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

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

March 2025

Abstract

We study the evolution of patenting in China from 1985-2019. We develop a new method to measure the importance of a patent for innovation, based on the use of a Large Language Model to process patent text data and a new theory of the innovation process. We also classify patent ownership using a comprehensive business registry. We highlight three insights. First, patents that are important for innovation have become less important on average. Second, knowledge within China has become more important than knowledge outside of China for directing innovation in China. Finally, knowledge produced by Chinese entities within China has become more important than knowledge produced by foreign entities.

*We are grateful to Dietmar Harhoff, Josh Lerner, Georg Licht, Mark Roberts, Heiwai Tang, and Xiaodong Zhu for their helpful comments and discussions. We thank participants at the HKIMR-ECB-BOFIT joint conference on Europe, Asia and the Changing Global Economy, NBER-Chinese Economy Working Group Fall Meeting 2024, 10th ZEW/MaCCI Conference on the Economics of Innovation and Patenting, and 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. Dai thanks funding support from the National Natural Science Foundation of China (grant numbers 72342031 and 72203252). 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

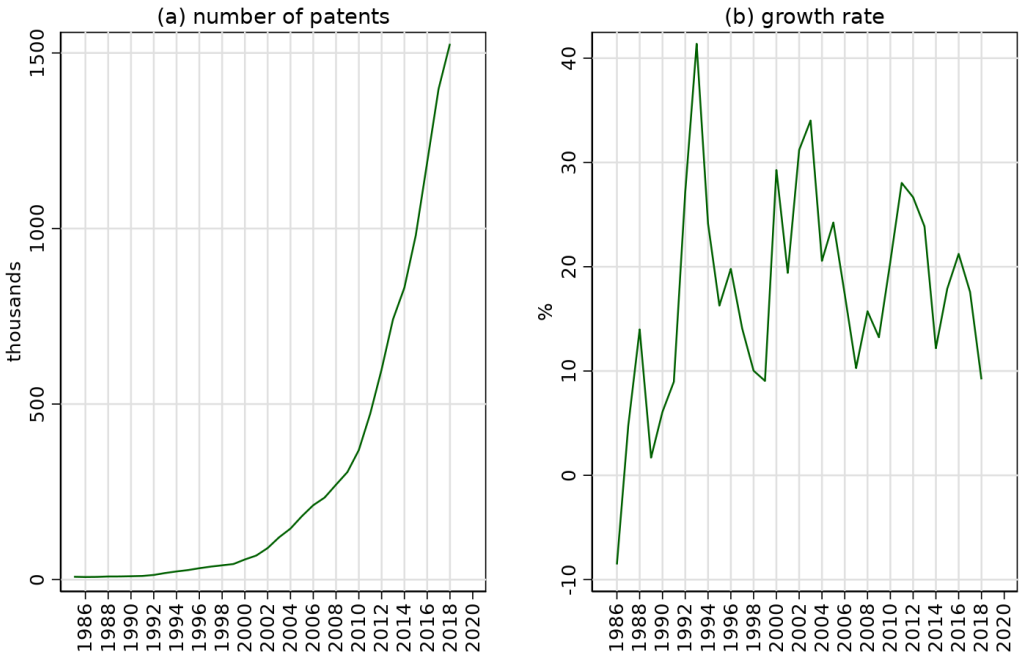
Patenting activity in China has accelerated substantially over the last two decades, far outpacing growth in places like the US. Figure 1.1 shows the number of invention patents applied for at the China National Intellectual Property Administration (CNIPA) in each year from 1985 to 2018 and the associated annual growth rates.[1] From 2000 to 2018, for example, 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). 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 how the quality of Chinese patenting has changed over time and quantify which sources of knowledge have been important for driving innovation in China.

Our analysis requires overcoming three important challenges. First, patents represent a vast and important source of codified knowledge, but one that is also 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 text 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 similarity between patents (e.g., the cosine similarity between the patent text embeddings), which in turn offer new ways to integrate patent text data into an analysis of innovation.

The second challenge, however, is that there is little existing theory about how patent text embeddings should be used to measure patent quality. To guide our analysis, we develop a simple model of innovation where the patent text embeddings are drawn from a dynamic stochastic process. The key idea is that the mean of the embedding distribution in the future is determined by the embeddings of patents in the past. This simple structure leads to novel insights about how to estimate the importance of a patent for innovation. Intuitively, we can now measure precisely the direction of innovation (i.e., the average direction in which patent embeddings are moving over time). Past patents whose embeddings are more similar to the direction of innovation are then inferred to be more important for the innovation process. As we show below, this new measure of patent importance is not only positively correlated with traditional measures of patent quality (e.g., citations and grant rates) but is also a stronger predictor of TFP and output at the firm level than traditional quality measures. Our approach thus provides a new metric to assess the importance of different sources of knowledge for Chinese innovation. For instance, we use this to compare the importance of patents in China versus patents outside

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

Figure 1.1: Number and growth rate of patents in China



of China (filed at the USPTO).

The third and final challenge is that standard patent data provide very little information about patentees beyond basic names and addresses. Hence, even if one can precisely measure the importance of a patent, it remains challenging to determine which types of patentees have been producing more important patents. To make progress on this front, 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). We can then compare the growth, composition, and importance of patenting across these patentee types.

Our analysis leads to three novel insights. First, patents that are important for innovation have become less important on average, which is due in part to the fact that patenting in China has become less innovative overall and more crowded over time. Second, knowledge within China has become more important than knowledge outside China for driving the direction of innovation in China. Finally, knowledge produced by Chinese entities within China (PIEs, SOEs, universities, and research institutes) has become more important than knowledge produced by

foreign entities (FIEs and overseas patentees).

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 describes the generation of patent text embeddings and shows how these can be used to measure the similarity of patenting activity in China and outside of China. Section 4 develops our model of innovation and shows how the theory leads to a new measure of patent importance based on the use of text embeddings, while section 5 describes how we take the model to the data. Section 6 provides validation of our new importance measure by showing how it correlates with patent citations and grant rates, as well as how it predicts TFP and output at the firm level. Section 7 then presents our main findings and section 8 concludes.

2 Data and basic patent statistics

We study an administrative database covering all invention patents applied for at the CNIPA from 1985 to 2019.[2] This includes Patent Cooperation Treaty (PCT) invention patents, which are filed almost exclusively by overseas applicants.[3] For brevity, we will henceforth refer to an invention patent application simply as a "patent".[4]

2.1 Classifying patentee types

Patent application documents 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 determine 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 name of each applicant in a patent to identify the *entity type* of the

[2]Note that our data is more comprehensive and accurate than some other databases for Chinese patents, e.g. downloaded from the CNIPA website. We restrict attention to invention patents, since these are typically considered to be the most innovative. 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%).

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

[4]Several patent offices provide online resources on the Chinese patent system, including patenting procedures and fees (CNIPA), patent examination and laws/policies (WIPO), China's patent numbering system (EPO), and an IPR toolkit for China (USPTO).

applicant. This allows us to differentiate between enterprises, universities, research institutes, individuals, and other domestic patentees (e.g., military). 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.[5] Finally, we combine the information about location, entity type, and ownership type to 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. Appendix A provides a detailed explanation of this classification procedure.

This taxonomy allows 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. For instance, Figure 2.1 shows the share of patents applied for in each year by patentee type. Before the mid-2000s, overseas patentees account for more than half of all patent applications in almost every year. However, from the early 2000s onward, there is rapid growth in the patenting activity by PIEs, which quickly become the dominant patentee type in terms of patent shares. By the late 2010s, PIEs account for around 40% of patent applications, whereas overseas patentees account for only around 10%. In contrast, the share of overseas patentees among granted USPTO utility patents actually increases from around 45% to 55% during the same time period (see Figure A.I in Appendix C). We also observe that the role of Chinese universities and SOE patentees increases slightly over time.

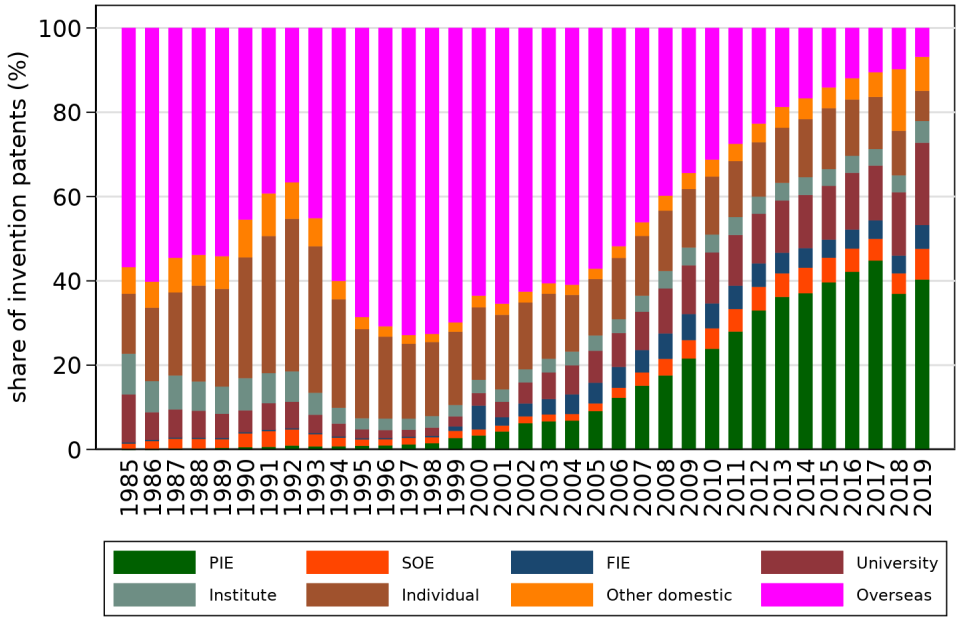
2.2 Technology classes

The most basic source of information about the knowledge that a patent represents is provided by a set of technology class codes reported by each patent, referred to as the International Patent Classification (IPC) codes. For example, IPC code "A01B 1/00" indicates that a patent is related to human necessities ("A"), specifically agriculture and forestry ("01"), more specifically soil working ("B"), and even more specifically hand tools ("1/00"). A patent can and often does report more than one IPC code but the main IPC code of each patent is also observed.

Figure 2.2 shows the share of patents applied for in each year by the section (first character) of each patent's main IPC code. Note that the technological composition of Chinese patents has been quite stable over time. The share of patents by 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)

[5] Of all the invention patents that we identify as having an "enterprise" entity 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 and are classified as "Other Domestic". Hence, the match between the patent data and the business registry is of reasonably high quality.

Figure 2.1: Patent shares by patentee type



(15.4%), F (Mechanical Engineering) (7.6%), E (Fixed Constructions) (3.9%), and D (Textiles; Paper) (1.7%).

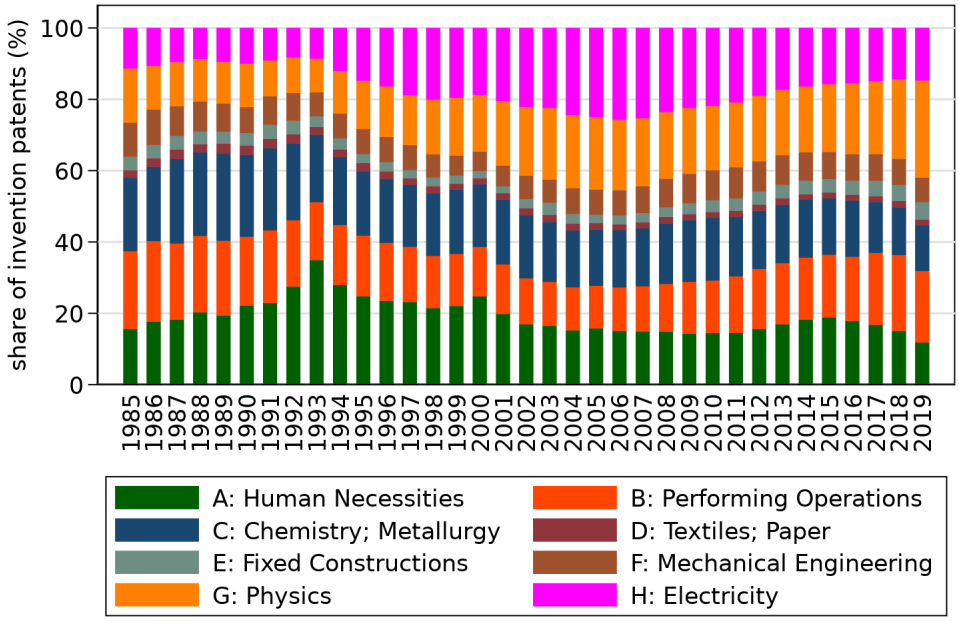
3 Patent text embeddings and patent similarities

3.1 Patent text embeddings

A patent's IPC codes are a coarse indicator of the knowledge that the patent embodies. Even at the finest level of disaggregation (e.g., "hand tools"), it is difficult if not impossible to discern from the IPC codes alone what the patent is really about. This information is contained, of course, in the textual descriptions of the patent – specifically, in the patent's abstract (a summary of the key elements of the patent) and claims (a detailed description of how the patent is claiming to be innovative). The text of a patent is hence a key source of codified knowledge, but this has traditionally been difficult to incorporate into quantitative analyses of patenting activity.

