University of Toronto
Department of Economics

Working Paper 770

The Anatomy of Chinese Innovation: Insights on Patent Quality and Ownership

By Philipp Böing, Loren Brandt, Ruochen Dai, Kevin Lim and Bettina Peters

March 8, 2024

The Anatomy of Chinese Innovation: Insights on Patent Quality and Ownership[*]

Philipp Boeing[†] Loren Brandt[‡] Ruochen Dai [§] Kevin Lim [¶] Bettina Peters [‖]

February 2024

Abstract

We study the evolution of patenting in China from 1985-2019. We use a Large Language Model to measure patent importance based on patent abstracts and classify patent ownership using a comprehensive business registry. We highlight four insights. First, average patent importance declined from 2000-2010 but has increased more recently. Second, private Chinese firms account for most of patenting growth whereas overseas patentees have played a diminishing role. Third, patentees have greatly reduced their dependence on foreign knowledge. Finally, Chinese and foreign patenting have become more similar in technological composition, but differences persist within technology classes as revealed by abstract similarities.

[*] We thank participants at the 7th Annual Bank of Canada-University of Toronto Conference on the Chinese Economy for invaluable feedback. Boeing and Peters acknowledge funding by the German Federal Ministry of Education and Research under grant number 01DO21006A. Brandt thanks funding support from the Noranda Chair at the University of Toronto. The responsibility for the content of this study lies with the authors.

[†] Goethe University Frankfurt, Frankfurt am Main, and ZEW – Leibniz Centre for European Economic Research, Mannheim (philipp.boeing@zew.de).

[‡] University of Toronto (loren.brandt@utoronto.ca).

[§] Central University of Finance and Economics (r.dai@cufe.edu.cn).

[¶] University of Toronto (kvn.lim@utoronto.ca).

[‖] ZEW – Leibniz Centre for European Economic Research, Mannheim (bettina.peters@zew.de).

1 Introduction

Patenting activity in China has accelerated substantially over the last two decades, far outpacing growth in places like the US. From 2000 to 2019, the number of invention patents granted by the China National Intellectual Property Administration (CNIPA) grew at an average annual rate of 23.5%, in contrast with an average growth rate of 4.3% at the US Patent and Trademark Office (USPTO).[1] In 2015, the number of patents granted at the CNIPA overtook the number of patents granted at the USPTO for the first time.

In this paper, we investigate the growth in patenting activity in China along four dimensions. First, to what extent has the quality of patenting in China changed over time? Second, which patentee types have been responsible for the rapid growth in Chinese patenting activity? Third, what sources of knowledge have new Chinese patents been building on and how much knowledge-sharing has there been between Chinese and foreign patentees? And fourth, how much overlap is there between Chinese and foreign patenting activity and how has this changed over time?

Answering these questions requires overcoming two important challenges. First, patents represent a vast and important source of codified knowledge, which is difficult to incorporate systematically into quantitative analyses of patenting growth and quality because it is almost exclusively in the form of text data. To make progress on this front, we leverage the capabilities of an industry-leading Large Language Model (LLM) to generate embeddings for the abstracts of every Chinese patent filed at the CNIPA. Embeddings are vectors that represent the meaning of a set of words in a vector space, where the mapping from text to vector is determined by an LLM that has been pre-trained on vast quantities of text data in multiple languages. These vector representations allow us to compute well-defined measures of patent similarity (cosine similarity between embeddings) and importance (similarity to the future minus similarity to the past). Our ability to make direct use of the semantic content of patent text allows us to shed new light on how patent quality has changed over time, as well as how the similarity between Chinese and foreign patenting has evolved.

The second challenge is that basic patent data provide very little information about patentees beyond basic names and addresses. To make progress in this regard, we leverage information on ownership of registered capital from a comprehensive business registry from China. Combining this with basic patent information allows us to differentiate between at least eight types of patenting entities: (i) privately-invested enterprises (PIEs); (ii) state-owned enterprises (SOEs); (iii) foreign-invested enterprises (FIEs); (iv) Chinese universities; (v) Chinese research institutes; (vi) individuals; (vii) other domestic patentees; and (viii) overseas patentees (those with an address outside of China). As we show, there are important differences in the growth and composition of patenting across these patentee types.

Our analysis leads to four novel insights. First, the importance of the average invention

[1] Invention patents in China are essentially equivalent to utility patents in the US.

patent in China declined from 2000 to 2010 but has been increasing in recent years. Second, private Chinese firms and universities account for most of the growth in Chinese patenting, whereas the role of overseas patentees has declined dramatically in terms of both levels and importance. Third, patentees in China have greatly reduced their dependence on foreign knowledge, a trend that began in the early 2000s. Finally, Chinese and foreign patenting have become more similar in terms of specialization across technology classes, but differences persist within technology classes as revealed by text similarities.

The remainder of this paper is organized as follows. Section 2 describes the key sources of data used in our analysis and explains how we classify patentee types using novel information from the Chinese business registry. Section 3 discusses the measurement of patent quality and how we utilize an LLM to incorporate patent text data into this measurement. Section 4 then presents our main findings in relation to the four key questions outlined above. Finally, section 5 concludes with some thoughts about the direction of future research on Chinese patenting and innovation more broadly.

2 Data and basic patent statistics

We study a comprehensive database covering all patents applied for at the CNIPA since 1985.[2] In what follows, we will restrict attention to invention patents, since these are typically considered to be the most innovative.[3] This includes Patent Cooperation Treaty (PCT) invention patents, which are filed almost exclusively by overseas applicants.[4] For brevity, we will henceforth refer to an invention patent application simply as a "patent".

Between 1985 and 2019, there is rapid growth in the number of patents at an average rate of 16.5% per year. Unlike for GDP and exports, this growth persists even after 2010. In terms of technology classes, the share of patents by main International Patent Classification (IPC) section in the average year, listed in descending order, is as follows: G (Physics) (20.3%), B (Performing Operations; Transporting) (17.8%), H (Electricity) (17.2%), A (Human Necessities) (16.2%), C (Chemistry; Metallurgy) (15.4%), F (Mechanical Engineering) (7.6%), E (Fixed Constructions) (3.9%), and D (Textiles; Paper) (1.7%).[5] Online Appendix A provides more detailed information about patent counts and shares by IPC section, as well as a breakdown of patents by product versus process classification, shares by location within China for domestic applicants,[6] and shares

[2] The CNIPA was founded in 1980 and was referred to as the State Intellectual Property Office (SIPO) for most of its existence.

[3] Around 40% of all patent applications in the average year are invention patent applications, with the remaining accounted for by utility patent applications (40%) and design patent applications (20%).

[4] China became a Patent Cooperation Treaty (PCT) contracting state on 1 January 1994.

[5] A patent can be associated with multiple IPC codes but each Chinese patent also reports its main IPC code, unlike patents at the USPTO.

[6] Domestic patent applications are highly concentrated in the coastal regions of China and this concentration

by country for foreign applicants.[7]

Patent application documents also provide information about the name of the patent applicant(s) and the address of the main applicant. Beyond this, however, little is known about who is responsible for patenting activity. For example, is the applicant of a patent an enterprise or a research institute? If it is an enterprise, who owns the firm? To make progress in this dimension, we define a taxonomy of what we refer to as the "patentee type" of a patent based on the intersection of three sources of information.

First, we identify from the address of the main patent applicant whether the *location* of the main patent applicant is domestic (reporting an address in China) or overseas (reporting an address outside of China). Second, for patents where the main applicant is domestic, we conduct a keyword search on the names of each applicant in a patent to differentiate between enterprises, universities, research institutes, individuals, and other domestic patentees (e.g., military). Appendix B provides a detailed description of this classification procedure. Third, we merge the data for domestic enterprise patents with a comprehensive registry of businesses in China. This allows us to observe ownership of registered capital for each entity in the business registry and hence to determine the *ownership type* of each enterprise applicant.[8]

Using these three sources of information, we thus define eight mutually exclusive *patentee types*: (i) privately-invested enterprises (PIEs); (ii) state-owned enterprises (SOEs); (iii) foreign-invested enterprises (FIEs); (iv) universities; (v) research institutes; (vi) individuals; (vii) other domestic patentees; and (viii) overseas patentees. This taxonomy will allow us to shed light on the actors responsible for patenting activity in China at a level of detail that has not previously been possible to examine.

3 Measuring patent quality

We now turn toward a central challenge: measuring the quality of a patent. We first describe how we use an LLM to generate embeddings for the text of each patent, how we utilize these embeddings to measure patent similarity and importance, and the advantages that this approach has relative to traditional approaches. We then validate our text-based measure of importance by documenting how it correlates with other commonly used measures of patent quality in the literature.

has been increasing over time. The role played by provinces in central China has also been increasing.

[7] Applications from the top five overseas locations – Japan, USA, Germany, South Korea, and Taiwan – represent around 80% of all overseas patent applications after 2000.

[8] Of all the invention patents that we identify as having an "enterprise" applicant name type, we are able to identify the ownership type of the patent from the business registry in around 88% of cases, while the remaining 12% cannot be matched to the business registry. Hence, the match between the patent data and the business registry is of high quality.

3.1 Text embeddings, patent similarity, and patent importance

Researchers studying patents have explored many different measures of patent quality, for example: the number of forward citations received by a patent (Hall et al. (2005), Kuhn et al. (2020)); the centrality of a patent in the citation network (Funk and Owen-Smith (2017), Park et al. (2023)); legal status changes (e.g., grants, unpaid renewal fees) (Hedge et al. (2023)); the timing of legal status changes (e.g., the time taken for a patent to be granted after application) (Chondrakis et al. (2021)); the number and length of patent claims (Marco et al. (2019)); and the existence of related patent filings at overseas patent offices (Harhoff et al. (2003)). Notice, however, that none of these measures make direct use of the semantic content of patent text, which is a key source of codified knowledge but that has been traditionally difficult to incorporate into quantitative analyses of patenting activity (earlier work in this direction includes Younge and Kuhn (2016) and Kelly et al. (2021).

To make progress, we utilize an industry-leading LLM to generate text embeddings for the abstract of every patent in the CNIPA database. Each embedding is a vector that represents the meaning of the abstract text, where the mapping from text to vector is determined by the LLM based on vast quantities of training data (e.g., all text that is publicly available on the internet). The model that we use in practice is the multilingual model from Cohere, a Canadian technology company that specializes in natural language processing (NLP) and LLMs.[9] The model represents each patent abstract with a 768-element vector, thereby reducing the extremely high-dimensional text data to a smaller but still high-dimensional space. In Online Appendix C, we provide a visualization of these embeddings using a dimensional reduction technique (t-Distributed Stochastic Neighbor Embedding) and show that patents sharing the same main IPC section tend to have abstract embeddings that are clustered together in similar areas of the vector space.

With these embeddings of the patent abstracts, we can compute well-defined measures of similarity between any pair of patents. We focus on a widely-used measure of similarity in natural language processing: cosine similarity. Consider two patents, i and j, and let $e_i \equiv \left[e_i^1 \cdots e_i^N\right]'$ and $e_j \equiv \left[e_j^1 \cdots e_j^N\right]'$ denote the embeddings of the abstracts for these two patents (where $N = 768$ is the dimension of each embedding). The cosine similarity of e_i and e_j is defined as follows:

$$\rho_{ij} = \frac{e_i \cdot e_j}{||e_i||||e_j||} = \frac{\sum_{k=1}^{N} e_i^k e_j^k}{\left[\sum_{k=1}^{N} \left(e_i^k\right)^2\right]^{\frac{1}{2}} \left[\sum_{k=1}^{N} \left(e_j^k\right)^2\right]^{\frac{1}{2}}} \tag{3.1}$$

The cosine similarity of any two vectors lies in the interval $[-1, 1]$ and is closer to 1 when the

[9] See https://txt.cohere.com/multilingual/ for a user-friendly introduction to the Cohere multilingual model.

two vectors are pointing in more similar directions in the vector space.[10]

Using this measure of cosine similarity, we can also construct text-based measures of patent quality. Let J^+_{st} and J^-_{st} denote the set of patents with main IPC section s that are applied for in the three years after and before year t respectively.[11] We define the *importance* of patent i with main IPC section s applied for in year t, I_{ist}, as follows:

$$\log I_{ist} = \frac{1}{|J^+_{st}|} \sum_{j \in J^+_{st}} \rho_{ij} - \frac{1}{|J^-_{st}|} \sum_{j \in J^-_{st}} \rho_{ij} \tag{3.2}$$

In words, the log importance of patent i is the difference between the average cosine similarity of its abstract embedding to all patents in the future three years and to all patents in the previous three years, where only patents that have the same main IPC section as patent i are considered. Intuitively, high textual similarity to future patents is indicative of a patent that has high impact on future patenting activity, whereas low textual similarity to past patents is indicative of a novel patent. An important patent is thus one that tends to have high impact and novelty.

