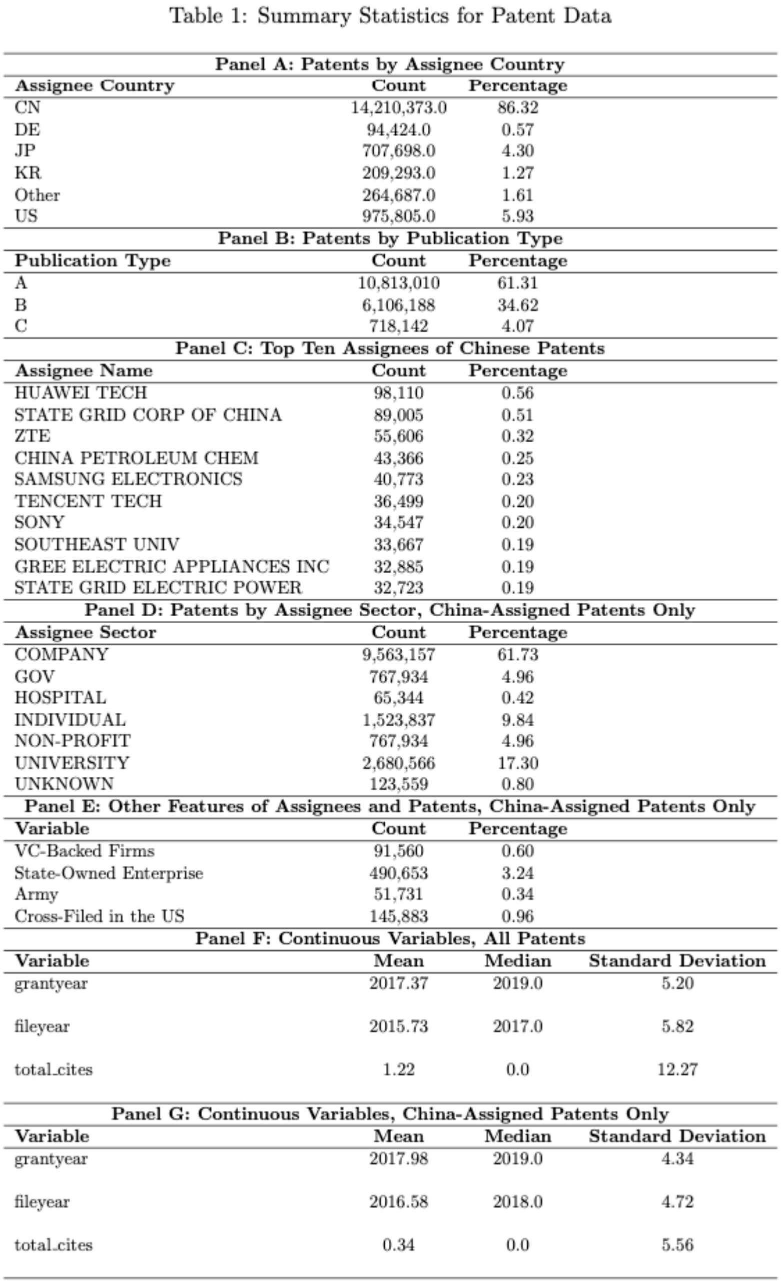


Table 1: Summary Statistics for Patent Data

Panel A: Patents by Assignee Country

Assignee Country	Count	Percentage
CN	14,210,373.0	86.32
DE	94,424.0	0.57
JP	707,698.0	4.30
KR	209,293.0	1.27
Other	264,687.0	1.61
US	975,805.0	5.93

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	98,110	0.56
STATE GRID CORP OF CHINA	89,005	0.51
ZTE	55,606	0.32
CHINA PETROLEUM CHEM	43,366	0.25
SAMSUNG ELECTRONICS	40,773	0.23
TENCENT TECH	36,499	0.20
SONY	34,547	0.20
SOUTHEAST UNIV	33,667	0.19
GREE ELECTRIC APPLIANCES INC	32,885	0.19
STATE GRID ELECTRIC POWER	32,723	0.19

Panel D: Patents by Assignee Sector, China-Assigned Patents Only

Assignee Sector	Count	Percentage
COMPANY	9,563,157	61.73
GOV	767,934	4.96
HOSPITAL	65,344	0.42
INDIVIDUAL	1,523,837	9.84
NON-PROFIT	767,934	4.96
UNIVERSITY	2,680,566	17.30
UNKNOWN	123,559	0.80

Panel E: Other Features of Assignees and Patents, China-Assigned Patents Only

Variable	Count	Percentage
VC-Backed Firms	91,560	0.60
State-Owned Enterprise	490,653	3.24
Army	51,731	0.34
Cross-Filed in the US	145,883	0.96