To make progress, we utilize an industry-leading LLM to generate *text embeddings* for the abstract and claims of every patent in the CNIPA database. Each embedding is a vector that represents the meaning of the 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 specifically is the *embed-multilingual-v2.0* model from

Figure 2.2: Patent shares by IPC section



Cohere, a Canadian technology company that specializes in natural language processing (NLP) and LLMs.[6] The model represents each text with a 768-element vector, thereby reducing the extremely high-dimensional text data to a smaller but still high-dimensional object. In practice, we embed the abstract and claims for each patent separately, use the abstract embeddings in our baseline analysis, and use the claims embeddings for robustness tests. To draw comparisons with patenting activity outside of China, we also generate embeddings for all granted utility patents at the USPTO.[7]

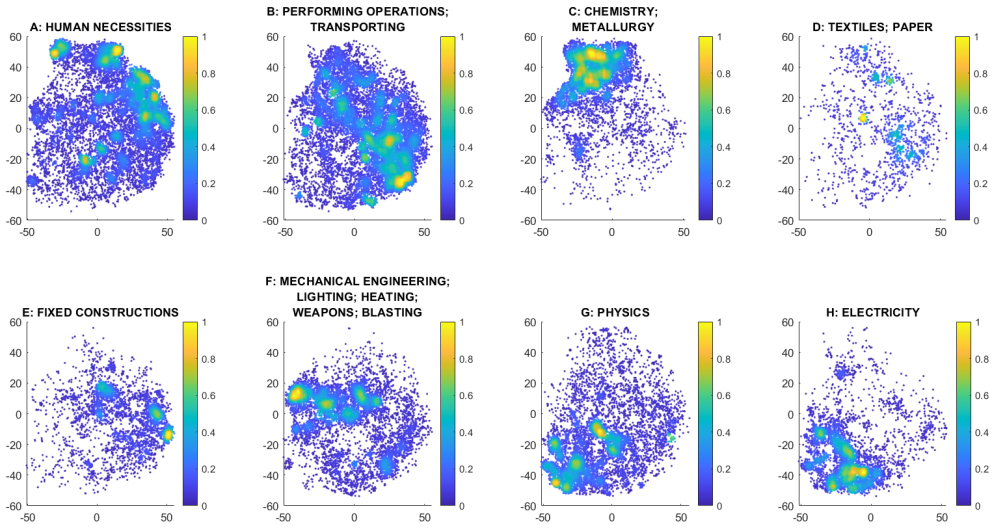
To visualize the text embeddings of the Chinese patent abstracts, we first reduce each 768-vector to two dimensions using a process known as t-Distributed Stochastic Neighbor Embedding (t-SNE), which is a technique for dimensionality reduction that is particularly well-suited for visualizing high-dimensional data. Figure 3.1 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.[8] Evidently, patents belonging to the same IPC section tend to have abstract embeddings that exist in similar areas of the reduced vector space. For example, the embeddings

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

[7]We focus on granted patents for the USPTO because data for these patents are more readily available. For instance, data for patent grants are available as far as back as 1976, whereas application documents are only available from the early 2000s onward.

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

Figure 3.1: 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.

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 differences and similarities in semantic meaning across patents in different technology classes.

It is also important to emphasize that the use of embeddings to measure the "meaning" of a patent is fundamentally different from approaches to textual analysis that are based on counts of keywords, such as the Term Frequency-Inverse Document Frequency (TF-IDF) method (e.g., as used by Younge and Kuhn (2016) and Kelly et al. (2021)). To illustrate, consider two hypothetical scenarios. In the first, all patents in the past contain only the key phrase "automobile", while those in the future contain only the keyword "electric vehicle". In the second, all patents in the past contain only the key phrase "horse carriage", while those in the future again contain only the keyword "electric vehicle". It is obvious that there is a greater leap in innovation in the second scenario compared with the first. However, keyword-based methods would not be able to discern this – one can only tell that the keywords in the past and future are different, but not *how different* they are. In contrast, the embeddings for "automobile", "electric vehicle", and "horse carriage" can be used to measure precisely the "distance" between

these keywords.

In the context of text embeddings, a widely-used measure of (inverse) distance is *cosine similarity*, which is defined as follows. Consider two patents, i and j, and let $\vec{x}_i \equiv \left[x_i^1 \cdots x_i^K\right]'$ and $\vec{x}_j \equiv \left[x_j^1 \cdots x_j^K\right]'$ denote the embeddings of the abstracts for these two patents (where $K = 768$ is the dimension of each embedding). The cosine similarity of $\vec{x}_i$ and $\vec{x}_j$ is:

$$c\left(\vec{x}_i, \vec{x}_j\right) = \frac{\vec{x}_i \cdot \vec{x}_j}{||\vec{x}_i||||\vec{x}_j||} = \frac{\sum_{k=1}^{K} \vec{x}_i^k \vec{x}_j^k}{\left[\sum_{k=1}^{K} \left(\vec{x}_i^k\right)^2\right]^{\frac{1}{2}} \left[\sum_{k=1}^{K} \left(\vec{x}_j^k\right)^2\right]^{\frac{1}{2}}} \tag{3.1}$$

The cosine similarity of any two vectors lies in the interval $[-1, 1]$ and is closer to 1 when the two vectors are pointing in more similar directions in the vector space.[9] This metric has the advantage of being invariant to the magnitudes of the embeddings, which tends to reflect the length of the text being embedded (in practice, we simply normalize each embedding $\vec{x}_i$ to have unit length, so that $c\left(\vec{x}_i, \vec{x}_j\right)$ is simply the dot product of $\vec{x}_i$ and $\vec{x}_j$). Now, for example, using the Cohere *embed-multilingual-v2.0* model, we find that c ("horse carriage", "electric vehicle") $= 0.90$ whereas c ("automobile", "electric vehicle") $= 0.93$, indicating that the model is able to recognize how different the different keywords are. Note that the standard deviation of cosine similarity scores between pairs of patent abstract embeddings in our data is on the order of 0.01, so a difference of 0.03 is quite large.

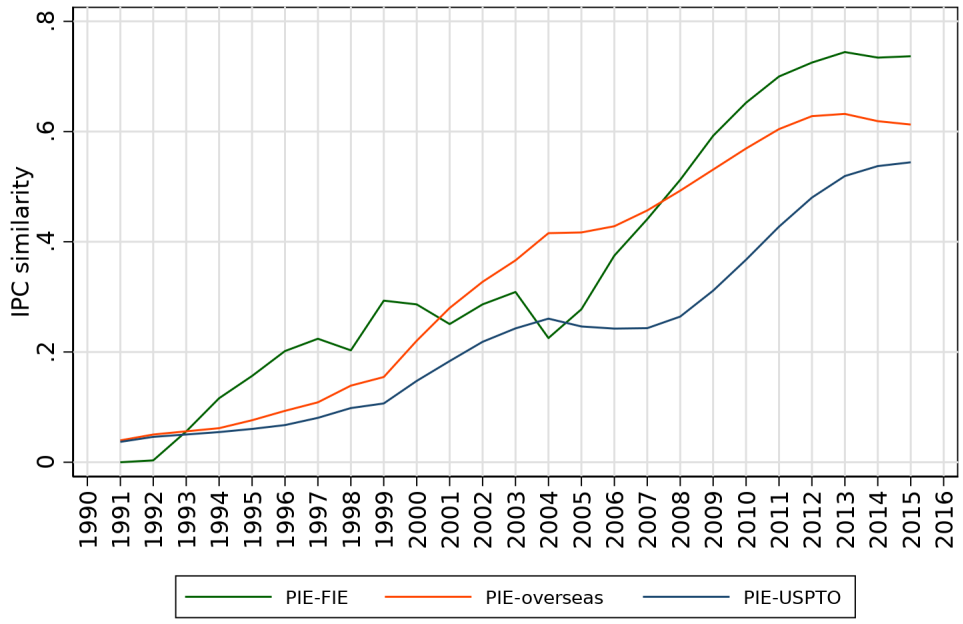
3.2 Patent similarities

To further illustrate the point that patent embeddings are able to capture information that is fundamentally different from that contained in traditional patent data, we now briefly examine how the similarity between patenting activity at the CNIPA and the USPTO has changed over time. The results that we present here will also be helpful in drawing a contrast between patent similarity and the measure of patent importance that we develop below.

Without relying on patent text data, the best that one could do would be to measure the similarity of patenting activity based on the technology classes associated with a patent. Specifically, for a given year and set of patents (e.g., those applied for by PIEs at the CNIPA or those granted at the USPTO), we compute the share of patents across IPCs. For each year and set of patents, this gives us a vector with dimension equal to the total number of unique IPCs and with each element corresponding to the share of patents accounted for by each IPC. We then compute the cosine similarity between these vectors across different pairs of patent sets in each year. Note that we do this using the full IPC code and hence we are extracting as much information as possible from these codes. Figure 3.2 shows how these similarities have changed

[9] When $K = 2$, $c\left(\vec{x}_i, \vec{x}_j\right)$ is the cosine of the angle between the vectors $\vec{x}_i$ and $\vec{x}_j$, hence the name.

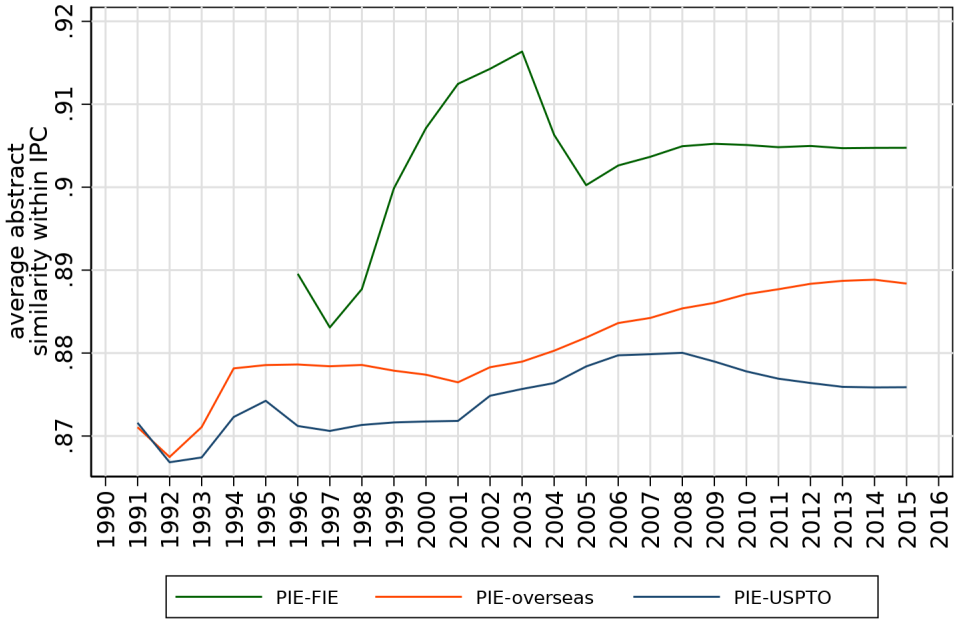
Figure 3.2: Patent similarities based on shares across IPCs



over time, comparing PIE patents with FIE, overseas, and granted USPTO patents (for visual clarity, we plot these measures as five-year rolling averages). There is a clear and persistent upward trend in all three similarity measures, suggesting that PIEs in China are converging toward patenting in more similar technologies as foreign entities.

With the patent text embeddings, however, we can also investigate how similar patenting activity is *within* IPCs. To do so, we simply compute the average cosine similarity of the abstract embeddings between different sets of patents within each IPC-year. Figure 3.3 shows these similarity scores averaged across IPCs within each year (weighted by how many pairs of patents are being compared in each IPC), again comparing PIE patents with FIE, overseas, and granted USPTO patents (and again plotted as five-year rolling averages). The picture that emerges is fundamentally different from one of clear convergence. We now see that within IPCs, PIE and FIE patenting become more similar in the late 1990s and early 2000s, but this similarity stops growing and even experiences a decline in later years. Furthermore, while there is slight growth in the similarity between PIE and overseas patents from 2001 onward, the similarity between PIE patents and granted USPTO patents stops growing around 2008 and declines thereafter. In sum, PIEs in China are patenting in more similar technology classes as foreign entities over time, but within a technology class, they are not necessarily patenting in more similar areas.

Figure 3.3: Patent similarities based on text embeddings within IPCs



4 A theory of patent importance

We now turn toward a central challenge: measuring the quality of a patent. 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)). Note, however, that none of these measures make direct use of the semantic content of patent text. Furthermore, the "quality" of a patent can mean many things: is a patent good for productivity, for example, or is it good for developing new products?

Since our interest in this paper is to investigate the sources of knowledge that drive innovation in China, we first focus on a specific notion of patent quality: a high-quality patent is one that is *important* for the direction of innovation. Note that important patents thus defined may not necessarily be the same as patents that are good for other economic outcomes, although below we will show that the measure of patent quality that we develop here is positively correlated with measures such as patent citations, firm TFP, and output.

4.1 A naive approach

How can patent text embeddings be used to determine which patents are important for innovation? A common conception in the literature is that a patent that is important for innovation should have two features: it should be similar to future patents ("impact") and different from past patents ("novelty"). For example, forward citations are by far the most widely used measure of impact, based on the assumption that if one patent cites another, then cited patent must have had some influence on the citing patent. Similarly, others have proposed that a lack of backward citations indicates that a patent is dissimilar to the past and hence is indicative of novelty (e.g., the so-called "radicalness" measure used by Ahuja and Lampert (2001) and Banerjee and Cole (2011)). More recently, Kelly et al. (2021) have also proposed that difference in the similarity of a patent's keywords to the future relative to the past is indicative of the patent's importance for innovation.

A candidate measure of the importance of a patent i applied for at time t based on the patent text embeddings would thus be:

$$\tilde{p}_{it} = \underbrace{c\left(\vec{x}_{it}, \vec{\mu}_{Ft}\right)}_{F_{it}} - \underbrace{c\left(\vec{x}_{it}, \vec{\mu}_{Bt}\right)}_{B_{it}} \tag{4.1}$$

where $\vec{x}_{it}$ is the embedding of patent i, $\vec{\mu}_{Ft}$ is an average embedding of some patents applied for after t, $\vec{\mu}_{Bt}$ is an average embedding of some patents applied for at or before t, and we let F_{it} and B_{it} denote the "forward" and "backward" cosine similarity of patent i, respectively.

There are two major limitations to measuring patent importance in this way, however. First, there is no theoretical underpinning for why this one particular transformation of the information contained in the embeddings is the best measure of importance. For example, why should F_{it} and B_{it} be equally weighted? Second, this approach evaluates the importance of a patent without regard for the existence of other contemporaneous patents that might also be important for innovation. Arguably, the set of competing ideas needs to be considered jointly to determine which particular ideas win out in shaping the future. Yet, without additional structure, it is not exactly clear how one should do this. Hence, we next develop a simple theory of patent importance.