This approach to measuring patent quality has several advantages over traditional approaches. First, it makes direct use of the semantic meaning of the patent text – information that is used by patent examiners to evaluate patent applications, for example – whereas almost all other existing measures of patent quality ignore the information contained in these text data. Second, the text of a patent is directly observable to researchers from the very first patent application document that is filed at the patent office. In contrast, forward citations received by a patent are typically not observed until several years after the patent has been applied for, simply because it takes time for other patents that are filed in the future to cite patents in the past.[12] Similarly, legal status changes (e.g., the granting of a patent) or related filings at overseas patent offices are ex post outcomes that take time to materialize after a patent is first applied for. Finally, our text-based measure of patent importance is less likely to suffer from well-known biases in other measures of patent quality. For example, patent applicants (Lampe, 2012) and examiners (Hegde and Sampat, 2009) may vary in their willingness to cite prior art for reasons that have nothing to do with patent quality.[13] Firms that share common ownership linkages, for

[10] When $N = 2$, $csim_{ij}$ is the cosine of the angle between the vectors e_i and e_j, hence the name.

[11] We use three-year windows in our baseline analysis to maximize the number of years that we are able to include in our analysis, although all our main findings are robust to using longer windows (we have explored windows of up to 7 years).

[12] In the Chinese context, backward citations are only observed for patents that are granted. Hence, to observe a forward citation for a patent, one must wait for a patent in the future to not only be applied for but also to be granted in order to observe whether this future patent makes a relevant backward citation.

[13] Several studies substantiate this concern by estimating the relationship between citations and economic outcomes. For example, Yin and Sun (2023) show that forward citations are not correlated with initial patent auction prices in China, while Wu et al. (2022) only find a significant relationship between patent citations and firms' total factor productivity for patents with more than ten claims. Similarly, restricting attention to citations

instance, may be more likely to cite each others' patents simply because of familiarity. Similarly, the granting of a patent is dependent on factors besides patent quality, such as the strictness of the patent examiner (Higham et al., 2021), heterogeneity across national examination practices (Bacchiocchi and Montobbio, 2010), and biases against foreign inventors (Webster et al., 2016). In China, national patenting targets (Sun et al., 2021), subsidies (Branstetter et al., 2023), and tax cuts (Wei et al., 2023) have contributed to distorted patenting activities, while excessive workload and low salaries for examiners potentially degrade the quality of examinations (Branstetter et al., 2023). In contrast, because our measure of importance is not based on ex post outcomes, it is much less likely to suffer from such biases.

3.2 Validating our measure of patent importance

To validate our text-based measure of patent importance, we regress the grant status of a patent (0 for ungranted, 1 for granted, multiplied by 100) on various measures of quality, including our importance measure (row i); whether a patent has any forward citations in the first seven years after its application year (row ii); the number of independent claims in the patent application (row iii); the average length of each claim in the patent application (row iv); and whether a patent has a priority filing at the USPTO (row v). The importance measure, the number of independent claims, and average claim length are standardized based on their respective distributions across all CNIPA invention patents in our data. For simplicity, we run these regressions on invention patents applied for at the CNIPA in one year only, 2010. All regressions include fixed effects for the main IPC section of each patent (1-digit) and the patentee type associated with the patent as described above.

Table 1 shows our results. As expected, our importance measure is positively associated with the grant status of a patent: a one standard deviation increase in the importance score translates into an increase in the grant rate of the patent by 4.8 percentage points. Also as expected, the other measures of patent quality are positively associated with the grant status of a patent. In addition, when we include all of the regressors in the regression simultaneously (column 6), the coefficients on each quality measure – including on the importance score – remain positive and significant. This suggests that these different measures are capturing different dimensions of patent quality. Finally, when we omit the importance score regressor but instead include each element of the 768-vector representation of a patent's abstract as additional controls (column 7), the share of variance explained by our regressors increases by around 30% (the R^2 statistic increases from 0.17 in column 6 to 0.23 in column 7), indicating that variation in the text of a patent's abstract is useful for explaining variation in grant statuses.[14]

generated by international search reports of PCT applications, Boeing and Mueller (2019) find that only foreign citations – and not domestic or self-citations – have a significant and positive relationship with R&D in China.

[14] Even though including the full abstract embedding in the regression increases the number of regressors significantly, note that we still have many more observations than regressors. Furthermore, note that the adjusted

Table 1: Relationship between patent grant status and other quality measures

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
i. importance	4.88					4.40	
	(0.08)					(0.08)	
ii. has forward citation		15.94				14.91	11.22
		(0.12)				(0.12)	(0.12)
iii. no. of independent claims			2.70			2.35	1.29
			(0.06)			(0.06)	(0.07)
iv. average claim length				2.13		1.88	1.31
				(0.04)		(0.04)	(0.04)
v. has USPTO filing					30.46	27.74	20.49
					(0.43)	(0.43)	(0.42)
embedding control	no	no	no	no	no	no	yes
observations (m)	0.98	0.98	0.96	0.96	0.98	0.96	0.96
R^2	0.15	0.16	0.14	0.14	0.15	0.17	0.23
adjusted R^2	0.15	0.16	0.14	0.14	0.15	0.17	0.23

Notes: Each column shows the results of a regression of the grant status of a patent (0 for ungranted, 1 for granted, multiplied by 100) on various measures of patent quality. Standard errors are shown in parentheses. The regressors in rows i, iii, and iv are standardized based on their respective distributions over all patents in our sample. Each regression includes fixed effects for the main IPC section of each patent and the patentee type associated with the patent. The "embedding control" refers to whether the full abstract embedding for a patent's abstract is included in the regression. The sample for all regressions is all CNIPA invention patents applied for in 2010.

We provide further validation of our importance measure in Online Appendix D by regressing various traditional measures of patent quality on a patent's importance percentile within each application year. We provide graphical evidence that patents with higher importance percentiles tend to have more forward citations, higher grant propensity, and shorter lags between application and grant dates.

4 Main findings

To describe our main findings, we will now index patentee types by p and will continue indexing IPC sections by s and application years by t. For brevity, we will refer, for example, to the set of patents belonging to patentees of type p with main IPC section s applied for in year t as pst-patents.

R^2 statistic (which adjusts the R^2 for the number of regressors) also increases by around 30%.

4.1 The quality of Chinese patenting

Importance. Panel (a) of Figure 1 shows how the distribution of log importance varies over time (see panel (a) of Figure A.IV in Online Appendix E for separate plots by each IPC section). For context, the plots also show how the importance of CNIPA patents for patenting at the USPTO has changed over time.[15] To aid with the interpretation of magnitudes, the plot reports scores that are standardized relative to the distribution of importance (for CNIPA patents relative to CNIPA patents) over all patents in our data.

We observe that average patent importance increased from 1997 to 2002, declined from 2002 to 2010, followed by a marked reversal of the downward trend in patent importance after 2010. Furthermore, as should be expected, CNIPA patents are more important for patenting activity at the CNIPA than at the USPTO, although there is a slight increase in the importance of CNIPA patents for USPTO patents from 2004 onward.

What explains these trends in patent importance? Consider all t-patents and let $\bar{I}_t$ denote the average importance of these patents. We show in Appendix G that for small changes, the log change in $\bar{I}_t$ can be decomposed into components arising from changes in the importance (intensive margin) and number of patents (extensive margin) by patentees of different types:

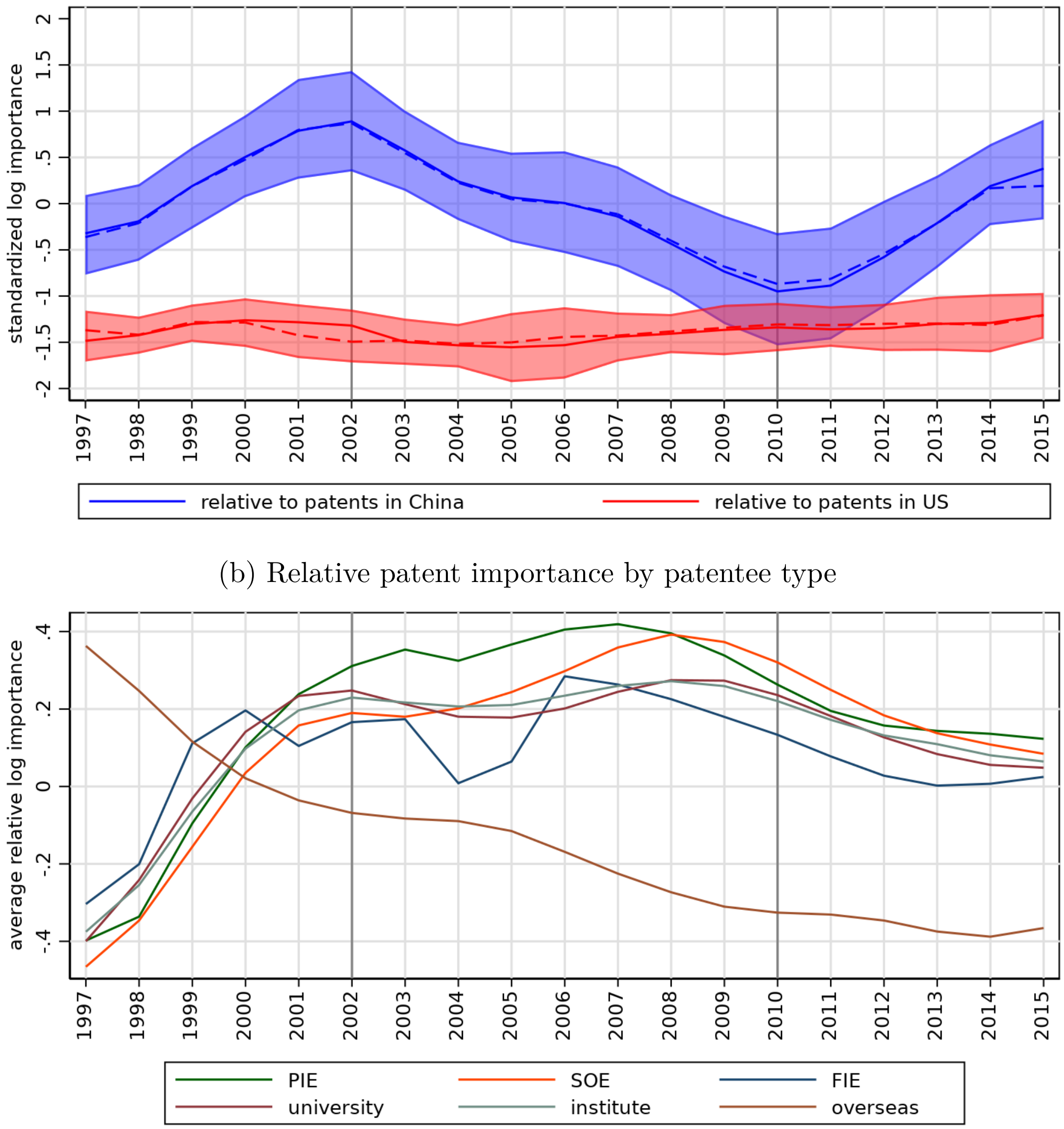
$$\hat{\bar{I}}_t = \sum_p [\underbrace{s_{pt} r_{pt} \hat{\bar{I}}_{pt}}_{\text{(i) intensive margin}} + \underbrace{s_{pt} \left(r_{pt} - 1 \right) \hat{N}_{pt}}_{\text{(ii) extensive margin}}] \tag{4.1}$$

where $\hat{\bar{I}}_{pt}$ denotes the log change in the average importance of pt-patents, $\hat{N}_{pt}$ denotes the log change in the number of pt-patents, s_{pt} is the share of t-patents accounted for by pt-patents, and r_{pt} is the average importance of pt-patents relative to all t-patents. Intuitively, average patent importance can increase if either: (i) the average importance of patents by a given patentee type p increases, especially if these patents account for a large share of patents (s_{pt} is high) and have high relative importance to begin with (r_{pt} is high); or (ii) there is growth in patents by patentee types that have higher-than-average importance ($r_{pt} > 1$).

Panel (a) of Table 2 shows the results of this decomposition for three separate time periods: 1997-2002, 2002-2010, and 2010-2015. As discussed above, patent importance increased during the first and third of these periods (at average annual rates of 28.3% and 27.7% respectively) but declined during the second period (at an average annual rate of -24.7%). Several key findings emerge. First, even though the decomposition in equation (4.1) technically applies to marginal changes, it provides a close approximation to the actual growth rates of average patent

[15] For each CNIPA patent, we again compute importance as in equation (3.2), except that the comparison sets J^+_{st} and J^-_{st} are for patents at the USPTO instead of at the CNIPA. Note that USPTO patents may be associated with multiple IPC classes but do not report their main IPC class. Hence, in computing the importance scores, we split each multi-class USPTO patent into fractions according to the share of IPC classes reported by the patent in each IPC section.