Panel F: Continuous Variables, All Patents

Variable	Mean	Median	Standard Deviation
grantyear	2017.37	2019.0	5.20
fileyear	2015.73	2017.0	5.82
total_cites	1.22	0.0	12.27

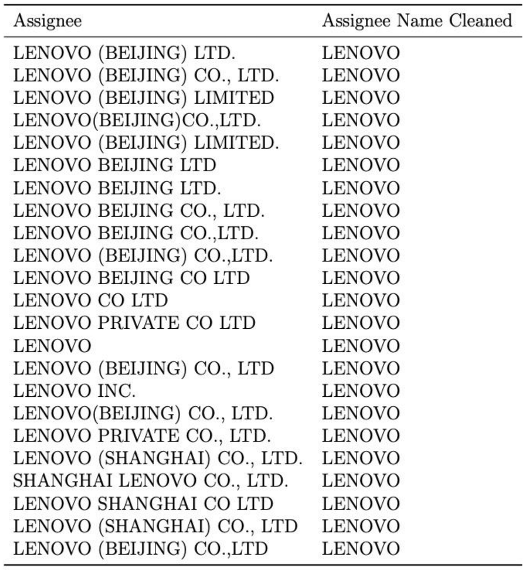
Panel G: Continuous Variables, China-Assigned Patents Only

Variable	Mean	Median	Standard Deviation
grantyear	2017.98	2019.0	4.34
fileyear	2016.58	2018.0	4.72
total_cites	0.34	0.0	5.56

Table 1: Summary statistics for Chinese patent publications.