4.2 The innovation process

Consider patenting in a single IPC class. We begin with the assumption that the embedding of each patent applied for in the class at time t is a random vector drawn from some distribution with mean $\vec{\mu}_t$ (a $K \times 1$ column vector) and support on the $(K-1)$-sphere (e.g., for $K = 2$, the support of the distribution is the unit circle). An example of such a distribution is the von-Mises Fisher distribution in $\mathbb{R}^K$, which is a multivariate normal distribution conditional on

the associated random vectors having unit length, although the exact functional form of the distribution will play no role in what follows.

We refer to $\vec{\mu}_t$ as the *state of knowledge* at time t, which captures the current location of patenting activity in the embedding space. Our goal is to understand how this state changes over time and to quantify the importance of each patent for such changes. To introduce the key ideas of the theory, we start with a parsimonious model of the innovation process, which we will enrich below before connecting the theory with the empirical analysis. Suppose then that the state of knowledge at $t+1$ is determined as follows:

$$\vec{\mu}_{t+1} = \rho_t \vec{\mu}_t + \sum_{i \in \Omega_t} p_{it} \vec{x}_{it} + \vec{\epsilon}_t \tag{4.2}$$

where $\vec{x}_{it}$ is the embedding of patent i applied for at time t, Ω_t is the set of patents applied for at t (with $N_t \equiv |\Omega_t|$), and $\vec{\epsilon}_t$ is a mean-zero error vector that is orthogonal to both $\vec{\mu}_t$ and $\{\vec{x}_{it}\}_{i \in \Omega_t}$. The future state of knowledge thus depends on its current state as well as on the ideas that are embodied in new patent applications. We refer to ρ_t as the *memory* of the innovation process, which is a standard AR(1) coefficient except that it is allowed to vary over time. Memory allows the state of knowledge at any time in the past or present to have an effect on the state of knowledge at any time in the future. We also refer to p_{it} as the *importance* of patent i, which will be our key measure of how important a patent is for innovation.

4.3 Patent importance

Since the error vector $\vec{\epsilon}_t$ is assumed to be orthogonal to the patent embeddings $\{\vec{x}_{it}\}_{i \in \Omega_t}$, the moment condition $\mathbb{E}\left[X_t' \vec{\epsilon}_t\right] = 0$ must hold, where X_t is a $K \times N_t$ matrix with column i equal to $\vec{x}_{it}$. The sample analogue of this moment condition can be written as:

$$X_t' X_t \vec{p}_t = X_t' \underbrace{(\vec{\mu}_{t+1} - \rho_t \vec{\mu}_t)}_{\vec{d}_{t+1}} \tag{4.3}$$

where $\vec{p}_t$ is the $1 \times N_t$ importance vector with element i equal to p_{it}. We define $\vec{d}_{t+1} \equiv \vec{\mu}_{t+1} - \rho_t \vec{\mu}_t$ as the *direction of innovation*, which captures the way in which the future is moving away from the present. Note that equation (4.3) is simply an ordinary least squares (OLS) moment condition, where $\vec{d}_{t+1}$ is the "dependent variable", each patent embedding $\vec{x}_{it}$ is an "independent variable" with coefficient p_{it}, and the "observations" are the K elements of the embeddings.

Now, note that since each patent embedding is normalized to have unit length, the matrix $X_t' X_t$ is simply the matrix of cosine similarities between the embeddings of patents applied for at date t. Furthermore, $X_t' \vec{\mu}$ for any unit vector $\vec{\mu}$ is simply a vector of cosine similarities between

the patent embeddings in X_t and μ. Hence, equation (4.3) can also be expressed as:

$$C_t \vec{p}_t' = \vec{F}_t' - \rho_t \vec{B}_t' \tag{4.4}$$

where C_t is the $N_t \times N_t$ cosine similarity matrix with (i,j)-element equal to $c(\vec{x}_{it}, \vec{x}_{jt})$, $\vec{F}_t$ is the $1 \times N_t$ forward similarity vector with element i equal to $c(\vec{x}_{it}, \vec{\mu}_{t+1})$, and $\vec{B}_t$ is the $1 \times N_t$ backward similarity vector with element i equal to $c(\vec{x}_{it}, \vec{\mu}_t)$. The i^{th} row of equation (4.4) in particular is:

$$p_{it} + \sum_{j \in \Omega_t \setminus \{i\}} c_{ijt} p_{jt} = F_{it} - \rho_t B_{it} \tag{4.5}$$

where $c_{ijt} \equiv c(\vec{x}_{it}, \vec{x}_{jt})$ is shorthand notation for the cosine similarity between patents i and j.

Equation (4.5) implies that patent i can be more similar to the direction of innovation if either patent i is more important or it is more similar to other important patents. In contrast with the "naive" measure of importance posited in equation (4.1), our measure now has two advantages. First, the relative weight on forward similarity F_{it} versus backward similarity B_{it} can be interpreted as depending on the memory of the innovation process (which we will estimate in the empirical implementation below). Second, equation (4.5) shows how it can be misleading to infer the importance of a patent based only on its own forward and backward similarity scores: a patent may look like it has high impact and novelty but may really just be similar to other patents that are truly impactful and novel. Our measure adjusts for this by taking into account the network of cosine similarities between patents.

To obtain a unique solution for the importance of each patent from equation (4.3), a necessary and sufficient condition is that the cosine similarity matrix $X_t{}' X_t$ is invertible. As we discuss below, invertibility will be a salient issue in taking the model to the data, but let us assume for now that invertibility holds in order to develop more intuition about what patent importance represents. With invertibility, the unique solution for the importance vector is:

$$\vec{p}_t' = \left(X_t' X_t\right)^{-1} X_t' \vec{d}_{t+1} \tag{4.6}$$

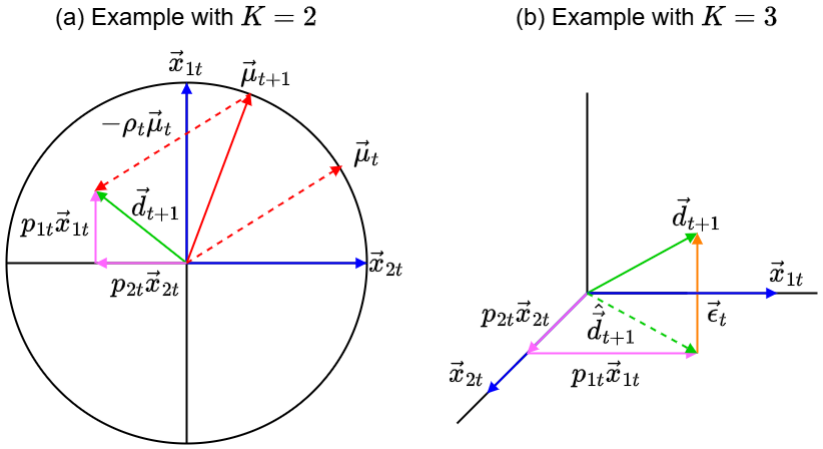
Hence, the direction of innovation predicted by the patent embeddings X_t is:

$$\hat{\vec{d}}_{t+1} = X_t \vec{p}_t' \tag{4.7}$$

$$= X_t \left(X_t' X_t\right)^{-1} X_t' \vec{d}_{t+1} \tag{4.8}$$

Now, note that $X_t \left(X_t' X_t\right)^{-1} X_t'$ is the *projection matrix* for the subspace spanned by the columns of X_t, i.e., the embeddings of the patents applied for at t. $\hat{\vec{d}}_{t+1}$ is thus the projection of the direction of innovation into this subspace and the importance scores $\{p_{it}\}_{i \in \Omega_t}$ are simply the unique set of weights on each patent embedding that allow us to construct this projection.

Figure 4.1: Illustration of patent importance

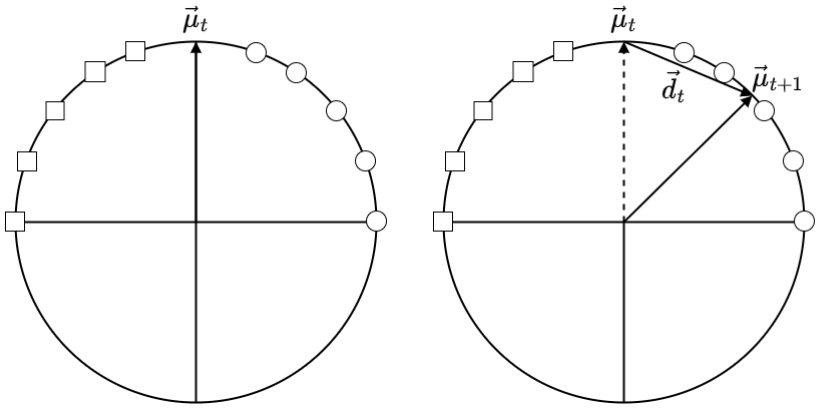


To illustrate, consider Figure 4.1. In panel (a), we show an example where $K = 2$ and there are exactly two patent embeddings, $\vec{x}_{1t}$ and $\vec{x}_{2t}$. In this case, the direction of innovation $\vec{d}_{t+1}$ already lies in the subspace spanned by $\vec{x}_{1t}$ and $\vec{x}_{2t}$. Hence, the projection $\hat{\vec{d}}_{t+1}$ is simply $\vec{d}_{t+1}$ itself and the magnitude of the error vector $\vec{\epsilon}_t$ is zero. It is clear that in this example, there is a unique pair of weights p_{1t} and p_{2t} that allows us to construct $\vec{d}_{t+1}$ using $\vec{x}_{1t}$ and $\vec{x}_{2t}$. Furthermore, note that the direction of innovation is pointing toward $\vec{x}_{1t}$ and away from $\vec{x}_{2t}$. Hence, patent 1 is assigned a positive importance score ($p_{1t} > 0$) and patent 2 is assigned a negative importance score ($p_{2t} < 0$).

In panel (b), we show a second example where $K = 3$ and we again have two patent embeddings, $\vec{x}_{1t}$ and $\vec{x}_{2t}$. In this case, the dimension of the embedding space is greater than the number of patents and it is possible for the direction of innovation not to lie in the subspace spanned by $\vec{x}_{1t}$ and $\vec{x}_{2t}$. The projection $\hat{\vec{d}}_{t+1}$, however, is the closest vector in this subspace to $\vec{d}_{t+1}$ and there is again a unique pair of weights p_{1t} and p_{2t} that allows us to construct $\hat{\vec{d}}_{t+1}$ using $\vec{x}_{1t}$ and $\vec{x}_{2t}$. The difference between $\vec{d}_{t+1}$ and $\hat{\vec{d}}_{t+1}$ is the error vector $\vec{\epsilon}_t$, which is orthogonal to the subspace spanned by $\vec{x}_{1t}$ and $\vec{x}_{2t}$ by assumption. In sum, one can think of the patent importance scores as weights on each patent that allow us to get "as close as possible" to the direction of innovation using the patent embeddings.

Finally, note that patent importance is best interpreted as a *relative* measure within a given year. To see why this is the case, suppose that the distribution of $\vec{x}_{it}$ is symmetric around its mean $\vec{\mu}_t$ (as it is in the von-Mises Fisher distribution, for example). In two dimensions, the distribution of $\vec{x}_{it}$ may then resemble the first diagram shown in Figure 4.2 below, where each patent embedding is represented by a square or circle. Now suppose that the the state of knowledge $\vec{\mu}_t$ were to change to some new state $\vec{\mu}_{t+1}$, for example by rotating clockwise as illustrated in the second diagram of Figure 4.2. Suppose also for exposition that $\rho_t = 1$ so that the direction of innovation $\vec{d}_t$ is simply the difference between $\vec{\mu}_{t+1}$ and $\vec{\mu}_t$. Note that the direction of innova-

Figure 4.2: Patent importance as a relative measure



tion points in the same direction as exactly half of the patent embeddings (the circles in Figure 4.2) and in the opposite direction as exactly half of the patent embeddings (the squares). Hence, the former group of patents would be assigned positive importance scores while the latter group of patents would be assigned negative importance scores. Furthermore, this is true regardless of the way in which the state of knowledge moves. As such, the average importance of patents in a given IPC-year will typically be close to zero and hence is an uninteresting statistic. Instead, what will matter is which patents within a given IPC-year are important and which are not.

4.4 Allowing for interactions across IPC classes and foreign influence

The version of the model described above makes two strong assumptions: that innovation within each IPC class occurs in isolation and that only knowledge in China is relevant for innovation in China. There are at least two features of our data, however, that strongly suggest these to be poor assumptions. First, as described above, patents often report secondary IPC codes besides their main IPC code, indicating that these other technology classes are also relevant for the knowledge embodied in a patent. Second, from the information about backward citations provided by each patent application document, we often see patents in one IPC class citing patents in other IPC classes. Similarly, we often see Chinese patents citing patents filed at foreign patent offices as well. Hence, we will now extend the model to relax both of these assumptions.

Let us first explicitly index a patent's main IPC class by g. Now consider innovation in IPC class g and suppose that instead of equation (4.2), the innovation process is described by:

$$\vec{\mu}^g_{t+1} = \rho^g_t \vec{\mu}^g_t + \sum_{i \in \Omega^g_t} p^g_{it} \vec{x}^g_{it} + \underbrace{\sum_{g' \in \Gamma_g} \gamma^{gg'}_t \vec{\mu}^{g'}_t}_{\text{IPC interactions}} + \underbrace{p^g_{US,t} \vec{\mu}^g_{US,t}}_{\text{foreign influence}} + \vec{\epsilon}^g_t \tag{4.9}$$

Note that we now allow the state of knowledge in other IPCs, $\vec{\mu}_t^{g'}$, to play a role, where Γ_g is the set of *related IPCs* that matter for innovation in g and $\gamma_t^{gg'}$ is the importance of g' for innovation in g. Similarly, we proxy for the influence of foreign knowledge on innovation in China by including the current state of knowledge for granted US utility patents in the same IPC class, $\vec{\mu}_{US,t}^g$, where $p_{US,t}^g$ measures the importance of US knowledge for innovation in China.

5 Empirical implementation

5.1 OLS identification and a LASSO estimator

Our goal is now to estimate the coefficients of the innovation process specified in equation (4.9) for every IPC and year: ρ_t^g, $\{p_{it}^g\}_{i\in\Omega_t}$, $\left\{\gamma_t^{gg'}\right\}_{g'\in\Gamma_g}$, and $p_{US,t}^g$. Given the discussion above, a natural approach would be to employ a simple OLS estimator. However, there are two issues with this method.