Figure 1: The evolution of patent importance


(a) Average patent importance

(b) Relative patent importance by patentee type

Notes: Panel (a) shows the mean (blue solid line), median (blue dotted line), and interquartile range (blue shaded area) of the log importance distribution for patents in each year. The corresponding lines and areas in red show the same outcomes but for the importance of CNIPA patents for USPTO patents. Panel (b) shows the rolling five-year average importance of patents applied for by different patentee types relative to the average importance within each year. All importance scores are standardized relative to the distribution of importance of CNIPA patents for CNIPA patents across all years.

importance. Second, almost all of the changes in average importance are driven by changes in the intensive margin measures (i.e., changes in patent importance for patentees of different types). Third, most of the growth in patent importance from 1997-2002 was driven by growing importance of overseas patents. Fourth, the decline in patent importance from 2002-2010 was also largely driven by the declining importance of overseas patents. Finally, the growth of patent importance from 2010-2015 was largely driven by growth in the importance of PIE patents. In sum, these findings highlight that overseas patents were key for patent importance growth initially but contributed much less in later years, with PIE patents becoming the dominant source of growth.

To provide more context for how the quality of Chinese patenting has changed over time, Online Appendix F documents our findings about how more traditional measures of patent quality have evolved, including the average number of forward citations per patent, citation centrality, legal status changes (e.g., whether a patent was examined or lapsed due to unpaid fees), the timing of legal status change events, the number and length of patent claims, and the propensity of priority filings at the USPTO.

4.2 Differences in patenting activity and quality by patentee type

Patent shares by patentee type. Panel (a) of Figure 2 shows the shares of invention patent applications accounted for by different patentee types. The most important observation here is the steady decline in the role played by overseas patentees and the growth in the share of patents accounted for by PIEs. In 1995, for example, overseas patentees account for sixty percent of all patents, with even larger shares in some sections (e.g., 90% in section H (Electricity)). By 2019, however, overseas patentees account for only 10% or less of patents in all IPC sections. In fact, the role played by overseas patentees declines not only in terms of patent shares but also in terms of patent growth rates. For example, the average annual growth rate in overseas patents falls from 16.1% between 1997-2002 to 9.9% between 2002-2010, then falls even further to 3.4% between 2010-2018. In contrast, PIEs become the most important patentee type by the end of our sample. In 2000, PIEs account for only a marginal share of patenting activity, but by 2019, they make up between 40-60% of patents across IPC sections. Chinese universities also account for a growing share of patents in IPC sections C (Chemistry; Metallurgy) and G (Physics). Unlike the declining growth rates observed for overseas patents, domestic patents grow at a fairly constant rate of around 26.9% between 1997-2018.

Patent quality by patentee type. Which patentee types produce the most important patents? To investigate, we construct the average log importance of *pt*-patents relative to the

Table 2: Decomposition of changes in patent importance

(a) Decomposition of changes in average patent importance by patentee type

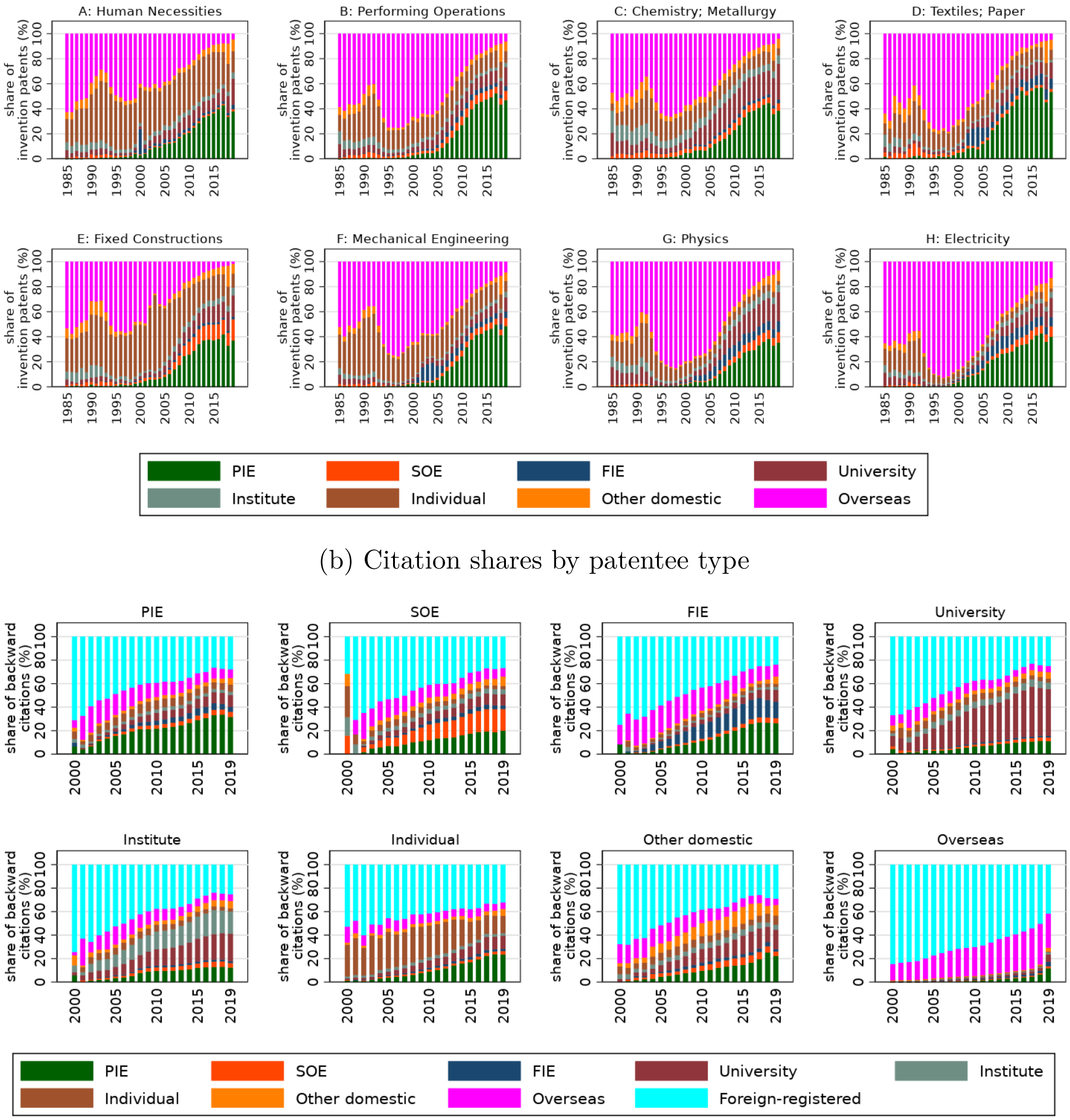
	1997-2002		2002-2010		2010-2015	
avg. growth	28.3		-24.7		27.7	
approx. avg. growth	30.9		-24.0		27.8	
decomposition	*int. margin*	*ext. margin*	*int. margin*	*ext. margin*	*int. margin*	*ext. margin*
	29.4	1.5	-26.4	2.4	26.1	1.6
PIEs	2.4	1.1	-5.1	1.9	10.6	1.4
SOEs	0.4	-0.0	-0.9	0.3	1.2	0.1
FIEs	1.2	0.5	-1.8	0.4	1.0	-0.1
universities	0.7	0.1	-2.3	0.2	3.0	-0.3
institutes	0.8	0.1	-1.0	0.1	1.0	-0.1
individuals	8.0	0.4	-3.7	0.0	4.7	0.8
other domestic	1.0	0.1	-0.8	0.1	1.4	0.0
overseas	15.0	-0.7	-11.0	-0.6	3.2	-0.3

(b) Decomposition of changes in relative patent importance by patentee type

	1997-2002				2002-2010				2010-2015			
	PIE	SOE	FIE	overseas	PIE	SOE	FIE	overseas	PIE	SOE	FIE	overseas
PIE	2.3	1.4	1.5	-0.4	1.4	1.5	3.7	-1.9	-0.9	-2.2	-2.5	-2.4
SOE	1.0	0.8	1.1	-0.3	0.5	0.6	0.9	-0.4	-0.7	-0.9	0.0	0.4
FIE	1.6	0.0	2.1	0.0	-0.8	0.1	-0.8	-0.2	-0.6	-0.5	0.4	0.5
univ.	2.3	3.2	2.4	-0.7	0.0	-0.7	0.6	-0.5	-0.8	-0.8	-0.5	0.0
inst.	1.5	1.4	1.6	-0.5	0.0	0.1	0.0	-0.1	-0.2	-0.2	-0.1	0.2
indiv.	3.4	1.8	3.8	-0.7	0.0	0.4	-0.1	0.1	0.5	0.3	0.1	-0.5
other	2.0	1.9	2.0	-0.7	0.3	0.2	0.4	-0.3	0.0	-0.2	-0.3	-0.4
overseas	0.2	2.5	-5.1	-5.3	-2.0	-0.7	-5.1	0.1	0.0	-0.3	-0.9	1.5
total	14.2	13.1	9.4	-8.6	-0.6	1.6	-0.4	-3.2	-2.8	-4.7	-2.2	-0.8

Notes: Panel (a) shows the results of the decomposition of changes in patent importance specified in equation (4.1). Panel (b) shows the results of the decomposition specified in equation (4.4). Each column in panel (b) corresponds to a patentee type p. The rows labeled "total" report the average annual change in the importance of p-patents relative to mean within IPC section and year. The other rows in panel (b) show the decomposition of this total change into changes in the bilateral importance of p-patents for p'-patents within the same IPC section.

Figure 2: Patents and citation shares by patentee type



(a) Patent shares by patentee type

(b) Citation shares by patentee type

Notes: Each plot in panel (a) shows the share of invention patents accounted for by each patentee type. Each plot in panel (b) shows the share of backward citations made by patents associated with a given patentee type that cite patents associated with different patentee types (including patents that are filed at patent offices other than the CNIPA). We exclude citations made to non-invention (design and utility) CNIPA patents.

corresponding average in year t as follows:

$$\log \tilde{I}_{pt} = \frac{1}{|J_{pt}|} \sum_{s} \sum_{i \in J_{pst}} \log I_{ist} - \frac{1}{|J_t|} \sum_{s} \sum_{i \in J_{st}} \log I_{ist} \tag{4.2}$$

where J_x denotes the set of x-patents for $x \in \{pst, pt, st, t\}$. Panel (b) of Figure 1 shows these relative importance scores for PIEs, SOEs, FIEs, universities, research institutes, and overseas patentees (see panel (b) of Figure A.IV in Online Appendix E for separate plots for each IPC section). In the early years of our sample, patents applied for by overseas patentees are the most important. However, we observe a clear decline over time in the relative importance of these patents, together with a steady increase in the relative importance of domestic patents from 1997 to 2007. As early as 2000, the average overseas patent is less important than the average patent applied for by domestic enterprises, universities, and institutes. Furthermore, by 2015, the gap in relative log importance between domestic and overseas patents is around half the standard deviation of the log importance distribution. For most of the 2000s, we also see that PIE patents have the highest relative importance, although SOE patents become more important on average between 2009 and 2013.

4.3 Dependencies between Chinese and foreign patenting

Text-based dependencies. What explains the declining importance of overseas patents relative to domestic patents in China? We now examine bilateral measures of importance between patents applied for by different patentee types. First, note that the log importance of a patent $i \in J_{st}$ defined in equation (3.2) can always be expressed as:

$$\log I_{ist} = \sum_{p} \underbrace{\left[\frac{1}{|J_{st}^{+}|} \sum_{j \in J_{pst}^{+}} \rho_{ij} - \frac{1}{|J_{st}^{-}|} \sum_{j \in J_{pst}^{-}} \rho_{ij} \right]}_{\log I_{ipst}} \tag{4.3}$$

where J_{pst}^{+} and J_{pst}^{-} denote the sets of ps-patents applied for in the three years after and before year t, respectively. The term in parentheses, which we denote by $\log I_{ipst}$, can be interpreted as the importance of patent i for ps-patents specifically.[16] It is then straightforward to show that the relative importance score defined in equation (4.2) can simply be decomposed as:

$$\log \tilde{I}_{pt} = \sum_{p'} \log \tilde{I}_{pp't} \tag{4.4}$$

[16] Note that we can also write this as $\log I_{ipst} = \frac{\left|J_{pst}^{+}\right|}{\left|J_{st}^{+}\right|} \left(\frac{1}{\left|J_{pst}^{+}\right|} \sum_{j \in J_{pst}^{+}} \rho_{ij} \right) - \frac{\left|J_{pst}^{-}\right|}{\left|J_{st}^{-}\right|} \left(\frac{1}{\left|J_{pst}^{-}\right|} \sum_{j \in J_{pst}^{-}} \rho_{ij} \right)$, which is simply the difference in the average cosine similarity of patent i to patents in J_{pst}^{+} versus those in J_{pst}^{-}, weighted by the share of s-patents in each period accounted for by ps-patents.

where $\log \tilde{I}_{pp't} \equiv \frac{1}{|J_{pt}|}\sum_s \sum_{i \in J_{pst}} \log I_{ip'st} - \frac{1}{|J_t|}\sum_s \sum_{i \in J_{st}} \log I_{ip'st}$ is the demeaned average log importance of pt-patents for p'-patents. This decomposition thus allows us to examine how the importance of patents belonging to a given patentee type p depends on their importance for patents belonging to all other patentee types p'.