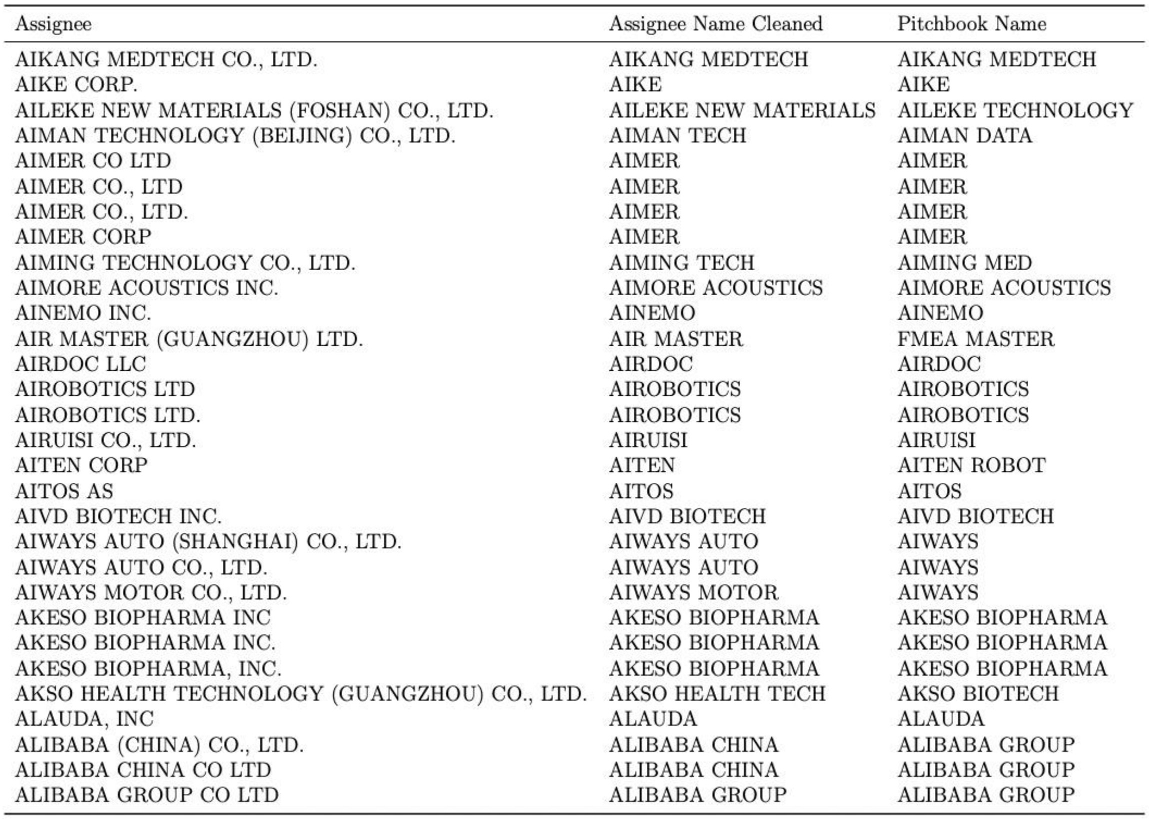
Appendix A: Database Documentation

Appendix B: 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

Appendix C: 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

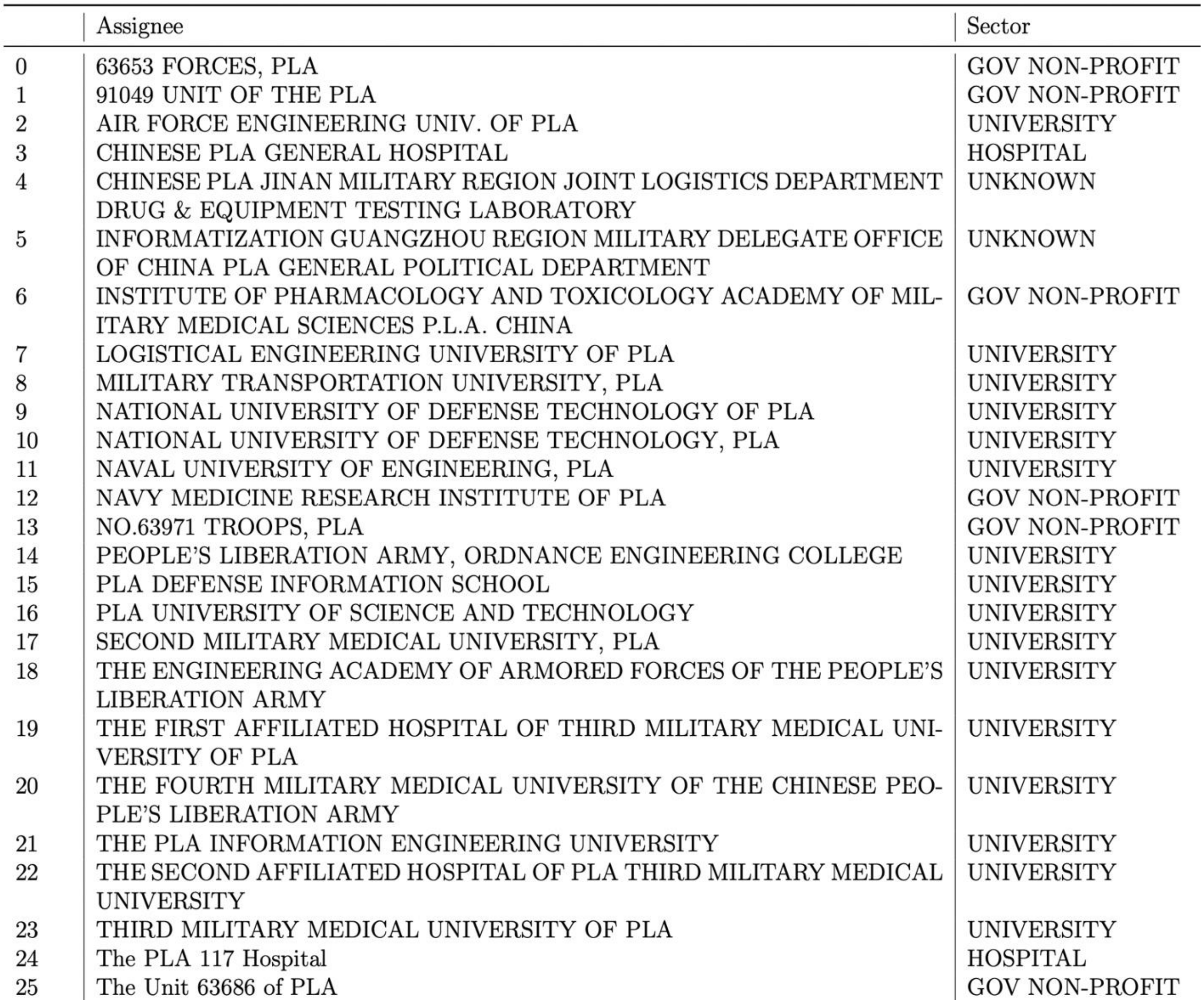
Appendix D: Patenting Entities Incorrectly Identified as Venture-Backed by PitchBook

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 PEOPLE & 39 ; S LIBERATION ARMY
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

Appendix E: 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
21	THE PLA INFORMATION ENGINEERING UNIVERSITY	UNIVERSITY
22	THE SECOND AFFILIATED HOSPITAL OF PLA THIRD MILITARY MEDICAL UNIVERSITY	UNIVERSITY
23	THIRD MILITARY MEDICAL UNIVERSITY OF PLA	UNIVERSITY
24	The PLA 117 Hospital	HOSPITAL
25	The Unit 63686 of PLA	GOV NON-PROFIT