The first issue is that for identification to hold under OLS, all of the embeddings on the right-hand side of equation (4.9) must be linearly independent. This may fail to be true for two reasons. First, there are a small number of cases in which there are duplicate abstracts within the same IPC-year. In such cases, we simply remove the duplicates from the estimation and split the estimated importance score for the remaining unique embedding equally across duplicates. Second, since $\vec{\mu}_t^g$ is defined as the mean of $\{\vec{x}_{it}^g\}_{i\in\Omega_t^g}$, for any realization of the latter there is likely to be a high degree of multicollinearity with the former. Hence, for the empirical model, we measure $\vec{\mu}_t^g$ as the average embedding in IPC g in years t, $t-1$, and $t-2$ instead (normalized to unit length), denoting this now by $\vec{\mu}_{Bt}^g$ to avoid confusion. For consistency, we also do the same for the secondary IPC embeddings $\vec{\mu}_t^{g'}$ and $\vec{\mu}_{US,t}^g$, replacing these with $\vec{\mu}_{Bt}^{g'}$ and $\vec{\mu}_{US,Bt}^g$, respectively. Similarly, we measure the future state of knowledge $\vec{\mu}_{t+1}^g$ as the average embedding in IPC g in years $t+1$, $t+2$, and $t+3$ instead (normalized to unit length), denoting this now by $\vec{\mu}_{Ft}^g$.

A second challenge with the OLS approach is that a necessary condition for linear independence of the embeddings is that $\mathcal{M}_t^g \leq K$, where $\mathcal{M}_t^g = 2 + N_t^g + |\Gamma_g|$ is the number of coefficients being estimated (note that $\mathcal{M}_t^g$ is the number of "regressors" and K is the number of "observations", so this is just the standard OLS condition that one cannot have fewer observations than regressors). This poses an important challenge, because the dimension of the embedding space in the Cohere *embed-multilingual-v2.0* model is fixed at $K = 768$, whereas the number of patents N_t^g applied for in any given IPC-year may be much larger. For example, ignoring the other coefficients, the patent importance scores are only uniquely determined by equation (4.6) when the cosine similarity matrix is invertible, which requires $N_t \leq K$. Out of the 865,401 unique IPC-years in our data, only 987 IPC-years fail to meet this criterion, but these cases account for 12.7% of all patent applications.

To deal with this issue, we hence discard the OLS approach and instead make use of a Least Absolute Shrinkage and Selection Operator (LASSO) estimator. Specifically, the coefficients of interest, $\chi_t^g \equiv \left\{\rho_t^g, \{p_{it}^g\}_{i\in\Omega_t}, \left\{\gamma_t^{gg'}\right\}_{g'\in\Gamma_g}, p_{US,t}^g\right\}$, are estimated as follows:

$$\min_{\chi_t^g}\left\{\|\vec{\mu}_{Ft}^g - \hat{\vec{\mu}}_{Ft}^g\|_2^2 + \alpha\|\vec{p}_t^g\|_1\right\} \tag{5.1}$$

where $\hat{\vec{\mu}}_{Ft}^g \equiv \rho_t^g\vec{\mu}_{Bt}^g + \sum_{i\in\Omega_t^g} p_{it}^g\vec{x}_{it}^g + \sum_{g'\in\Gamma_g}\gamma_t^{gg'}\vec{\mu}_{Bt}^{g'} + p_{US,t}^g\vec{\mu}_{US,Bt}^g$ is the predicted future state of knowledge. This adds to the standard OLS sum-of-squared-errors objective (the first term in equation (5.1)) an ℓ_1-penalty (the second term) that promotes sparsity in the estimated patent importance scores, with the parameter α governing the strength of this penalty. This has a straightforward economic interpretation: the ℓ_1-norm of the importance vector (i.e., the sum of the absolute values of $\{p_{it}^g\}_{i\in\Omega_t^g}$) can be viewed as a measure of the cost of *attention* that inventors have to pay to the current set of patents when choosing the direction of innovation. Hence, the ℓ_1-penalty can be interpreted as a constraint that inventors have a limited attention span. Since only the embeddings $\{\vec{x}_{it}^g\}_{i\in\Omega_t^g}$ are at the individual patent level (whereas the other embeddings are averages across multiple patents), we apply the ℓ_1-penalty only to the importance vector $\vec{p}_t^g$ and not to the other coefficients in the estimation.

To select the value of α, a standard approach is to use *cross-validation*, where the dataset is first split into n groups, one of which is treated as test data and the remaining $n-1$ are treated as training data. Given a candidate value of α, the LASSO model is estimated using the training data and the loss (the value of the objective function) is then evaluated on the test data given the estimated coefficients. This is done n times, with each iteration using a different group of data as the test. The value of α is then selected numerically to minimize the average loss across all iterations. While this can be implemented for each IPC-year separately in principle, this greatly increases the computational time for generating the LASSO estimates in practice. Hence, we use cross-validation on a small random sample of IPC-years to pick a value for α that we then use for all IPC-years. This process leads to a value of $\alpha = 10^{-9}$, so the LASSO estimator is in a sense very close to the OLS estimator. For the cases in which OLS is identified (i.e., when $\mathcal{M}_t^g \leq K$ and there is no multicollinearity), we also compute the OLS estimates of the coefficients and compare them to the LASSO estimates for robustness.

5.2 The relevant set of secondary IPC classes

The final issue that needs to be addressed before taking the model to data is defining the related set of IPCs, Γ_g, for each IPC g. While one may wish to impose as few restrictions on this set as possible, one should also note that there are 75,025 unique IPC codes in our data and hence some *a priori* restriction on Γ_g is necessary given the discussion about identification and sparsity above. As mentioned, there are two sources of information that may be helpful in

determining Γ_g: the set of secondary IPCs listed by a patent and the set of IPCs other than its own that a patent cites.

Formally, let $S_{sec}^{gg'}$ and $S_{bcit}^{gg'}$ denote the shares of patents with main IPC g that list g' as a secondary IPC and that cite at least one patent with main IPC g', respectively. One might then deem $g' \in \Gamma_g$ if either $S_{sec}^{gg'}$ or $S_{bcit}^{gg'}$ is above a certain threshold. Note that the secondary IPCs and cited IPCs partially overlap. For example, on the set of all IPC pairs gg' where both $S_{sec}^{gg'} > 0$ and $S_{bcit}^{gg'} > 0$, the correlation between the two shares is 0.46. In other words, IPCs that are listed as secondary classes also tend to get cited.[10] However, the network of citations between IPCs is much sparser than the network defined by the secondary IPCs. For instance, $S_{sec}^{gg'}$ is strictly positive for around 4.5 million unique pairs of IPCs, whereas this number is only 1.8 million for $S_{bcit}^{gg'}$. Similarly, out of all 75,025 unique IPCs, 93% (69,503 IPCs) have patents that report at least one secondary IPC, whereas only 75% (56,513 IPCs) have patents that cite patents in at least one other IPC besides their own. In light of this, we choose to define Γ_g on the basis of the secondary IPC network.

Now, note that most patents report only a small number of secondary IPCs. For example, if we define Γ_g to be the set of IPCs where $S_{sec}^{gg'}$ is greater than 10%, which is already a fairly low threshold, then the average patent is in an IPC with $|\Gamma_g| = 2$ and the patent at the 99^{th} percentile of this measure has $|\Gamma_g| = 15$. Hence, we adopt a simpler rule and define Γ_g to be the top ten secondary IPCs for g on the basis of $S_{sec}^{gg'}$.[11]

One may be concerned that limiting the number of related IPCs to at most ten is imposing too strong a restriction on how technologies in different IPC classes can influence each other, especially since the maximum possible number of related IPC pairs is on the order of 5.6 billion (75,025 × 75,024) and we are allowing for only around 750,000. Nonetheless, even with this level of sparsity in the related-IPC network, around two-thirds of all IPCs are connected to each other either directly or indirectly in the so-called "giant component" of the network. Furthermore, these IPCs account for 97% of all patents. Hence, even with $|\Gamma_g| \leq 10$, almost all patents in our data are connected to each other at least indirectly in the related-IPC network.

6 Validation of patent importance

Before presenting our main findings, we validate our measure of patent importance by showing how it correlates with other economic outcomes. First, we investigate how patent importance correlates with two other commonly-used measures of patent quality that are used in the literature: the number of forward citations received by a patent and the grant status of a patent. Second, we investigate which of these measures of patent quality – importance, forward

[10] We also find a positive correlation of 0.48 between $S_{sec}^{gg'}$ and $S_{fcit}^{gg'}$, which is the share of patents with main IPC g' that get at least one citation from a patent with main IPC g.

[11] If there are less than ten secondary IPCs g' for which $S_{sec}^{gg'} > 0$, then $|\Gamma_g| < 10$.

citations, and grant status – best predict TFP and output at the firm level. One can view the first exercise as a form of validation internal to the context of patents and the second as a form of external validation. Finally, we briefly discuss the difference between patent similarity (as presented in section 3.2) and patent importance.

6.1 Internal validation: correlations with forward citations and grant status

For the internal validation exercise, we first demean patent importance, number of forward citations received, and patent grant status by IPC-year effects.[12] This effectively controls for IPC-specific time trends that might otherwise obfuscate any relationship between the patent outcomes. For instance, patents filed in earlier years may have more forward citations than patents filed in later years simply because the former have had a longer time to get cited. Similarly, the average patent grant rate may vary over time due to changes in the patent examination system that are unrelated to patent importance or patent citation behavior. Next, we assign patents into 50 percentile groups based on their demeaned importance score (so that each group accounts for 2% of all patents). We then regress the demeaned number of forward citations and grant status on dummy indicators for a patent's importance percentile group.

Figures 6.1 and 6.2 show plots of the estimated coefficients for the demeaned number of forward citations and grant status, respectively, where the error bars denote 95% confidence intervals. In Figure 6.1, we observe a strikingly monotonic relationship between forward citations and a patent's importance percentile group: patents that have higher importance within an IPC-year tend to receive more forward citations than other patents within the same IPC-year. For instance, patents at the 90^{th} percentile of the within-IPC-year distribution of patent importance (percentile group 45 on the x-axis of Figure 6.1) have, on average, 0.11 more citations than patents at the 10^{th} percentile of the distribution (percentile group 5). While this difference might seem small, the average number of forward citations per patent in our data is 0.44 with a standard deviation of 1.21, so a difference of 0.11 is sizable.

In Figure 6.2, we also observe a positive relationship between the grant status of a patent and its importance percentile group: patents that have higher importance within an IPC-year are more likely to be granted than other patents within the same IPC-year. Here, however, we observe an interesting pattern that is different from the case of forward citations – once a patent is "important enough", being more important is not associated with a higher grant rate. In particular, there appears to be a threshold around the median of the distribution (percentile group 25 on the x-axis of Figure 6.2), where patents with importance above the threshold have around a 2-3 percentage point higher grant rate than patents with importance below the threshold, but above the threshold, more important patents do not have a significantly different

[12] We exclude self-citations, where a self-citation is defined as one in which there is at least one common applicant name between the cited and citing patent.

Figure 6.1: Correlation between patent importance and forward citations

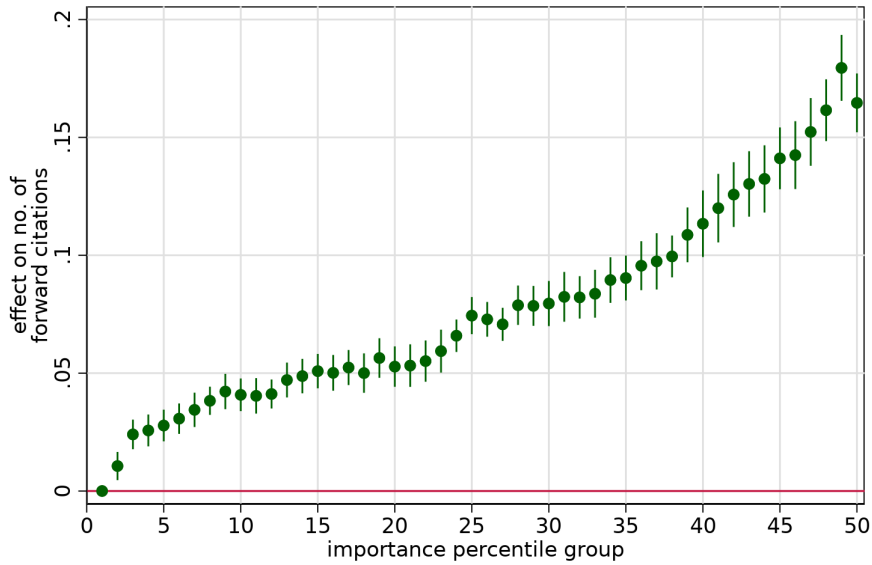
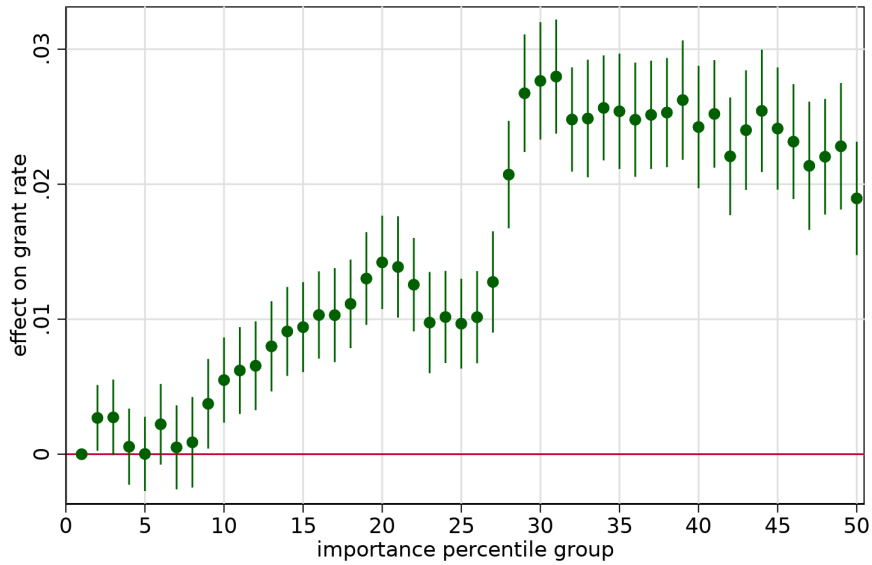


Figure 6.2: Correlation between patent importance and grant status



grant rate than less important patents. This is perhaps reflective of the patent examination process: to be granted, a patent simply has to be "innovative enough" to clear the examination bar and being more innovative beyond this does not lead to a higher probability of being granted.