Panel (b) of Table 2 reports the results of this decomposition in changes for the same time periods as before: 1997-2002, 2002-2010, and 2010-2015. For brevity, we report these results only for patents by domestic enterprises (PIEs, SOEs, FIEs) and overseas applicants. For each patentee type p (shown in the columns), the table reports the annual changes of both $\log \tilde{I}_{pt}$ and $\log \tilde{I}_{pp't}$. For example, the first column shows that between 1997-2002, the relative log importance of PIE patents grew at an annual rate of 14.2%. Of this, 2.3% is attributable to growth in the importance of PIE patents for other PIE patents.

In the first period, 1997-2002, we observe that the growing relative importance of domestic enterprise patents is driven mainly by the growing importance of these patents for patenting by PIEs, universities, and individuals. On the other hand, the relative importance of overseas patents declines largely because these patents become less important for themselves. In the second period, 2002-2010, we observe much smaller changes in relative importance over time. Domestic enterprise patents continue to grow in importance for PIE patents, whereas the importance of overseas patents for PIE patents falls further. In the third period, 2011-2015, the relative importance of domestic enterprise and overseas patents declines slightly, which is mostly driven by the falling importance of these patents for PIE patenting.

Citation-based dependencies. Panel (b) of Figure 2 shows the share of backward citations made by patents associated with a given patentee type that cite patents associated with different patentee types. We include here the share of citations that cite foreign-registered patents (i.e., at patent offices other than the CNIPA) but exclude citations of non-invention (design and utility) CNIPA patents. A key observation is that the share of backward citations to foreign-registered patents declines steadily for all domestic patentee types from 2000 onward. In contrast, all domestic patentee types increase their dependence on PIE patents in terms of backward citations. In this sense, domestic patentees in China have become much less dependent on foreign knowledge for their patenting activity. Note that the overseas patentees predominantly cite other overseas patents as well as foreign-registered patents and that this changes very little from 2000 to 2019. Furthermore, there is a noticeable propensity for patentees to cite patents within their own type – for example, the share of FIE citations that cite FIE patents is greater than share of PIE citations that cite FIE patents.

4.4 Overlap between Chinese and foreign patenting

Overlap across technology classes. To explore the extent to which different patentee types are patenting in the same technology classes, we first compute the share of patents for each patentee type that is accounted for by each IPC 4-digit category. For each patentee type, this gives us a vector of shares where the dimension of the vector is the number of 4-digit IPC categories. We then compute the cosine similarity between these vectors for each pair of patentee types. To look into how these similarities change over time, we compute these measures separately for patents in each application year. Panel (a) of Figure 3 shows how the similarities between PIE patents and FIE, overseas, and granted USPTO patents evolves over time. We observe that in 2000, the composition of PIE patenting activity in terms of IPC 4-digit categories is most similar to the composition of patenting activity by overseas patentees. Over time, however, PIE patent composition become more similar to SOE, FIE, and even USPTO patent composition (from 2008 onwards). By 2019, these similarity scores are higher and much closer to each other than they were in the past, indicating that by the end of the sample, different patentee types are essentially patenting in more similar technology classes.

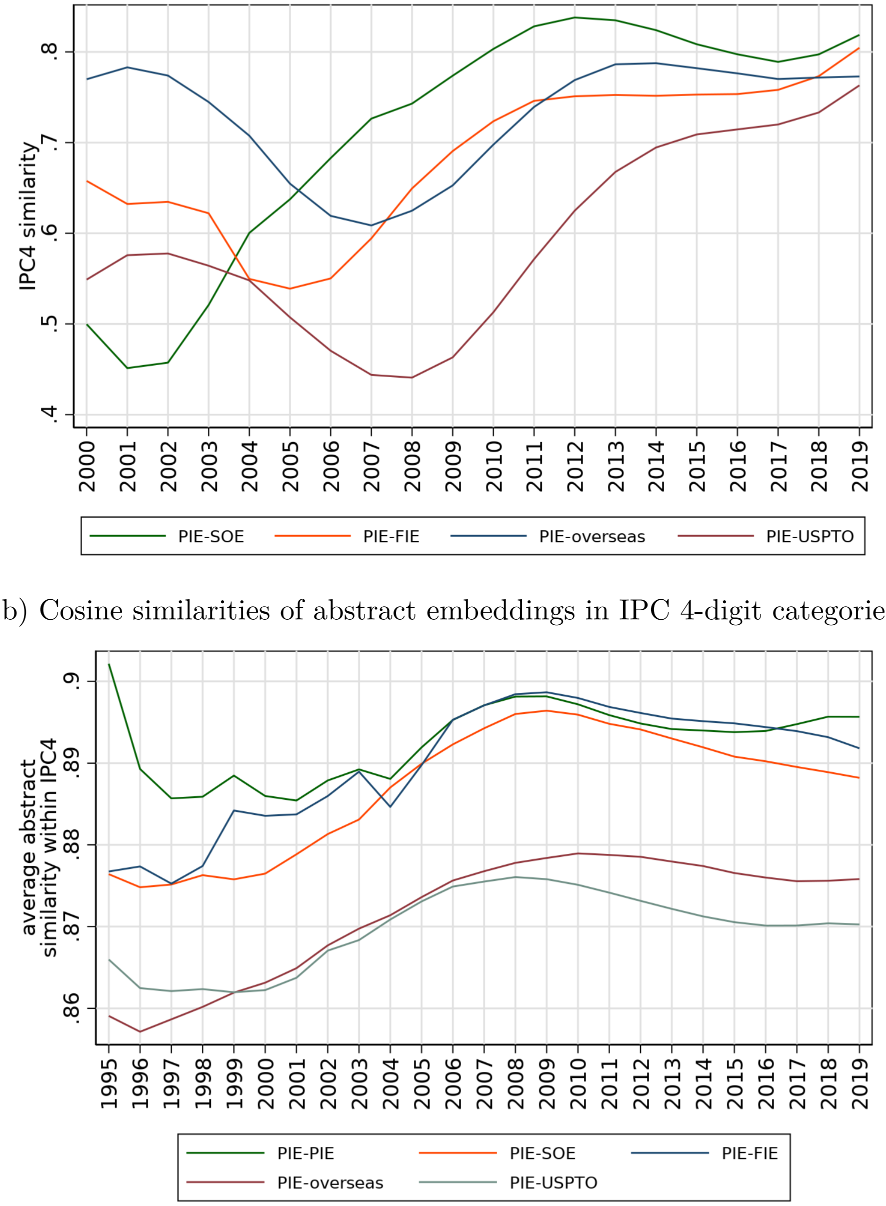
Overlap in text similarity within technology classes. Even though the composition of patents becomes more similar across patentee types in terms of technology classes, there may be differences in patenting activity within technology classes. To investigate, we compute the average cosine similarity of the abstract embeddings between patents associated with each pair of patentee types within an IPC 4-digit category. Panel (b) of Figure 3 shows the average of these similarity scores across IPC 4-digit categories for PIE patents relative to PIE, SOE, FIE, overseas and USPTO patents. We observe that PIE patent abstracts within an IPC 4-digit category are most similar to the abstracts of other enterprise patents in China. In contrast, they are less similar to the abstracts of overseas and USPTO patents. These similarity scores increase from 1995 to 2008, indicating that PIE patents are becoming more similar to the patents of other patentee types. However, the gap in similarity between enterprise patents versus overseas and USPTO patents persists, highlighting that even though PIEs are patenting in more similar technology classes as overseas and USPTO patentees, there are important differences in this patenting activity within technology classes.[17]

5 Conclusion

We conclude with four key takeaways. First, using our novel measure of text-based importance, we find that patent importance declined from 2002-2010, driven largely by the falling

[17] The standard deviation of the cosine similarity between a pair of patents in a given year is on the order of 10^{-2}. Hence, the gap in similarity between enterprise patents versus overseas and USPTO patents is around one to two standard deviations.

Figure 3: Similarities between patenting activity by different patentee types



(a) Cosine similarities of IPC 4-digit patent shares

(b) Cosine similarities of abstract embeddings in IPC 4-digit categories

Notes: Panel (a) shows shows the cosine similarity between PIE and SOE, FIE, overseas, and USPTO patent composition in terms of their mix across IPC 4-digit categories. Panel (b) shows the average cosine similarity between PIE and PIE, SOE, FIE, overseas, and USPTO patents within an IPC 4-digit category based on their abstract embeddings. For visual clarity, we plot rolling five-year averages that include the current year and past four years.

importance of overseas patents, but has been increasing since 2010 due mostly to the rising importance of PIE patents. Second, PIEs account for most of the growth in Chinese patenting activity, whereas the role of overseas patentees has declined dramatically in terms of both shares and relative importance. Third, overseas patents have become much less important for patenting by domestic enterprises in China. Patentees in China have also greatly reduced their dependence on foreign knowledge as measured by citations, which is not a recent phenomenon. Finally, Chinese and foreign patenting have become more similar in terms of specialization across technology classes, but differences persist within technology classes as revealed by text similarities.

Several directions for future research are promising. First, an important question that arises from our analysis is why the role of overseas patents has sharply diminished in China. Some of this may be expected given improvements in China's own innovative capabilities, but the extent of the decline appears out of line with the growing importance of the Chinese market in a global context more broadly. Second, our approach to measuring patent importance in China can also be applied to other contexts, such as patenting at the USPTO. This will not only be helpful for examining the forces underlying the decline in the growth of overseas patents in China, but also to provide context for the trends in Chinese patent importance that we have documented here. Third, we have conducted the analysis in this paper largely at the patent level, but aggregating patent information to the applicant level will also be important for identifying the firms, universities, and institutes that are the most key for Chinese patenting activity. Fourth, innovation is a key focus of Chinese industrial policy and hence it is essential to understand how patenting and innovative activity more broadly are being directed by policies such as those targeting strategic and emerging industries. Finally, our method of employing LLMs to study patent abstracts can also be extended to incorporate information contained in patent claims. This source of text data is much larger in scale and may be helpful in constructing more refined measures of patent importance and similarity.

References

Bacchiocchi, E. and F. Montobbio (2010). International knowledge diffusion and home-bias effect: Do uspto and epo patent citations tell the same story? Scandinavian Journal of Economics 112(3), 441–470.

Boeing, P. and E. Mueller (2019). Measuring china's patent quality: Development and validation of isr indices. China Economic Review (57).

Branstetter, L. G., D. Hanley, and H. Zhang (2023). Unleashing the dragon: The case for patent reform in china. Working paper.

Chondrakis, G., E. Melero, and M. Sako (2021). The effect of coordination requirements on sourcing decisions: Evidence from patent prosecution services. Strategic Management Journal 43, 1141–1169.

Funk, R. J. and J. Owen-Smith (2017). A dynamic network measure of technological change. Management Science 63(3), 791–817.

Hall, B. H., A. Jaffe, and M. Trajtenberg (2005). Market value and patent citations. RAND Journal of Economics 36, 16–38.

Harhoff, D., F. M. Scherer, and K. Vopel (2003). Citations, family size, opposition and the value of patent rights. Research Policy 32(8), 1343–1363.

Hedge, D., K. Herkenhoff, and C. Zhu (2023). Patent publication and innovation. Journal of Political Economy 131(7), 1845–1903.

Hegde, D. and B. Sampat (2009). Examiner citations, applicant citations, and the private value of patents. Economics Letters 105, 287–289.

Higham, K., G. de Rassenfosse, and A. B. Jaffe (2021). Patent quality: Towards a systematic framework for analysis and measurement. Research Policy 50(4).

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.

Kuhn, J., K. Younge, and A. Marco (2020). Patent citations reexamined. RAND Journal of Economics 51(1), 109–132.

Lampe, R. (2012). Strategic citation. Review of Economics and Statistics 94(1), 320–333.

Marco, A. C., J. D. Sarnoff, and C. A. W. deGrazia (2019). Patent claims and patent scope. Research Policy 48(9).

Park, M., E. Leahey, and R. J. Funk (2023). Papers and patents are becoming less disruptive over time. Nature 613(7942), 138–144.

Sun, Z., Z. Lei, B. D. Wright, M. Cohen, and T. Liu (2021). Government targets, end-of-year patenting rush and innovative performance in china. Nature Biotechnology (39), 1068–1075.

Webster, E., P. H. Jensen, and A. Palangkaraya (2016). Patent examination outcomes and the national treament principle. Mimeo.

Wei, S.-J., J. Xu, G. Yin, and X. Zhang (2023). Mild government failure. Working paper.

Wu, H., J. Lin, and H.-M. Wu (2022). Investigating the real effect of china's patent surge: New evidence from firm-level patent quality data. Journal of Economic Behavior and Organization (204), 422–442.

Yin, Z. and Z. Sun (2023). Predicting the value of Chinese patents using patent characteristics: evidence based on a Chinese patent auction. Industrial and Corporate Change, 1286–1304.

Younge, K. and J. Kuhn (2016). Patent-to-patent similarity: A vector space model. Mimeo.

A Basic patent statistics

Technology classes. Panel (i) of Table A.I categorizes patents by their main International Patent Classification (IPC) section (1-digit). The composition of invention patents by technology class changes little over time. Sections A (Human Necessities), B (Performing Operations; Transporting), C (Chemistry; Metallurgy), G (Physics) and H (Electricity) account for the largest shares of invention patents. Patenting in sections A (Human Necessities) and C (Chemistry; Metallurgy) become slightly less prevalent after 2000, whereas the opposite is true for patenting in section G (Physics). Patenting in section H (Electricity) was also relatively more prevalent from 2000-2010 compared with after 2010.

Product versus process patents. Panel (ii) of Table A.I categorizes patents into process versus product patents based on the text of a patent's claims. First, a set of keywords referring to prior claims is used to identify dependent claims in a patent and separate them from independent claims. Second, each independent claim is searched for a set of 60 positive keywords that are associated with process innovation. This search is carried out in the claim ' s preamble (identified by the position of the so-called transitional phrase, a comma, or a colon – in that order), or in the entire claim text if no preamble is found. If there are several keyword matches, the keyword coming last in the selected text is used. If at least one such keyword is found, the claim is labeled as "process"; if not, the claim is labeled as "product". We then define process (product) patents as patents for which every independent claim is a process (product) claim. Patents that have both process and product claims are labeled as "mixed". We observe that the majority of patents in China are either product or mixed patents, i.e., they have at least one product claim. These account for more than 80% of all patents in the average year.

Geography. Panels (i)-(iii) of Table A.II categorize patents by the location of the primary patent applicant (based on the reported address). The shift in patenting activity away from overseas applicants (those with an address outside of China) toward domestic applicants (those with an address in China) is clear from panel (i): overseas applicants account for more than half of all patent applications from 1985-1989, more than two-thirds by the late 1990s, but only 11.2% from 2015-2019. Domestic patent applications are highly concentrated in the coastal regions of China and this concentration has been increasing over time. For example, the top five locations in China ranked by the total number of patent applications from 1985-2019 are Beijing and four coastal provinces: Jiangsu, Guangdong, Zhejiang, and Shandong.[18] These five locations accounted for around 35% of all invention patent applications in China before 2005 and around 55% after 2005. The role played by the next top five locations in China – which are mainly

[18] Locations within China are classified as administrative divisions at the province level. Most of these are provinces (e.g., Jiangsu, Guangdong) but some are direct-administered municipalities (e.g., Beijing, Shanghai).

Table A.I: Basic patent statistics: IPC sections and product vs. process

	’85-’89	’90-’94	’94-’99	’00-’04	’05-’09	’10-’14	’15-’19	all years
all invention patents	41	75	182	482	1,203	2,181	7,129	11,292
(i) invention patents by main IPC section								
A: Human Necessities	8	21	41	85	179	341	1,155	1,831
	(18.3)	(28.2)	(22.8)	(17.6)	(14.9)	(15.6)	(16.2)	(16.2)
B: Performing Operations	9	14	28	62	158	357	1,379	2,005
	(21.6)	(17.8)	(15.4)	(12.8)	(13.1)	(16.4)	(19.3)	(17.8)
C: Chemistry; Metallurgy	9	15	32	82	199	362	1,035	1,735
	(22.7)	(20.5)	(17.7)	(17.0)	(16.6)	(16.6)	(14.5)	(15.4)
D: Textiles; Paper	1	2	4	9	21	39	117	193
	(2.5)	(2.4)	(2.0)	(1.9)	(1.7)	(1.8)	(1.6)	(1.7)
E: Fixed Constructions	2	3	4	13	32	78	303	435
	(3.7)	(3.3)	(2.4)	(2.6)	(2.7)	(3.6)	(4.3)	(3.9)
F: Mechanical Engineering	4	5	12	31	92	185	530	859
	(8.8)	(7.3)	(6.6)	(6.6)	(7.7)	(8.5)	(7.4)	(7.6)
G: Physics	5	8	27	93	230	398	1,529	2,290
	(12.6)	(10.7)	(14.8)	(19.3)	(19.1)	(18.2)	(21.4)	(20.3)
H: Electricity	4	7	33	107	291	422	1,080	1,945
	(9.9)	(9.9)	(18.3)	(22.3)	(24.2)	(19.3)	(15.2)	(17.2)
(ii) invention patents by process/product type								
process	7	15	27	78	217	405	1,243	1,990
	(16.6)	(20.0)	(14.7)	(16.1)	(18.0)	(18.6)	(17.4)	(17.6)
product	17	35	79	190	474	1,003	3,356	5,152
	(40.3)	(46.7)	(43.3)	(39.4)	(39.4)	(46.0)	(47.1)	(45.6)
mixed	11	25	76	214	503	751	2,507	4,086
	(27.1)	(33.3)	(41.9)	(44.4)	(41.8)	(34.4)	(35.2)	(36.2)
unknown	7	0	0	1	10	23	23	63
	(15.9)	(0.0)	(0.0)	(0.1)	(0.8)	(1.1)	(0.3)	(0.6)

Notes: All patent counts are reported in thousands. Parentheses report shares out of all invention patents.

provinces in central China – has also been increasing over time. For overseas patent applications, concentration is even higher. For instance, applications from the top five overseas locations – Japan, USA, Germany, South Korea, and Taiwan – represent around 80% of all overseas patent applications after 2000. As we document in detail below, there is also a significant shift over time away from patenting by overseas applicants and toward patenting by domestic applicants.

B Identifying applicant name types

We first assign each unique applicant name *appname* in the patent database an applicant type *appnametype* based on the following sequential keyword search.

1. If *appname* contains 解放军, assign *appnametype* “military”.

2. If *appname* contains {公司, 会社, 独立行政法人, 董事会, 集团, 有限责任, 股份, 企业, 公

Table A.II: Basic patent statistics: geography

	’85-’89	’90-’94	’94-’99	’00-’04	’05-’09	’10-’14	’15-’19	all years
	(ii) invention patents by domestic vs. overseas							
domestic	18	40	53	183	669	1,661	6,330	8,953
	(44.2)	(52.5)	(28.9)	(37.9)	(55.6)	(76.2)	(88.8)	(79.3)
overseas	23	36	129	299	534	519	796	2,336
	(55.8)	(47.5)	(71.1)	(26.1)	(44.4)	(23.8)	(11.2)	(20.7)
	(ii) domestic invention patents by region							
top 1-5 Chinese locations	6	14	19	81	378	946	3,407	4,851
	(34.6)	(35.9)	(36.5)	(44.2)	(56.4)	(56.9)	(53.8)	(54.2)
top 6-10 Chinese locations	4	7	10	40	132	330	1,408	1,929
	(19.6)	(17.7)	(18.3)	(22.0)	(19.7)	(19.8)	(22.2)	(21.5)
other Chinese locations	8	18	24	62	160	386	1,514	2,172
	(45.8)	(46.4)	(45.2)	(33.8)	(23.9)	(23.2)	(23.9)	(24.3)
	(iii) overseas invention patents by region							
top 1-5 overseas locations	15	26	98	235	429	421	632	1,857
	(67.1)	(72.2)	(76.3)	(78.7)	(80.4)	(81.0)	(79.4)	(79.5)
top 6-10 overseas locations	5	6	20	40	63	55	82	271
	(19.7)	(17.0)	(15.4)	(13.5)	(11.7)	(10.6)	(10.3)	(11.6)
other overseas locations	3	4	11	23	42	43	82	208
	(13.3)	(10.8)	(8.3)	(7.8)	(7.9)	(8.3)	(10.3)	(8.9)

Notes: All patent counts are reported in thousands. Parentheses report shares out of all invention patents (panel (i)), domestic invention patents (panel (ii)), and overseas invention patents (panel (iii)) in the same time period. In panels (ii) and (iii), locations are ranked based on total invention patent applications, 1985-2019. Top Chinese locations in order: Jiangsu, Guangdong, Beijing, Zhejiang, Shandong, Anhui, Shanghai, Sichuan, Hubei, and Shaanxi. Top overseas locations in order: Japan, USA, Germany, South Korea, Taiwan, France, Netherlands, Switzerland, UK, and Sweden.

室, 工司, 公社, 商店, 工业}, contains 厂 and has length greater than three characters, or ends in 站, assign *appnametype* “enterprise”.

3. If *appname* contains 部, 厅, 局 and has length greater than three characters, assign *appnametype* “government”.

4. If *appname* contains {大学, 学校}, assign *appnametype* “university”.

5. If *appname* contains {研究所, 技术院, 设计院, 检定所, 验所, 技术协会, 科院, 科所, 科研所, 检定所}, assign *appnametype* “institute”.

6. If *appname* contains 学院, assign *appnametype* “university”.

7. If *appname* contains 研究, assign *appnametype* “institute”.

8. If *appname* contains {中心, 医院, 协会, 制造院, 基金会, 工技协, 学会, 技术委, 技术室, 服务处, 委员会}, assign *appnametype* “organization”.

9. If *appname* has length no greater than three characters, assign *appnametype* “individual”.

10. If *appname* does not satisfy any of the above, assign *appnametype* “other”.

Note that the keyword search is sequential in the sense that if an applicant name has already been assigned an applicant type at some step in the process, it is removed from the set of applicant names that remain to be assigned at later steps in the process. For example, 中国科学院微生物研究所 (Institute of Microbiology, Chinese Academy of Sciences) is identified as an institute at step 5 while 福建工程学院 (Fujian College of Engineering) is identified as a university at step 6, even though both names contain the characters 学院.

The next step is to assign each patent an applicant name type *pappnametype* based on the applicant names listed in the patent application. Note that patent application name types *pappnametype* and application name types *appnametype* can differ because some patents have multiple applicants. The assignment of *pappnametype* is based on the following procedure.

1. If at least one applicant name type is “enterprise” and no applicant name types are in {“university”, “institute”, “organization”, “government”, “military”}, assign *pappnametype* “enterprise”.

2. If at least one applicant name type is “university” and no applicant name types are in {“company”, “institute”, “organization”, “government”, “military”}, assign *pappnametype* “university”.

3. If at least one applicant name type is “institute” and no applicant name types are in {“company”, “university”, “organization”, “government”, “military”}, assign *pappnametype* “institute”.

4. If at least one applicant name type is “organization” and no applicant name types are in {“company”, “university”, “institute”, “government”, “military”}, assign *pappnametype* “organization”.

5. If at least one applicant name type is “government” and no applicant name types are in {“company”, “university”, “institute”, “organization”, “military”}, assign *pappnametype* “government”.

6. If at least one applicant name type is “military” and no applicant name types are in {“company”, “university”, “institute”, “organization”, “government”}, assign *pappnametype* “military”.

7. If at least one applicant name type is “individual” and no applicant name types are in {“company”, “university”, “institute”, “organization”, “government”, “military”}, assign *pappnametype* “individual”.

8. If no applicant name types are in {“company”, “university”, “institute”, “organization”, “government”, “military”, “individual”}, assign *pappnametype* “other”.

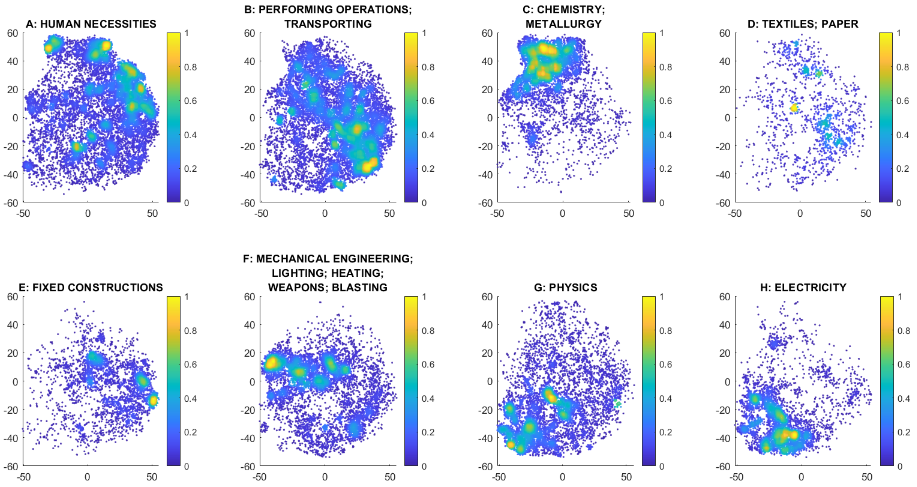
9. Assign all remaining patents *pappnametype* “mixed”.