6.2 External validation: predicting firm TFP and output

Figures 6.1 and 6.2 establish that patent importance, forward citations, and grant status are positively correlated within an IPC-year. However, without a reference point external to the patent data, it is impossible to assess which of these measures is a "better" indicator of patent quality, if any. To investigate, we merge the patent data with data on above-scale firms from the Chinese National Bureau of Statistics (NBS) for the years 1998-2007. In this dataset, we have 1,784,156 firm-years and 438,707 firms with information on TFP and output. Of these, 52,033 firm-years and 16,183 firms have patents. The firms in these data account for around 25% of total output in the sample.

For each patent, we first determine whether it is in the top p% of the importance distribution within its application year (a "top-p% important patent") and in the top p% of the forward citation distribution within its application year (a "top-p% cited patent"), for a given threshold p. For each firm f and year t in the merged data, we then measure the following: (i) the stock of top-p% important patents applied for by f before and up to t, $N_{ft}^{topimp,p}$; (ii) the stock of top-p% cited patents applied for by f before and up to t, $N_{ft}^{topcit,p}$; and (iii) the stock of active patents owned by f at t, N_{ft}^{active}, i.e., patents applied for by f and granted before and up to t that have not been terminated as of t.[13] Appendix Table A.I provides some summary statistics of these patent counts by year.

We then estimate the following regression via OLS:

$$y_{ft} = \beta^{topimp,p} N_{ft}^{topimp,p} + \beta^{topcit,p} N_{ft}^{topcit,p} + \beta^{active} N_{ft}^{active} + \gamma \delta_{ft} + \alpha_{h(f)t} + \epsilon_{ft} \tag{6.1}$$

where y_{ft} is a firm-year outcome of interest (e.g., log TFP or output), δ_{ft} is a dummy indicator equal to one if firm f has applied for at least one patent before and up to t, $\alpha_{h(f)t}$ is an industry-year fixed effect (where $h(f)$ denotes the 4-digit industry code of firm f), and ϵ_{ft} is a residual. We will primarily be interested in comparing the regression coefficients $\beta^{topimp,p}$, $\beta^{topcit,p}$, and β^{active}, since this will be informative about which characteristics of a firm's patents best predict firm performance: having more important patents, having more highly-cited patents, or having more active patents.

Table 6.1 report our estimated regression coefficients from equation (6.1) with t-statistics shown in parentheses, where the dependent variable y_{ft} is treated as log TFP in panel (a) and as log output in panel (b). Each column in each of the tables reports results for a different value

[13] Patents may terminate because a firm fails to pay an annual renewal fee, for example.

of the top-patent threshold, p. Note that we also scale the dependent variable by a factor of 100 so that the magnitudes of the regression coefficients can be interpreted as percentage points (p.p.). We highlight four key observations.

First, the estimated coefficient on the stock of a firm's top-p% important patents, $\beta^{topimp,p}$, is always positive and significant at the 99% confidence level. For instance, each additional patent in the top 2% of the within-year importance distribution is associated with an increase in TFP of 0.40 p.p. (row i, column $p = 2$ of panel (a)) and an increase in output of 5.42 p.p. (row i, column $p = 2$ of panel (b)). In other words, firms that own important patents are likely to be larger and more productive. Second, our estimate of $\beta^{topimp,p}$ is monotonically decreasing in the threshold percentile p. For example, each additional patent in the top 2%, 5%, 10%, 25%, and 50% of the within-year importance distribution is associated with an increase in TFP of 0.40, 0.31, 0.17, 0.05, and 0.03 p.p., respectively (row i of panel (a)). In other words, more important patents are more strongly correlated with firm TFP and output. These first two observations provide external validation that our measure of importance is capturing a characteristic of patents that is positively associated with measures of firm performance.

Third, in every specification, our estimate of $\beta^{topimp,p}$ is greater than our estimates of the coefficients on the other patent stock regressors, $\beta^{topcit,p}$ and β^{active}. For example, comparing rows i-iii of panel (a), we see that the estimated value of $\beta^{topcit,p}$ is around one-third the value of $\beta^{topimp,p}$, while the estimated value of β^{active} is negative. Similarly, comparing rows i-iii of panel (b), we observe that the estimates of $\beta^{topcit,p}$ and β^{active} are at least an order of magnitude smaller than our estimate of $\beta^{topimp,p}$. Finally, our estimate of $\beta^{topimp,p}$ decays much more slowly with the percentile threshold p compared with our estimate of $\beta^{topcit,p}$. For example, in panel (b), $\beta^{topimp,p}$ is positive and significant even for the $p = 50$ threshold, whereas $\beta^{topcit,p}$ is positive and significant only for the specifications where $p \leq 10$. In sum, our measure of patent importance is not only positively associated with firm performance, but it is also a better predictor of firm TFP and output than traditional measures of patent quality based on forward citations and the legal status of a patent.

In Appendix B, we provide results for three alternative versions of the regressions discussed above. In the first alternative, when constructing patent stocks in year t, we include only patents applied for between $t-4$ and t. In other words, we consider only "new" patents applied for by a firm. Results for this version are provided in Appendix Table A.II. In the second alternative, we add firm fixed effects to the regressions. Results for this version are provided in Appendix Table A.III. In both of these versions, we find that the four observations highlighted above continue to hold in almost all cases (the sole exception is that the coefficients on $N_{ft}^{topimp,p}$ are no longer statistically significant at the 99% confidence level when the outcome is log TFP and firm fixed effects are included in the regression).

Finally, in the third alternative, we estimate the same baseline specification as in equation (6.1) but add as a regressor the stock of a firm's top-p% important patents where importance

Table 6.1: Regressions of firm TFP on patent stocks

(a) outcome: log TFP (×100); fixed effects: industry-year year					
$p =$	2	5	10	25	50
i. stock of top-p% important patents, $\beta^{topimp,p}$	0.40	0.31	0.17	0.05	0.03
	(3.87)	(5.18)	(5.39)	(4.68)	(4.16)
ii. stock of top-p% cited patents, $\beta^{topcit,p}$	0.14	0.08	0.05	0.02	-0.00
	(2.88)	(3.17)	(3.22)	(2.21)	(-0.24)
iii. stock of active patents, β^{active}	-0.03	-0.07	-0.09	-0.09	-0.09
	(-1.48)	(-2.91)	(-3.55)	(-3.46)	(-3.25)
iv. has patents, γ	3.66	3.63	3.63	3.66	3.68
	(22.33)	(22.14)	(22.15)	(22.35)	(22.50)
observations (m)	1.78	1.78	1.78	1.78	1.78
R^2	0.73	0.73	0.73	0.73	0.73
adjusted R^2	0.73	0.73	0.73	0.73	0.73
(b) outcome: log output (×100); fixed effects: industry-year					
$p =$	2	5	10	25	50
i. stock of top-p% important patents, $\beta^{topimp,p}$	5.42	5.11	2.69	0.76	0.29
	(15.02)	(23.95)	(24.27)	(18.72)	(11.18)
ii. stock of top-p% cited patents, $\beta^{topcit,p}$	0.59	0.47	0.34	-0.01	-0.13
	(3.43)	(5.15)	(5.79)	(-0.19)	(-4.14)
iii. stock of active patents, β^{active}	0.35	-0.28	-0.62	-0.32	-0.01
	(4.90)	(-3.36)	(-6.91)	(-3.32)	(-0.10)
iv. has patents, γ	110.64	110.09	110.17	110.64	111.01
	(191.48)	(190.30)	(190.53)	(191.57)	(192.37)
observations (m)	1.78	1.78	1.78	1.78	1.78
R^2	0.15	0.15	0.15	0.15	0.15
adjusted R^2	0.15	0.15	0.15	0.15	0.15

Notes: This table reports the results of estimating equation (6.1) by OLS, where the dependent variable is log TFP in panel (a) and log output in panel (b). t-statistics are shown in parentheses.

is based on the naive measure defined in equation (4.1). This allows us to assess whether our model-based measure of importance is superior to the naive measure in terms of predicting firm TFP and output. Results for this version are provided in Appendix Table A.IV. We find that in almost all cases, the estimated coefficient on our baseline measure of top-p% important patents is larger than the corresponding coefficient based on the naive measure of importance (the sole exception is for the specification where the outcome is log TFP and top patents are defined at the 2% threshold). Furthermore, in many specifications, the estimated coefficient on the naive measure of important patents is negative or statistically insignificant. The advantages that our baseline measure of patent importance has over the naive measure are thus meaningful for predicting firm performance.

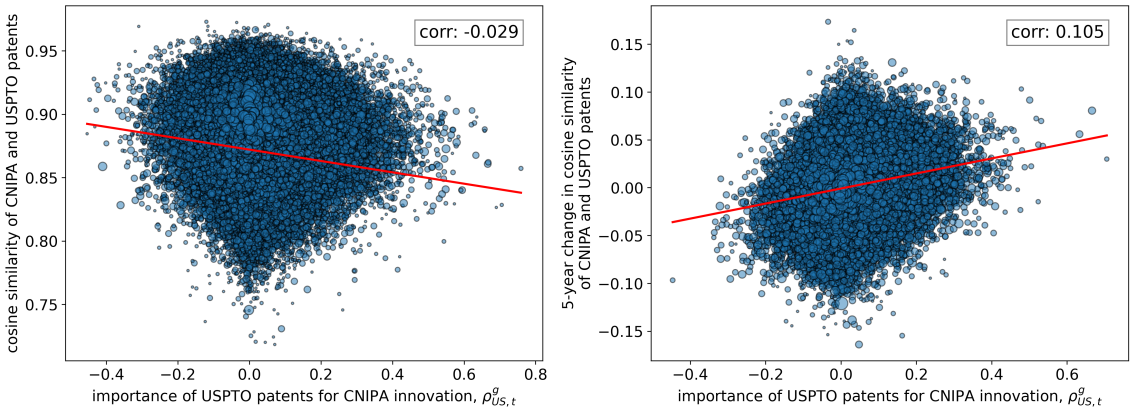
6.3 Patent importance versus patent similarity

We conclude this section by making the point that patent similarity and patent importance are two related but fundamentally different concepts. If one observes patents in China becoming more similar to patents in the US, one may be tempted to infer that US patents are somehow becoming more influential in shaping Chinese innovation. In the left panel of Figure 6.3, we show the relationship between the importance of USPTO patents for innovation in China within an IPC-year, $\rho^g_{US,t}$, and the average abstract cosine similarity between CNIPA and USPTO patents in the same IPC-year. In the right panel, we show the relationship between $\rho^g_{US,t}$ and the change in average abstract cosine similarity from $t-2$ to $t+2$. In each figure, the size of a marker reflects the number of patents pairs being compared (i.e., the number of CNIPA patents times the number of USPTO patents) and the best-fit line is from a regression of the y-axis variable on the x-axis variable weighted by number of patent pairs. We see that patent importance is essentially unrelated to abstract similarity: being more similar to the current state of knowledge is not the same thing as being more similar to the direction of innovation. Furthermore, patent importance is only weakly positively correlated with changes in similarity over time. Hence, the measure of patent importance motivated by our model of the innovation process is a fundamentally different measure from simple similarity scores.

7 Results

We have now provided three forms of validation that text embeddings can be used to capture meaningful variation across patents and to measure patent importance: (i) patents in the same IPC section have embeddings that are more similar to each other (Figure 3.1); (ii) our measure of patent importance is positively correlated with other traditional measures of patent quality (Figures 6.1 and 6.2); and (iii) our measure of patent importance is a better predictor of firm TFP and output than traditional measures of patent quality (Table 6.1). We now present our

Figure 6.3: Patent importance versus patent similarity



main findings about how patent importance has evolved over time in China.

7.1 The distribution of patent importance

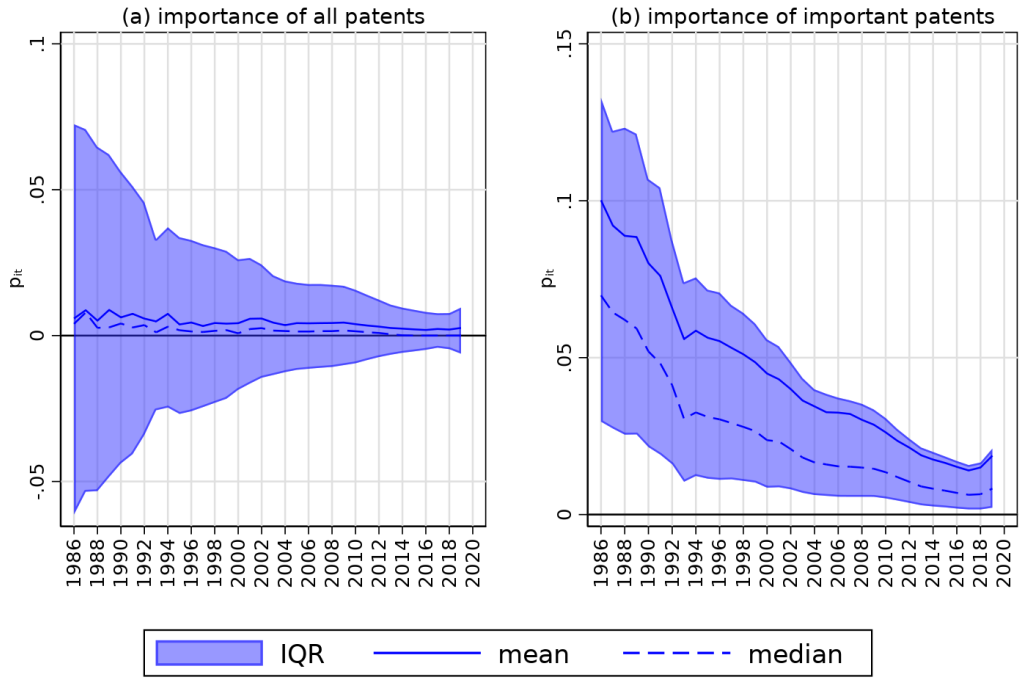
Panel (a) of Figure 7.1 shows how the distribution of our estimated patent importance measure p_{it} changes over time. As expected given the discussion in section 4.3, the average value of patent importance is consistently close to zero in every year and is not a meaningful statistic. What is meaningful, however, is that the *dispersion* of the patent importance distribution is declining over time and, closely related, that the average importance of important patents, defined as those with $p_{it} > 0$, is also declining over time as shown in panel (b) of Figure 7.1. What explains these trends? We propose two explanations.