C Visualization of patent abstract embeddings

To visualize the text embeddings of Chinese patent abstracts, we first reduce each 768-vector to two dimensions using a process known as t-Distributed Stochastic Neighbor Embedding (t-SNE), which is a technique for dimensionality reduction that is particularly well-suited for visualizing high-dimensional data. Figure A.I shows heatmaps of the two-dimensional t-SNE reduction of each abstract embedding by the main IPC section for each CNIPA patent from 1985 to 2000.[19] Evidently, patents belonging to the same IPC section tend to have abstract embeddings that exist in similar areas of the vector space. For example, the abstract embeddings for patents in sections G (Physics) and H (Electricity) – technology classes which are arguably more similar to each other than to other classes – tend to be clustered in the southwest corner of the two-dimensional space, whereas the embeddings for patents in section B (Performing Operations; Transporting) tend to be clustered in the southeast corner and those for section C (Chemistry; Metallurgy) tend to be clustered in the northwest corner. This observed clustering based on IPC section is indicative that the embeddings of the patent abstracts are picking up on differences and similarities in semantic meaning across patents in different technology classes.

[19] The t-SNE reduction of an vector depends on the set of all vectors that are being reduced. Since the computation time require for the t-SNE decomposition scales quickly with the number of vetors being reduced, we show this figure only for patents in the earlier years of our data.

Figure A.I: Two-dimensional t-SNE reduction of CNIPA patent abstract embeddings, 1985-2000



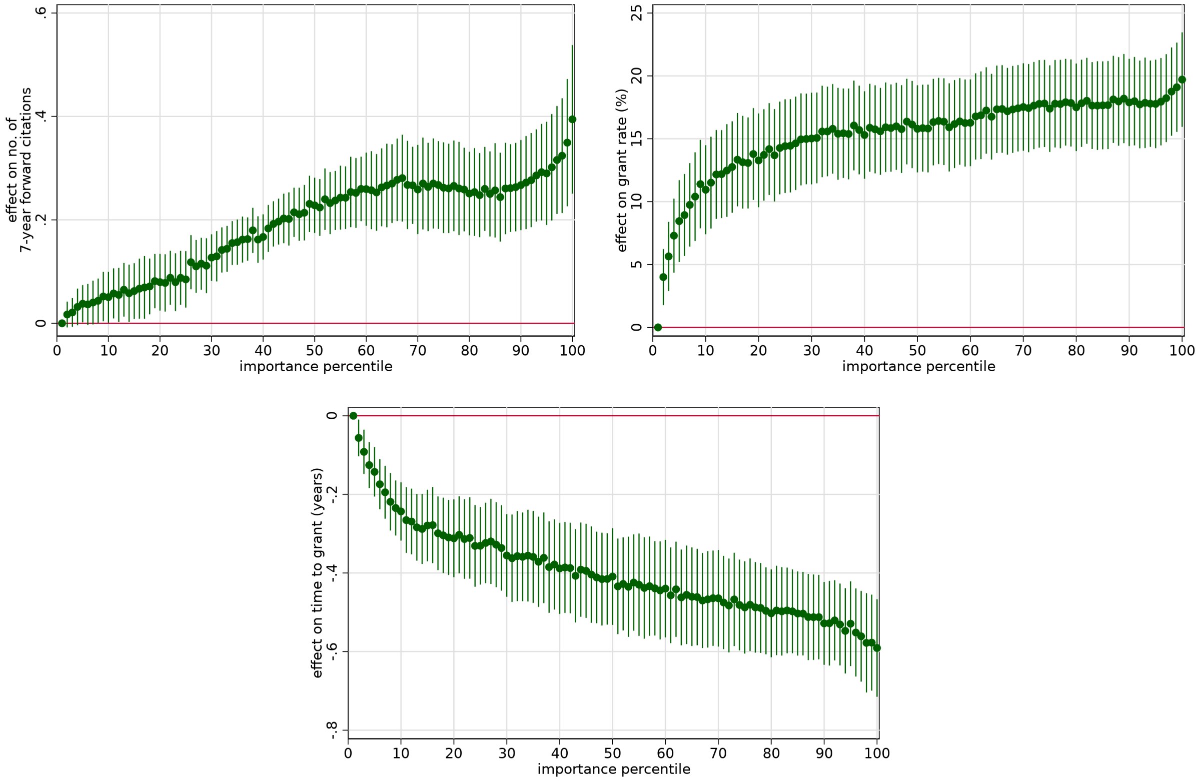
Notes: This figure shows the two-dimensional t-SNE reductions of the abstract embeddings for all CNIPA invention patents from 1985-2000 by the main IPC section of each patent.

D Scatter plots of patent importance versus other measures of quality

We investigate how our measure of patent importance correlates with other traditional measures of patent quality. We first compute the percentile of each patent's importance score amongst all patents applied for in the same calendar year. We then consider three different measures of patent quality – the number of forward citations received within seven years of application, grant status, and years between application and grant date – and regress each of these on the patent's within-year importance percentile. Since most of the backward citations that we observe appear in the early 2000s and we consider 7-year forward citations as an outcome, we restrict the sample for these regressions to all CNIPA invention patents applied for between 1997 and 2012. As the three quality outcomes typically differ across technology classes, patentee types, and time, we also include time-interacted dummy variables for a patent's main IPC subclass (4-digit) and the patentee type associated with the patent.

Figure A.II shows the estimated regression coefficients and associated standard errors (clustered at the 4-digit IPC level) for each importance percentile for the three quality measures. Our measure of patent importance is positively associated with the number of forward citations that a patent receives and the likelihood that the patent is granted. On the other hand, patent importance is negatively associated with the time that it takes for granted patents to be approved for granting, suggesting that more important patents are granted faster because there

Figure A.II: Relationship between importance, forward citations, grant status, and grant lag



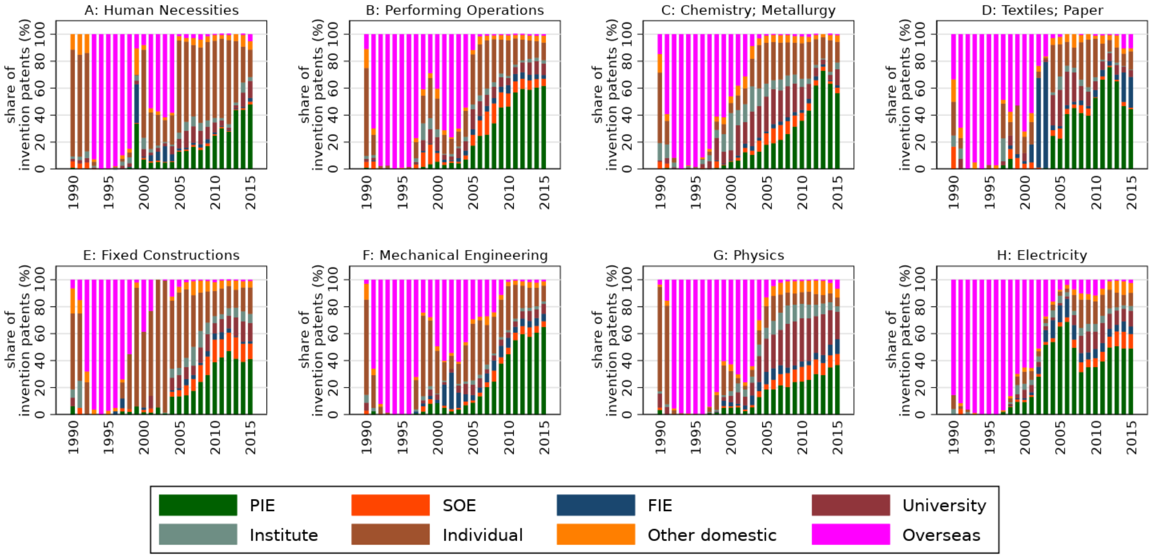
Notes: Each plot shows regression coefficients from a regression of a measure of patent quality – the number of forward citations received within seven years of application, grant status, and years between application and grant date – on the importance percentile of the patent within the year of the patent's application date. Each regression also includes time-interacted dummies for the patent's main 4-digit IPC code and the patentee type associated with the patent. Importance is defined as in equation (3.2). The sample for each regression is the set of CNIPA invention patents from 1997 to 2012. Bars indicate standard errors, which are clustered at the 4-digit IPC level.

is less relevant prior art to be examined. These relationships are not perfectly monotonic, but patent importance is clearly correlated in the way that one might expect with standard measures of patent quality that have been considered in the literature – a quite remarkable observation given that the importance score is based only on how patent text compares with text in the future versus text in the past.

E Trends in patent shares and importance by patentee type

Figure A.III shows the same information as panel (a) of Figure 2 in the main text – the share of patents accounted for by each patentee type within each IPC section – except that here we restrict attention to only the top 5% of patents in each IPC section and year based on patent

Figure A.III: Patents and shares by patentee type for



Notes: Each plot shows the share of invention patents accounted for by each patentee type, restricting attention only to the top 5% of patents in each IPC section and year based on patent importance.

importance. We observe that the decline in the share of overseas patents occurs more rapidly amongst these top patents compared with all patents. For example, by 2005, overseas patents account for less than 5% of the most important patents in each IPC section.

Figure A.IV shows the same information as Figure 1 in the main text – how the distribution of log importance (panel (a)) and the relative importance of patents by different patentee types (panel (b)) vary over time – but shows results for each IPC section separately, where patents are classified according to their main IPC section.

In panel (a), we observe that the timing of the decline in average importance in the early 2000s differs slightly across sections. For example, the declines occurred earlier in sections E (Fixed Constructions) and F (Mechanical Engineering). Nonetheless, in all IPC sections, average patent importance was lower in 2010 than in 2000. In addition, the reversal of the downward trend in patent importance after 2010 is most prominent in sections A (Human Necessities) and C (Chemistry; Metallurgy), where the importance of the average patent increases by about two standard deviations of the importance distribution within five years. Regarding the importance of Chinese patents for patenting at the USPTO, we observe that in section C (Chemistry; Metallurgy), there is an initial increase in the late 1990s, followed by a decline from 2002 to 2006 and then a second episode of rising importance from 2006 to 2009. In section G (Physics), there is a gradual rise beginning in the early 2000s, while in section H (Electricity), there is a brief episode of rising importance between 2002 and 2006, although this is followed by a period of decline.

In panel (b), we observe that the decline in the relative importance of overseas patents reverses in some sections (e.g., D (Textiles; Paper) and E (Fixed Constructions)) after 2010. Nonetheless, by 2015, overseas patents have lower average importance compared with patents applied for by the other patentee types in all IPC sections.

F Traditional measures of patent quality in China

Number of forward citations. Figure A.V shows how the average number of forward citations that CNIPA invention patents receive (from other CNIPA patents) changes over time. From 2000 onward, the majority of a patent's forward citations are received within 6 years of application and almost all citations are received within 8 years. The average patent applied for in 2000 receives around one citation within 6 years and this average is fairly constant until 2010. After 2010, however, this average declines steadily. This is largely due to censoring, since we only observe backward citations made by patents that are granted up to 2019, which highlights one of the key limitations of using forward citation counts as a measure of patent quality.[20]