First, the rate at which the state of knowledge in China is changing has slowed over time. To examine this formally, let us define the *innovativeness* of IPC g at time t as:

$$I_t^g = \frac{1}{2}\left[1 - c\left(\vec{\mu}_{Ft}^g, \vec{\mu}_{Bt}^g\right)\right] \tag{7.1}$$

where again $\vec{\mu}_{Ft}^g$ and $\vec{\mu}_{Bt}^g$ are the average patent embeddings (normalized to unit length) in IPC g in years $\{t+1, t+2, t+3\}$ and $\{t-2, t-1, t\}$, respectively. Note that I_t^g is larger when $\vec{\mu}_{Ft}^g$ and $\vec{\mu}_{Bt}^g$ are less similar and hence innovativeness is a measure of how much the future state of knowledge differs from the past. Note also that $I_t^g \in [0, 1]$, since $c\left(\vec{\mu}_{Ft}^g, \vec{\mu}_{Bt}^g\right) \in [-1, 1]$. Panel (a) of Figure 7.2 shows how the distribution of innovativeness changes over time at the patent level (i.e., the distribution of $I_{it} \equiv I_t^{g(i)}$ where $g\left(i\right)$ is the main IPC of patent i). Evidently, the innovativeness of Chinese patenting has declined steadily over time, indicating that the state of knowledge is changing at a slower pace. Recall that the patent importance scores are the weights assigned to the patent embeddings in order to "explain" the direction of innovation. Hence, when innovativeness is low, there is less that needs to be explained in the first place and

Figure 7.1: The distribution of patent importance over time

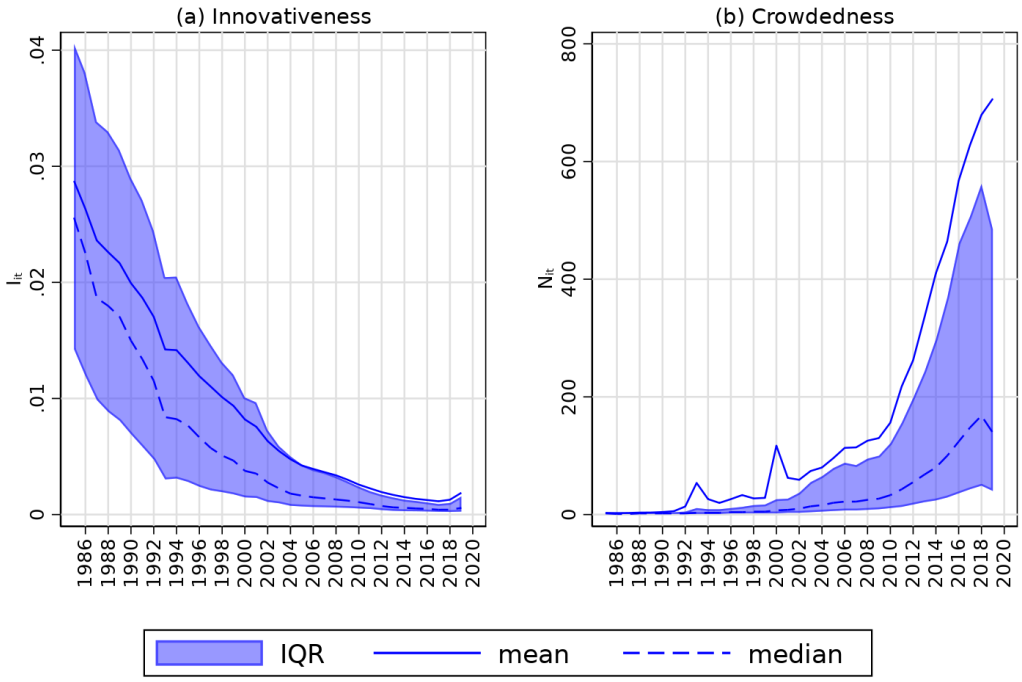


the magnitudes of the patent importance scores tend to be small.

The second explanation for the decline in the dispersion of patent importance and the importance of important patents is that Chinese patenting has become much more *crowded* over time. More formally, for each patent i in IPC $g(i)$ applied for at time t, we measure the total number of patents applied for in the same IPC-year, $N_{it} \equiv N_t^{g(i)}$. In panel (b) of Figure 7.2, we plot the distribution of N_{it} over time. Clearly, the crowdedness of Chinese patenting has increased substantially. For example, the average patent in 2005 has around 100 other patents in the same IPC-year, whereas this increases by about seven-fold by 2019. The crowdedness of innovation matters for the distribution of patent importance because when there are many candidate patents that can explain the direction of innovation, each individual patent is less likely to matter a lot.

To verify that the innovativeness and crowdedness of patenting are related to patent importance as discussed above, we consider two regressions. First, at the patent-year level, we regress importance (p_{it}) on innovativeness (I_{it}) and crowdedness (measured as $\log N_{it}$) for the set of important patents ($p_{it} > 0$). Columns (1) and (2) of Table 7.1 report the estimated standardized beta coefficients (t-statistics shown in parentheses) with and without the inclusion of IPC fixed effects, respectively. Second, at the IPC-year level, we compute the average importance of important patents, $\bar{p}_t^{g+} \equiv \frac{1}{|\Omega_t^{g+}|} \sum_{i \in \Omega_t^{g+}} p_{it}$, where $\Omega_t^{g+} \equiv \{i \in \Omega_t^g : p_{it} > 0\}$. We then regress

Figure 7.2: Innovativeness and crowdedness of Chinese patenting



$\bar{p}_t^{g+}$ on innovativeness (I_t^g) and crowdedness ($\log N_t^g$) at the IPC-year level. Columns (3) and (4) of Table 7.1 report the estimated standardized beta coefficients, again with and without the inclusion of IPC fixed effects, respectively.

Clearly, there is a close empirical relationship between patent importance and the innovativeness and crowdedness measures. For example, at the patent-year level within an IPC, an increase in innovativeness by one standard deviation is associated with an increase in patent importance by 43% of a standard deviation (row i, column (1) of Table 7.1). Similarly, an increase in crowdedness by one standard deviation is associated with a decrease in patent importance by 22% of a standard deviation (row ii, column (1) of Table 7.1). Hence, our finding that the average importance of important patents is declining over time is closely related to the fact that patenting in China is becoming less innovative and more crowded. At the same time, we note that there is still a great deal of heterogeneity across patents in terms of importance despite these clear aggregate trends. For instance, variation in innovativeness and crowdedness explain only 9% of the within-IPC variation in patent importance (last row of column (1) in Table 7.1).

It is also important to note that the decline in innovativeness and increase in crowdedness of patenting is not unique to China. Figure A.II in Appendix C shows that similar patterns are observed since 1976 for granted USPTO utility patents. As a point of comparison, the average level of innovativeness for China in 2010 (around 0.02) is similar to the level of innovativeness

Table 7.1: Regressions of patent importance on innovativeness and crowdedness

	(1)	(2)	(3)	(4)
observation =	patent-year	patent-year	IPC-year	IPC-year
i. innovativeness	0.43	0.51	0.52	0.54
	(557.97)	(874.56)	(183.27)	(218.56)
ii. crowdedness	-0.22	-0.25	-0.16	-0.21
	(-245.99)	(-597.01)	(-90.88)	(-138.60)
IPC fixed effects	yes	no	yes	no
observations (m)	6.01	6.01	0.52	0.52
R^2	0.36	0.28	0.39	0.24
adjusted R^2	0.35	0.28	0.32	0.24
within R^2	0.09	-	0.14	-

Notes: Columns (1) and (2) report the results of regressing patent importance (p_{it}) on innovativeness (I_{it}) and crowdedness ($\log N_{it}$) at the patent-year level for the sample of important patents ($p_{it} > 0$). Columns (3) and (4) report the results of regressing the average importance of important patents ($\bar{p}_t^{g+}$) on innovativeness (I_t^g) and crowdedness ($\log N_t^g$) at the IPC-year level. Coefficients are reported as standardized beta coefficients and t-statistics are shown in parentheses.

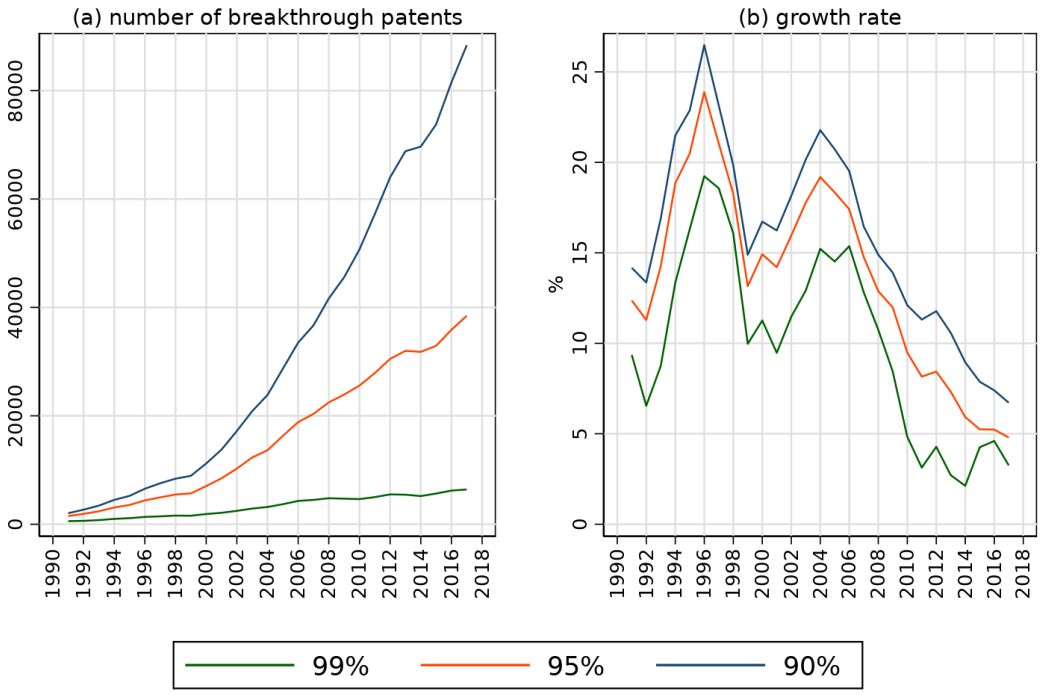
reached in the US for granted utility patents in the mid 1990s. Similarly, the average level of crowdedness in China (around 200 patents) is similar to the average level of crowdedness in the US in the early 2000s, although one should bear in mind that we are counting all patent applications for China but only granted patents for the US.

Furthermore, note that the decline in the average importance of important patents occurs in parallel with rapid growth in the number of patents overall (Figure 1.1). Hence, the stock of highly important patents may well be increasing. To investigate, we define a *breakthrough* patent as one that has an importance score above the k^{th} percentile of the importance distribution over the entire sample of patents (so that the threshold does not vary across years). Figure 7.3 shows the number of breakthrough patents in each year for $k \in \{99, 95, 90\}$ and the associated annual growth rates. We indeed find that the number of breakthrough patents is increasing over time, with higher rates of growth in the early 2000s of around 15-20% per year and lower but still positive rates of growth in later years.

7.2 The importance of Chinese versus foreign knowledge

We next turn our attention to investigating which sources of knowledge are the most important for shaping the direction of innovation in China. The first margin that we consider is the

Figure 7.3: Number and growth rate of breakthrough patents



relative importance of knowledge embodied in patents filed in China versus outside of China. In particular, we compare here the estimated importance of Chinese patents p^g_{it} with the estimated importance of US patents $p^g_{US,t}$.

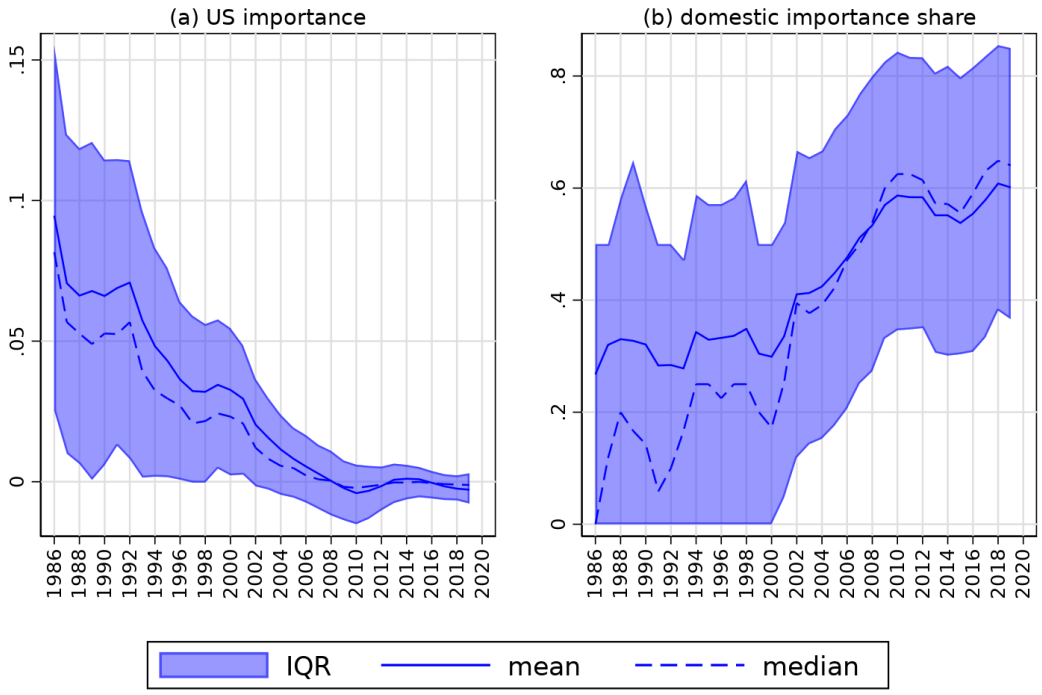
Panel (a) of Figure 7.4 shows how the distribution of US importance, $p^g_{US,t}$ varies over time. Much like the distribution of importance for important Chinese patents, we observe that the importance of US knowledge is steadily declining. This decline may be explained in part by the same forces as discussed above – i.e., less innovative and more crowded patenting in China – but some of the decline may also reflect changes in the *relative* importance of knowledge embedded in Chinese versus US patents.