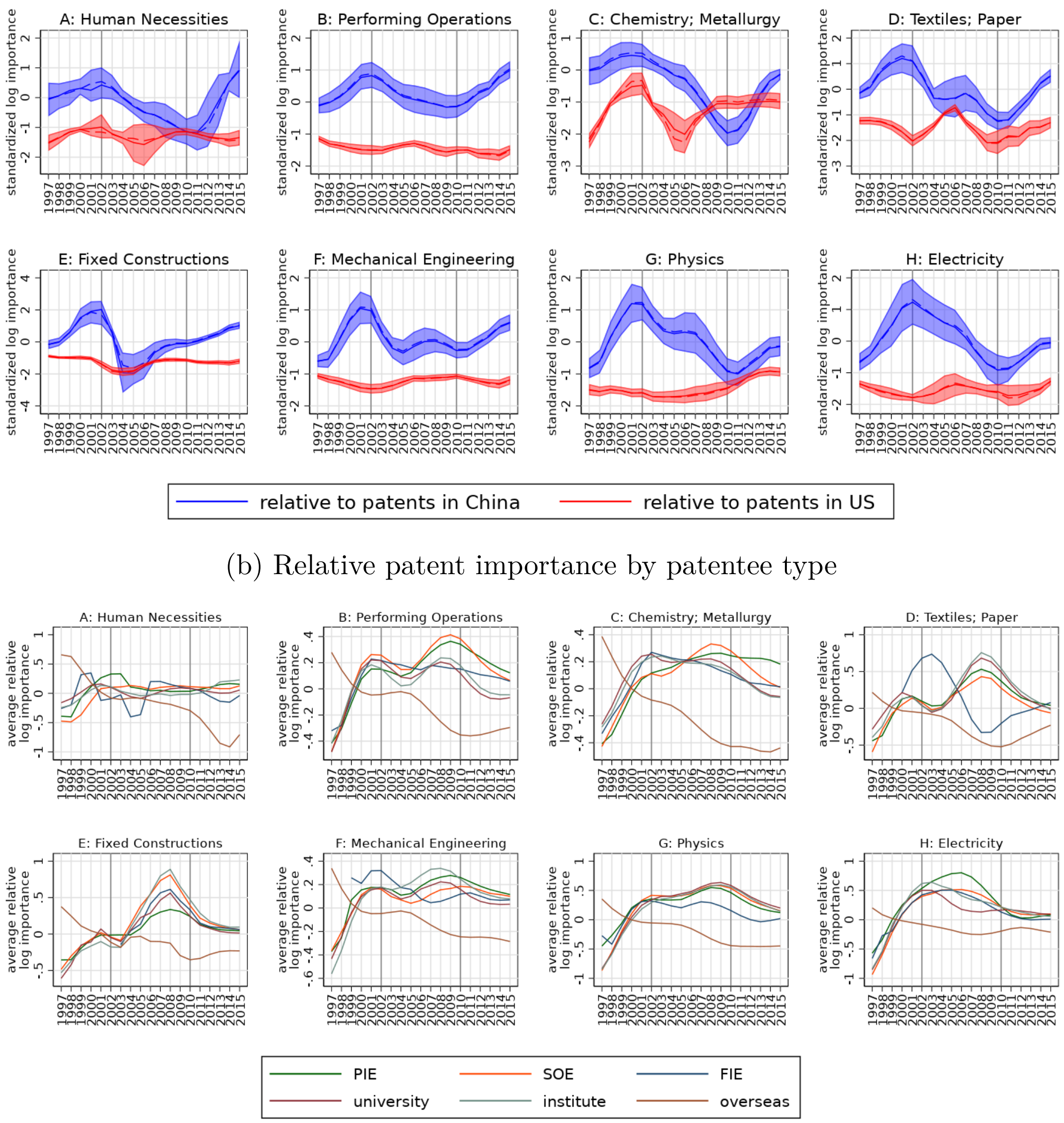
Citation centrality. A common approach to measuring patent quality is to compute measures of centrality based on citation linkages. To do this, we first define a directed graph where a node in the graph is a 3-digit IPC category and the strength of the link from node A to node B is the share of backward citations by patents in node B that cite patents in node A. Based on this graph, we then compute the eigenvector centrality of each node.[21] Figure A.VI shows how these centrality measures compare for patents from 2000-2010 versus patents from 2011-2019. We observe that eigenvector centralities are very constant over time – technology classes that are central from 2000-2010 also tend to be central from 2011-2019. Examples of IPC categories with high centrality scores are G01 (Measuring), G06 (Computing), H01 (Electric Elements), H04 (Electric Communication Techniques), and A61 (Medical/Veterinary Science). On the other hand, examples of IPC categories with low centrality scores are A22 (Butchering), A42 (Headwear), C13 (Sugar), and D07 (Ropes).

Legal status changes. After a patent application is submitted, various events can occur that alter its legal status. We can then categorize patents into eight mutually exclusive cases based on the most common legal status change events. First, patents that have not been granted fall

[20] There is also censoring in the earlier years of our sample, as we observe fewer backward citations per patent application for patents that are applied for before 2000. Hence, patents applied for in earlier years only receive citations much further into the future.

[21] The eigenvector centrality of a node i is defined recursively as $\lambda c_i = \sum_{j=1}^{N} w_{ij} c_j$, where N is the number of nodes in the graph, w_{ij} is the weight of the link from node i to j, and λ is the largest eigenvalue of the weighted adjacency matrix (a matrix with ij-element equal to w_{ij}).

Figure A.IV: The evolution of patent importance by patentee type

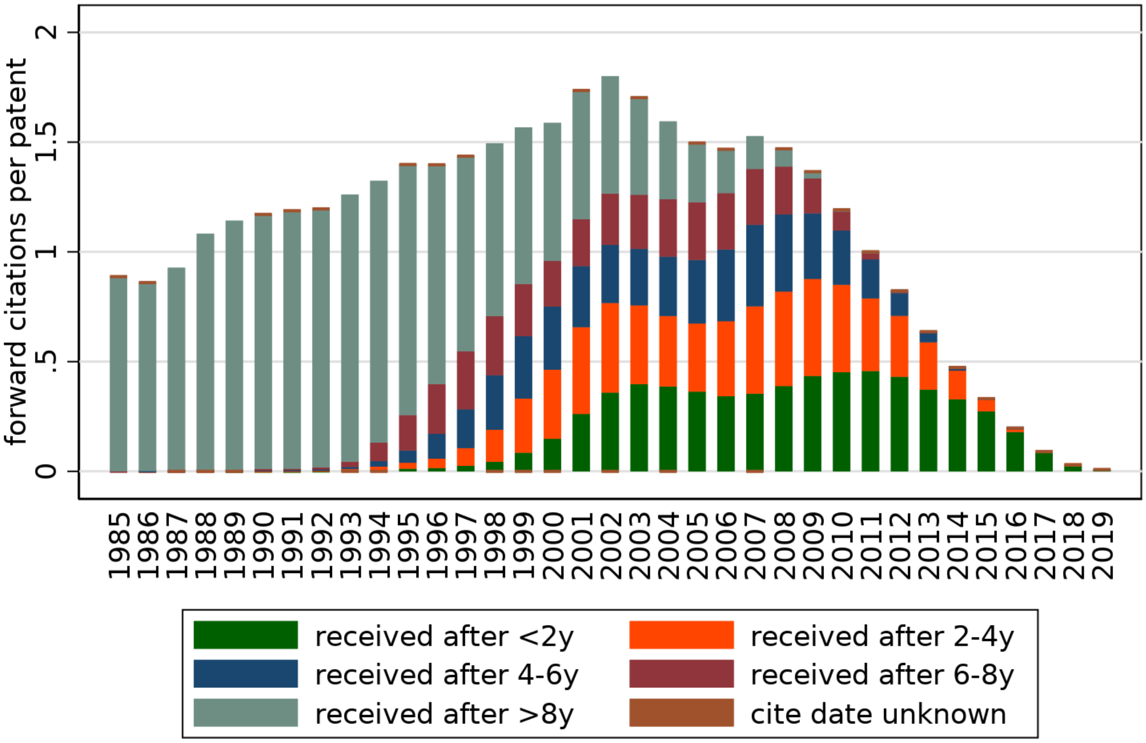


(a) Average patent importance

(b) Relative patent importance by patentee type

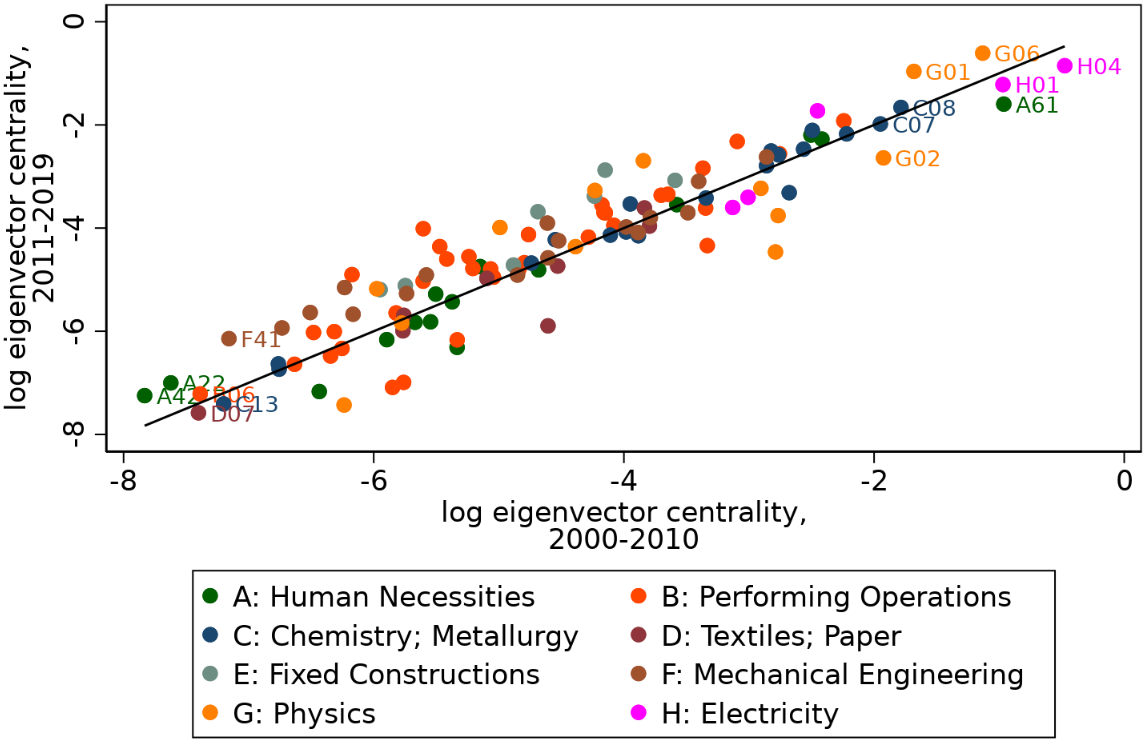
Notes: Panel (a) shows the mean (blue solid line), median (blue dotted line), and interquartile range (blue shaded area) of the log importance distribution for patents in each year and IPC section. The corresponding lines and areas in red show the same outcomes but for the importance of CNIPA patents for USPTO patents. Panel (b) shows the average importance for patents applied for by different patentee types relative to the average importance within each year and IPC section. All importance scores are standardized relative to the distribution of importance of CNIPA patents for CNIPA patents across all years.

Figure A.V: Average number of forward citations per patent



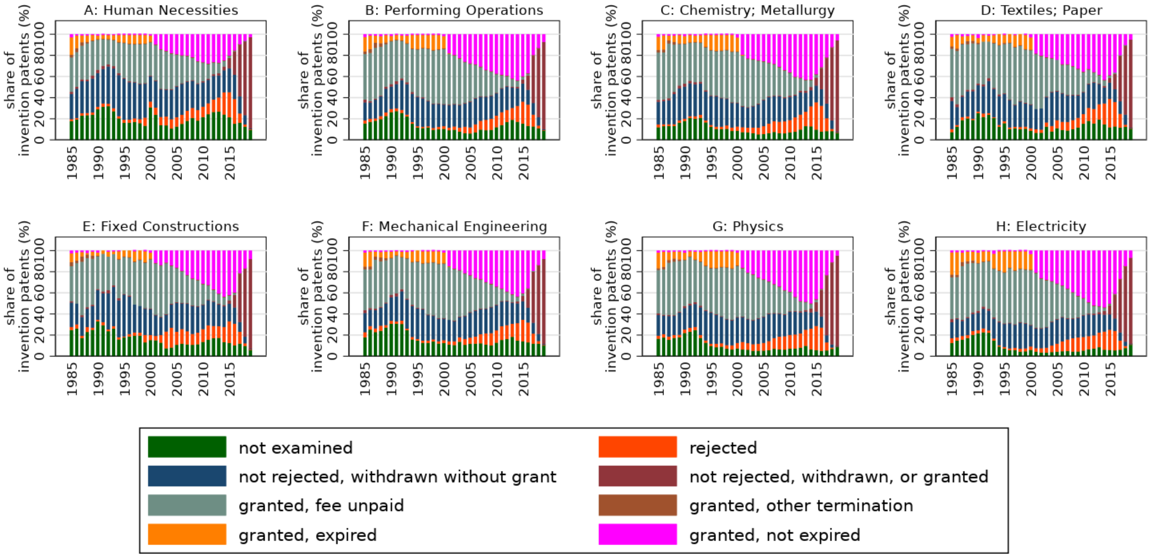
Notes: This figure shows the average number of forward citations that CNIPA invention patents receive from other CNIPA invention patents in each year. The bars are colored according to the time between the application date of the cited patent and the date that the citation is made.

Figure A.VI: Eigenvector centralities by IPC 3-digit category



Notes: This figure shows the eigenvector centralities in a directed graph where a node is a 3-digit IPC code and the strength of the link from node A to node B is the share of backward citations by patents in node B that cite patents in node A. The x-axis displays centrality scores for patents from 2000-2010 and the y-axis displays centrality scores of patents from 2011-2019.