To investigate the latter, we measure for each IPC g and year t the share of Chinese patents with an importance score greater than the importance of US patents:

$$\mathcal{D}^g_t \equiv \frac{1}{N^g_t} \sum_{i \in \Omega^g_t} \mathbb{1}_{\left[p^g_{it} > p^g_{US,t}\right]} \tag{7.2}$$

We henceforth refer to this measure as the *domestic importance share*. Panel (b) of Figure 7.4 plots the distribution of the domestic importance share over time at the patent level (i.e., the distribution of $\mathcal{D}_{it} \equiv \mathcal{D}^{g(i)}_t$). We observe that the average domestic importance share is roughly constant from 1985 to 2000 at around 30%. Throughout the 2000s, however, there is a sustained

Figure 7.4: US importance and the domestic importance share



increase in the measure, so that by 2010, the average share is just under 60%. From 2010 onward, the domestic importance share levels off and again remains roughly constant at around 60%. The 2000s were thus a key period during which innovation in China became increasingly directed by knowledge within China as opposed to knowledge outside.

Figure 7.4 also shows that there is substantial heterogeneity across IPCs in the domestic importance share. To examine how the relative importance of Chinese versus US knowledge differs across technology classes, Table 7.2 reports how the domestic importance share has changed over time for 23 technology groups defined by IPC class (the first three characters of the IPC code). Column (1) reports the share of all patents accounted for by each IPC group G, S^G. Columns (2) and (3) report the domestic importance share for each IPC group G pooling all patents across 2000-2009 ($\mathcal{D}^G_{00-09}$) and 2010-2019 ($\mathcal{D}^G_{10-19}$), respectively. Column (4) reports the log change in the domestic importance share across these two time periods, $\Delta \log \mathcal{D} \equiv \log \mathcal{D}^G_{10-19} - \log \mathcal{D}^G_{00-09}$, and column (5) reports the log change in the domestic importance share for each IPC group G relative to the same log change across all patents.

We highlight two observations. First, there is an increase in the domestic importance share for all 23 IPC groups from 2000-2009 to 2010-2019 (all values in column (4) are positive). The relative shift in the direction of innovation toward Chinese versus foreign knowledge is hence not limited to only a few technology classes but rather is a widespread phenomenon in China.

Second, even though there is an increase in the domestic importance share for all IPC groups, there is also substantial heterogeneity across groups in the rate of this increase. The domestic importance share grows the fastest in technologies such as Paper (D21), Health; Life-saving (A61-A62), Metallurgy (C21-C30), and Chemistry (C01-C14), whereas it grows the slowest in Personal or Domestic Articles (A41-A47), Engines or Pumps (F01-F04), Printing (B41-B4), Transporting (B60-B68), and Engineering in General (F15-F17).

We also consider whether Chinese government policy may have played a role in shaping which technologies experienced faster growth in the domestic importance share. We focus on a key government policy, the Medium- and Long-Term Program for Science and Technology Development (MLP), which was officially launched by the State Council in February 2006. This program specified China's science and technology goals for the period 2006-2020. Among other things, the MLP identified a set of 62 "priority technologies" that were considered important for both economic development and national security and that should hence receive more policy support. For example, in the energy sector, an MLP priority was "Low-cost, large-scale development and utilization of renewable energy resources", while in the transportation sector, a priority was "High-speed rail transport systems".

To link the MLP priority technologies to our patent data, we proceed as follows. First, we obtain text descriptions (several paragraphs) of each MLP priority from an outline of the MLP published by the State Council. We then generate the text embedding of the description for each MLP priorty (call this the *priority embedding*). Next, we compute the average patent abstract embedding in each IPC from 2000-2010 (call this the *IPC embedding*). For each MLP priority x IPC pair, we then compute the cosine similarity between *priority embedding* and *IPC embedding*. This gives us a measure of how similar patents in each IPC are compared with the MLP priorities. To classify an IPC as being an MLP priority or not, we compute the average cosine similarity across the top three most similar priority technologies for each IPC (call this score *max priority*). This is a measure of how similar each IPC is to its three "closest" priority technologies. Finally, we define an IPC to be an "MLP priority IPC" if *max priority* is above the 90^{th} percentile of the distribution across all patents. The results we report below are also robust to defining the threshold at the 75^{th} and 95^{th} percentiles.

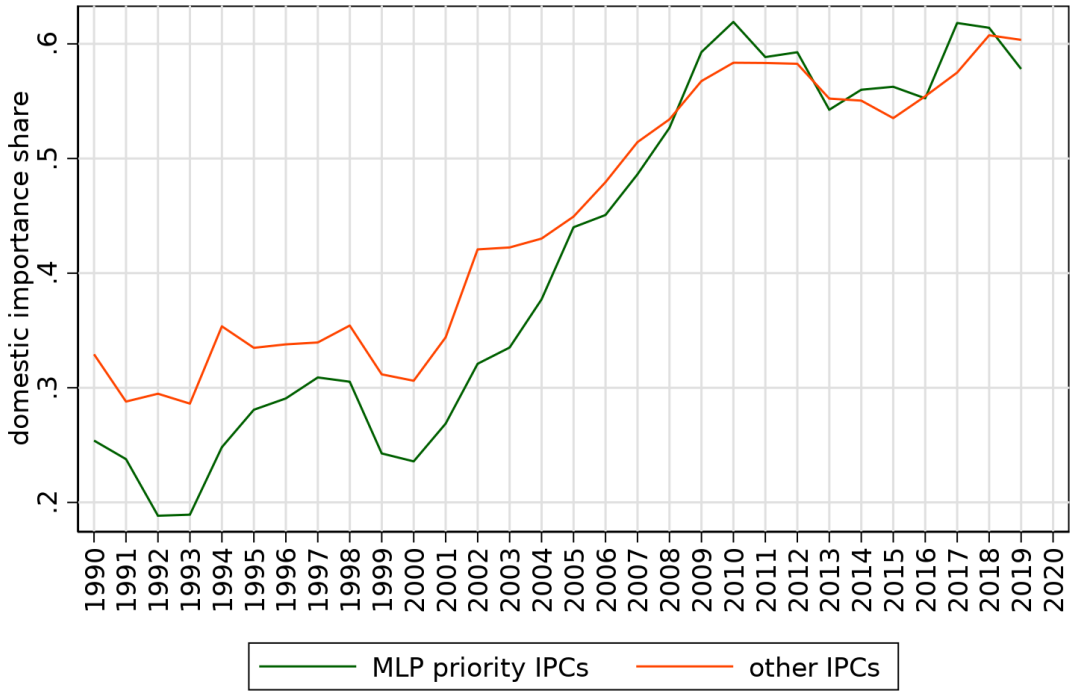
Figure 7.5 shows how the domestic importance share changes over time for MLP priority IPCs and non-priority IPCs. We observe that MLP priority IPCs initially have a lower domestic importance share than the other IPCs. However, the gap narrows starting in the early 2000s, such that by 2008, the MLP priority IPCs have caught up with the other IPCs in terms of the domestic importance share. After 2008, the MLP priority technologies have a greater domestic importance share than the non-priority IPCs in most years. While this is not causal evidence, it clearly establishes that the technologies targeted by the MLP were also those for which the direction of innovation was trending more toward Chinese patents than US patents.

Table 7.2: Domestic importance shares by IPC group

		(1)	(2)	(3)	(4)	(5)
IPC group	Classes	S^G	$\mathcal{D}^G_{00-09}$	$\mathcal{D}^G_{10-19}$	$\Delta \log \mathcal{D}^G$	$\Delta \log \hat{\mathcal{D}}^G$
Paper	D21	0.2	42.8	57.1	28.8	11.3
Health; Life-saving	A61-A62	7.1	46.3	60.8	27.2	9.8
Metallurgy	C21-C30	2.1	47.3	60.2	24.1	6.7
Chemistry	C01-C14	13.3	48.5	60.6	22.4	4.9
Weapons; Blasting	F41-F42	0.1	41.8	51.0	19.9	2.5
Microstructural Technology	B81-B82	0.1	48.0	58.4	19.5	2.0
Amusement	A63	0.5	48.0	58.2	19.2	1.7
Nuclear Physics	G21	0.1	41.0	49.3	18.5	1.0
Foodstuffs; Tobacco	A21-A24	3.4	49.7	59.7	18.2	0.8
Electric Elements and Techniques	H01-H05	17.3	47.0	56.0	17.7	0.2
Earth or Rock Drilling; Mining	E21	0.7	50.2	59.0	16.2	-1.3
Separating; Mixing	B01-B09	4.0	51.7	60.0	14.9	-2.5
Non-nuclear Physics	G01-G16	20.2	47.8	55.3	14.7	-2.8
Lighting; Heating	F21-F28	3.3	48.2	55.6	14.4	-3.0
Shaping	B21-B33	6.9	50.7	58.4	14.2	-3.2
Agriculture	A01	2.8	49.4	56.7	13.9	-3.5
Textiles or Flexible Materials	D01-D07	1.5	48.2	54.7	12.6	-4.8
Building	E01-E06	3.1	50.8	57.6	12.5	-4.9
Engineering in General	F15-F17	2.3	52.1	58.7	12.0	-5.4
Transporting	B60-B68	6,0	50.1	56.0	11.0	-6.5
Printing	B41-B44	0.7	49.8	55.0	10.0	-7.4
Engines or Pumps	F01-F04	1.9	52.0	56.7	8.5	-8.9
Personal or Domestic Articles	A41-A47	2.5	50.9	55.0	7.7	-9.7

Notes: Column (1) reports the share of patents across all years accounted for by each IPC group G, S^G. Columns (2) and (3) report the domestic importance share pooling all patents across 2000-2009 and 2010-2019, respectively. Column (4) reports the log difference between columns (3) and (2). Column (5) subtracts from column (4) the log change in the domestic importance share from 2000-2009 to 2010-2019 for all patents. Values in columns (1)-(3) are reported as percentages. Rows are sorted in descending order of the values in column (5).

Figure 7.5: Domestic importance shares for MLP priority and non-priority technologies



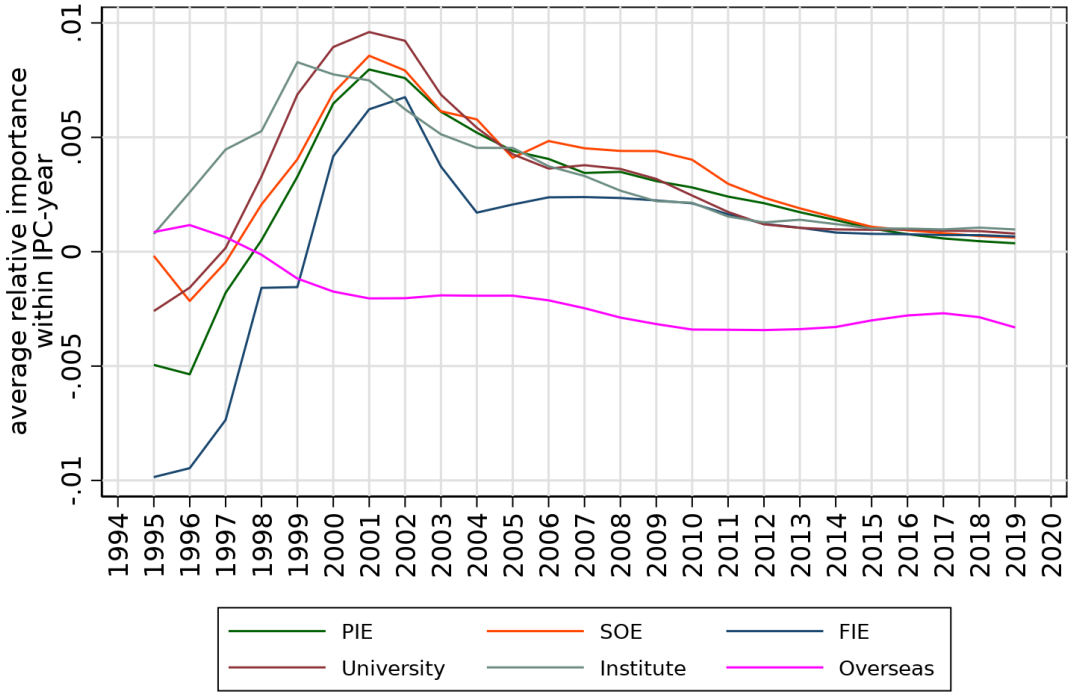
7.3 Relative importance by patentee types within China

We now examine which sources of knowledge within China have been more important for driving the direction of innovation in China. We base our analysis on the patentee types described in section 2.1. Recall that there are important changes over time in the shares of patents accounted for by different patentee types, with rapid growth in the PIE share together with a marked decline in the share for overseas patentees.

To compare the importance of patents across patentee types, we first compute the importance of each patent parsed of IPC-year effects. We then average this measure over all patents applied for by a given patentee type in a given year. This measures the *relative* importance of patents by each patentee type within an IPC-year. Figure 7.6 shows how this relative measure changes over time. For visual clarity, we omit the "Individual" and "Other Domestic" patentee types and plot five-year rolling averages of the relative importance scores.