Figure A.VII: Patent shares by legal status

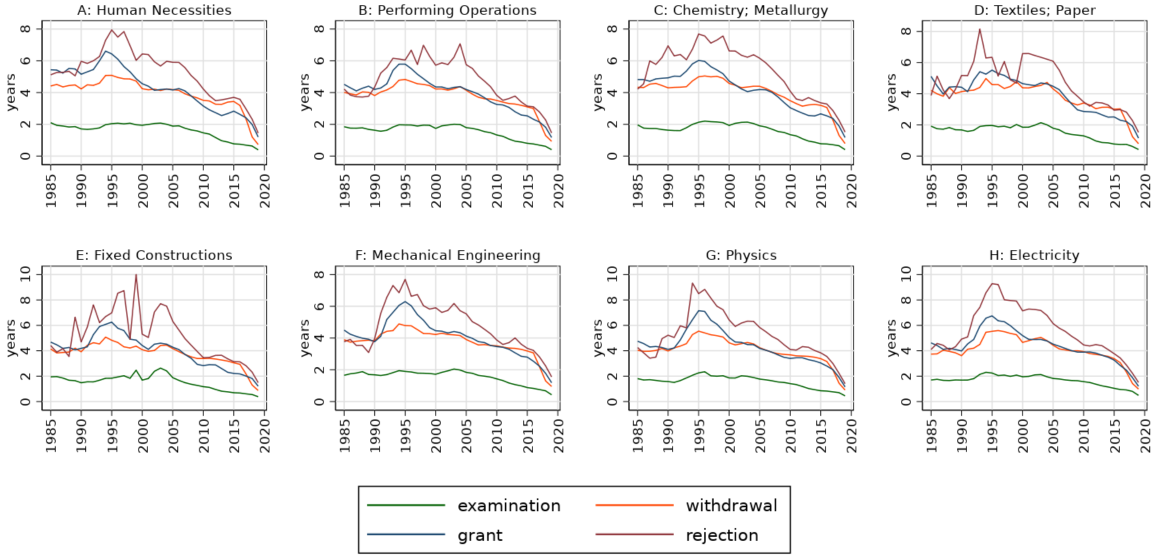


Notes: This figure shows the shares of invention patent applications within each IPC section and year based on eight mutually-exclusive legal status events.

into one of four categories: (i) those that have not been examined; (ii) those that have been examined but that have been rejected; (iii) those that have been examined, were not rejected, but were withdrawn before being granted; and (iv) those that were examined, have not been rejected or withdrawn, and have not been granted. Second, patents that have been granted also fall into one of four categories: (i) those that were terminated because of the failure to pay a renewal fee; (ii) those that were terminated for reasons other than unpaid renewal fees and expiration; (iii) those that were terminated due to expiration; and (iv) those that were still active as of the latest date in our data sample.

Figure A.VII shows the share of patents that fall into each of these eight categories in each year, by the main IPC section of each patent. Patent grant rates are between 40-60% in most cases. We observe increasing grant rates in all IPC sections throughout the 1990s, followed by falling grant rates from 2000 to 2015. Many ungranted patents never request examination and many are withdrawn without being rejected. The patent rejection rate also increases steadily from 2000 to 2015. For example, across all IPC sections, the rejection rate was less than 5% in 2000 but around 20% in 2015. Amongst granted patents, the primary reason for termination of patent rights before the expiration of a patent is the non-payment of renewal fees. For example, more than two-thirds of the patents that were applied for in 2000 and that were eventually granted were terminated before expiration for this reason. This share falls steadily from 2000 to 2015, although this is due in part to censoring, since the average non-payment event occurs

Figure A.VIII: Timing of legal status events



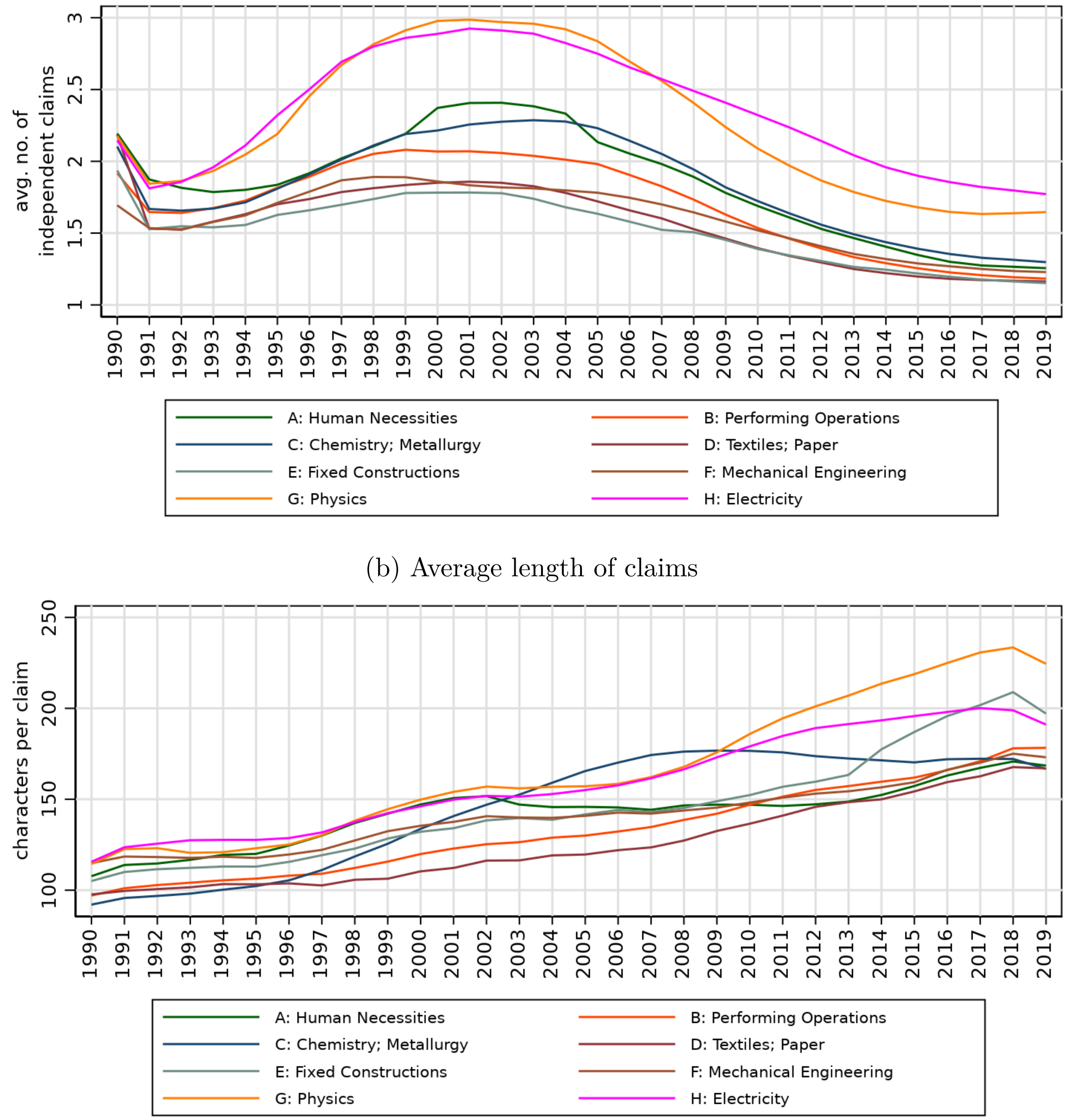
Notes: This figure shows the average time between the application date of a patent and the date that the patent is examined, withdrawn, granted, and rejected (conditional on each of these events occurring for a patent). Average durations are shown by IPC section and application year.

around ten years after the application date of a patent.

Timing of legal status change events. Figure A.VIII shows the average number of years that it takes for examination, withdrawal, granting, and rejection to occur after the application date of a patent. An important observation is that the average time taken for a patent to be granted falls steadily from an average of around 6 years in 1995 to around 3 years in 2015, a trend that is consistent across all IPC sections. This is due in part to the fact that the time taken for a patent to receive examination also falls after 2005.

Number and length of claims. Panel (a) of Figure A.IX shows the average number of independent claims per patent application by patent application year, while panel (b) shows the average number of Chinese characters per claim for the average patent application. Throughout the 1990s, the average number of independent claims per patent rises in all IPC sections. This is particularly prominent in sections G (Physics) and H (Electricity), where the average number of independent claims grows from just under two in 1991 to around three in 2000. Beginning in the early 2000s, however, we observe a steady decline in this average across all IPC sections. On the other hand, the average length of each patent claim (number of Chinese characters) rises steadily throughout our sample period in all IPC sections, increasing by between 50-100% from 1990 to 2019.

Figure A.IX: Number and length of claims per patent

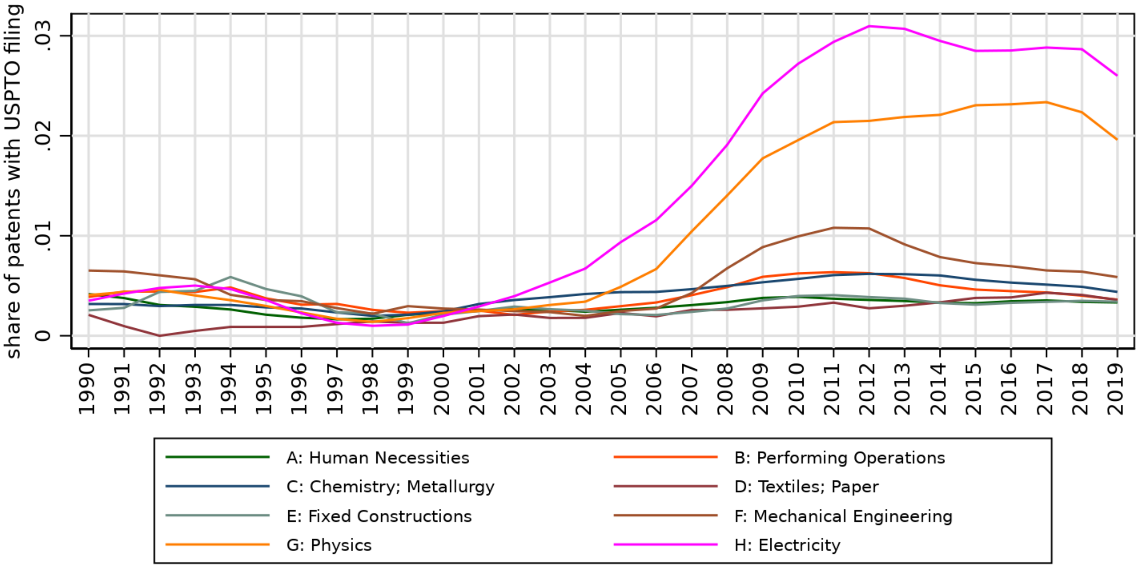


(a) Number of independent claims

(b) Average length of claims

Notes: Panel (a) shows the average number of independent claims associated with a patent. Panel (b) shows the average number of Chinese characters per independent claim for the average patent. All results are shown by by IPC section and application year.

Figure A.X: Share of patents with USPTO priority filings



Notes: This figure shows the share of CNIPA patents within each IPC section and application year that also have a priority filing at the USPTO.

Priority filings at the USPTO. Figure A.X shows the share of CNIPA invention patents that have a related priority filing at the USPTO. This share is typically very small – below 1% in most IPC sections throughout our sample – although there are significant increases in sections G (Physics) and H (Electricity). In these two sections, the share of patents with USPTO filings grows from around 0.3% in 2000 to 2.2% and 3.0% respectively by 2012. There is a smaller increase in this share in section F (Mechanical Engineering) as well.

G Decomposing changes in average patent importance

Consider a set of observations Ω_t with $|\Omega_t| = N_t$. Consider also a partition of Ω_t into J classes, $\{\Omega_{jt}\}_{j=1}^{J}$ with $|\Omega_{jt}| = N_{jt}$. Suppose we are interested in the average of some variable X_t over all observations $i \in \Omega_t$:

$$\bar{X}_t = \frac{1}{N_t} \sum_{i \in \Omega_t} X_{it} \tag{G.1}$$

We can write this equivalently as:

$$\bar{X}_t = \sum_{j=1}^{J} s_{jt} \bar{X}_{jt} \tag{G.2}$$

where $s_{jt} \equiv \frac{N_{jt}}{N_t}$ is the share of observations accounted for by class j and $\bar{X}_{jt} \equiv \frac{1}{N_{jt}} \sum_{i \in \Omega_{jt}} X_{it}$ is the average outcome within class j. Totally differentiating, we obtain:

$$\hat{\bar{X}}_t = \sum_{j=1}^{J} \left[s_{jt} r_{jt} \hat{\bar{X}}_{jt} + s_{jt} \left(r_{jt} - 1 \right) \hat{N}_{jt} \right] \tag{G.3}$$

where $\hat{y} \equiv \frac{dy}{y}$ indicates the marginal log change in variable y and $r_{jt} \equiv \frac{\bar{X}_{jt}}{\bar{X}_t}$ is the ratio of the average outcome in class j to the average outcome overall.