We highlight two key observations. First, before 2000, patents filed by Chinese universities and research institutes tended to be the most important. By 2005, however, SOE patents assume the top position in terms of relative importance. In almost all years, SOE patents are relatively more important than private sector patents, while PIE patents are also typically more important than FIE patents. Second, in the mid-1990s, overseas patents had slightly higher than

Figure 7.6: Relative patent importance by patentee type



average importance. However, the relative importance of these patents declines steadily, so that from 2000 onward, overseas patents are less important than patents applied for by enterprises, universities, and research institutes in China. A potential explanation for this is that overseas patentees perhaps mainly file patents in China to protect the intellectual property of goods that are being exported to the Chinese market and hence these patents are not important for innovation in China, although one would need patentee-level export data to investigate this thoroughly. In Appendix C, Figure A.III, we show that overseas patents at the USPTO have also been less important for US innovation than domestic patents since the mid-1980s, although the gap in importance is smaller than it is for China.[14] Patents applied for at the USPTO by Chinese applicants are also less important for US innovation than patents owned by non-Chinese overseas applicants, although there was an increase in the relative importance of Chinese patents in the 2000s.

In Appendix D, we show that patent importance can be approximately decomposed as:

$$p_{it}^{g} \approx \sum_{s} S_{Ft}^{gs} p_{it}^{gs} \tag{7.3}$$

[14]These estimates are generated by estimating the same model specified in equation (4.9) for granted USPTO utility patents, replacing the average USPTO embedding on the right-hand side with the average CNIPA embedding.

where p_{it}^{gs} is the importance of patent i for patenting by patentee type s and S_{Ft}^{s} is the share of future patents attributed to patentee type s in IPC g.[15] This decomposition allows us to further examine how patent importance varies by patentee type pairs. In Figure 7.7, each matrix shows, for a particular five-year window, the average importance of patents by patentee type s (rows) for patenting by patentee type s' (columns), defined as $\tilde{p}_t^{ss'} = \frac{1}{|\Omega_t^s|}\sum_g \sum_{i\in\Omega_t^{gs}} p_{it}^{gs'}$ where Ω_t^{gs} is the set of patents in IPC g applied for by patentee type s and $\Omega_t^s \equiv \cup_g \Omega_t^{gs}$. The column on the right of each matrix shows the average importance of patents for each patentee type, $\frac{1}{|\Omega_t^s|}\sum_g \sum_{i\in\Omega_t^{gs}} p_{it}^{g}$. For visual clarity, all importance scores are multiplied by 1000.

As expected, we find that patents by a given patentee type are always the most important for themselves (the diagonals of each matrix are almost always larger than the off-diagonals). In particular, we observe that overseas patents are always important for themselves (and in fact are increasingly more so from 2000-2019) but are almost never important for patenting by domestic patentees. This sheds light on why the average relative importance of overseas patents falls overall: these patents are only important for themselves and account for a smaller share of total patents over time. We also observe that among domestic patentees, the average importance of overseas patents is highest for FIEs, which again is to be expected. Patents by PIEs and SOEs are also important for each other, but the importance of SOE patenting for PIE patenting tends to be higher than the other way around. FIE patents tend to be important for PIE patenting but less so for SOE patenting. University and research institute patents are important for each other as well as for PIE and SOE patenting, but tend not to be important for FIE patenting.

Finally, we note that measures of average patent importance may mask differences across patentee types among the most important patents. Hence, we now examine the representation of each patentee type among the breakthrough patents discussed at the end of section 7.1. For patentee type s in year t, we measure this representation using two statistics:

$$r_{st} = \sum_g S_{st}^g \left(S_{st}^{g*} - S_t^{g*}\right), \qquad \bar{r}_{st} = S_{st}^* - S_t^* \tag{7.4}$$

where S_{st}^g is share of patents by s in IPC g, S_{st}^{g*} is the share of patents by s in IPC g that are breakthrough patents, S_t^{g*} is the share of all patents in IPC g that are breakthrough patents, S_{st}^* is the share of patents by s in all IPCs that are breakthrough patents, and S_t^* is the share of all patents in all IPCs that are breakthrough patents. Hence, r_{st} measures whether patentee type s is over- or under-represented among breakthrough patents *within* an IPC on average, which is directly comparable with the relative importance scores within IPC-years in Figure 7.6, whereas $\bar{r}_{st}$ simply measures whether patentee type s is over- or under-represented among breakthrough patents overall.

[15] This decomposition is exact if patent importance is estimated via OLS and hence is an approximation only because we estimate importance using the LASSO estimator. Nonetheless, the decomposition is almost exact: the correlation between p_{it} and the right-hand side of (7.3) is 0.98.

Figure 7.7: Patent importance by patentee type pairs

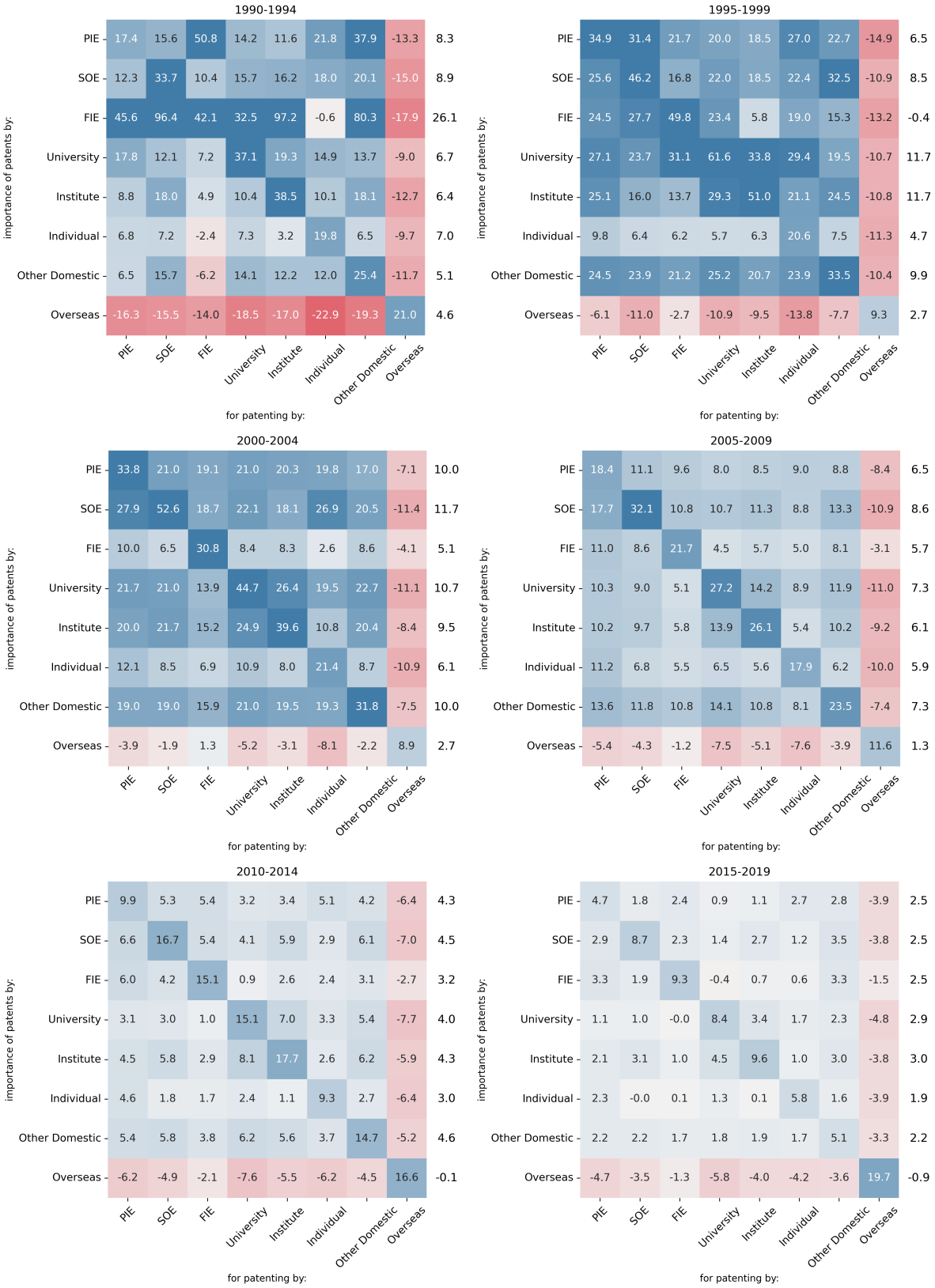
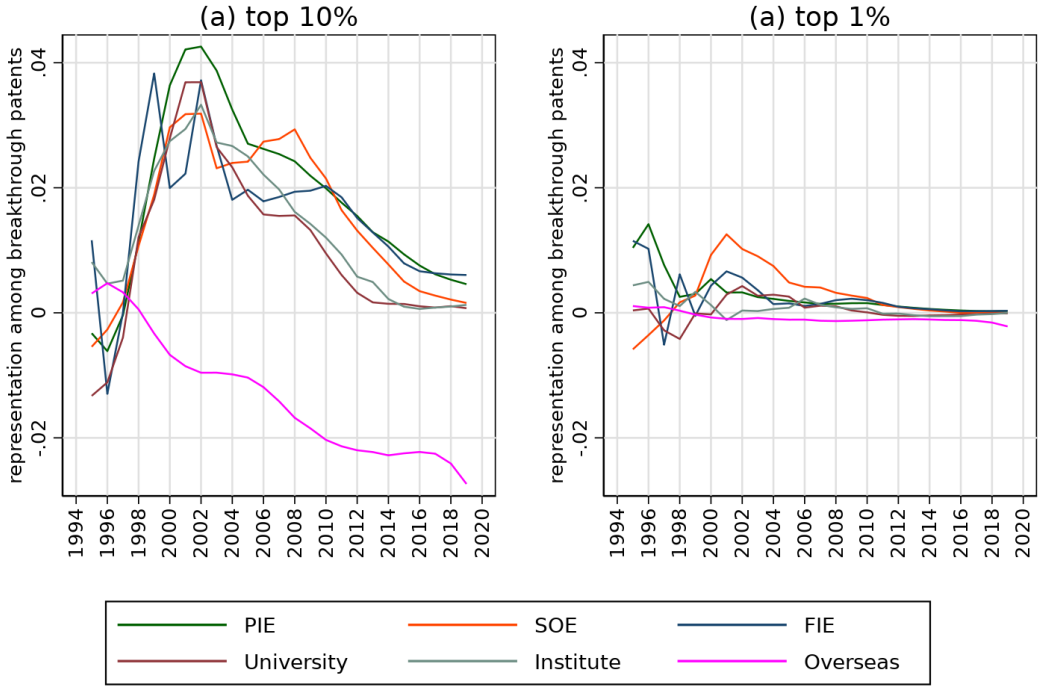
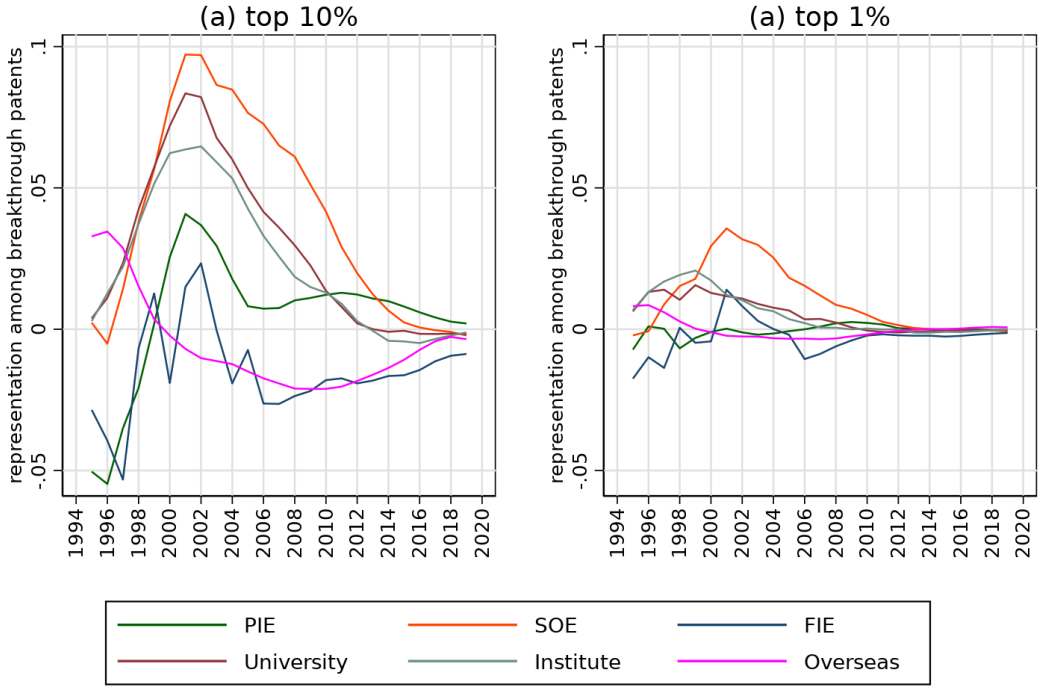


Figure 7.8: Representation among breakthrough patents within IPC by patentee type



Figures 7.8 and 7.9 plot the representation measures r_{st} and $\bar{r}_{st}$, respectively, where breakthrough patents are defined as either those in the top 10% (panel (a)) or top 1% (panel (b)) of the importance distribution across all years. Again, For visual clarity, we omit the "Individual" and "Other Domestic" patentee types and plot five-year rolling averages. Consistent with the patterns discussed above, we see that in Figure 7.8 that overseas patents are increasingly under-represented among breakthrough patents within IPCs, although differences in representation across patentee types are smaller when breakthrough patents are defined at the 1% threshold. There are also interesting differences between the patterns based on average importance and those based on representation among breakthrough patents. For instance, even though all domestic patentee types have quite similar relative importance by the end of the sample in Figure 7.6, we find that PIEs and FIEs are the most over-represented among breakthrough patents within IPCs, whereas government related entities (SOEs, universities, and research institutes) are less represented. Finally, in Figure 7.9, we see a somewhat different picture: when considering representation among breakthrough patents overall, SOE, university, and research institute patents are the most over-represented, whereas FIE and overseas patents are under-represented, reflecting compositional differences across IPCs between patentee types.

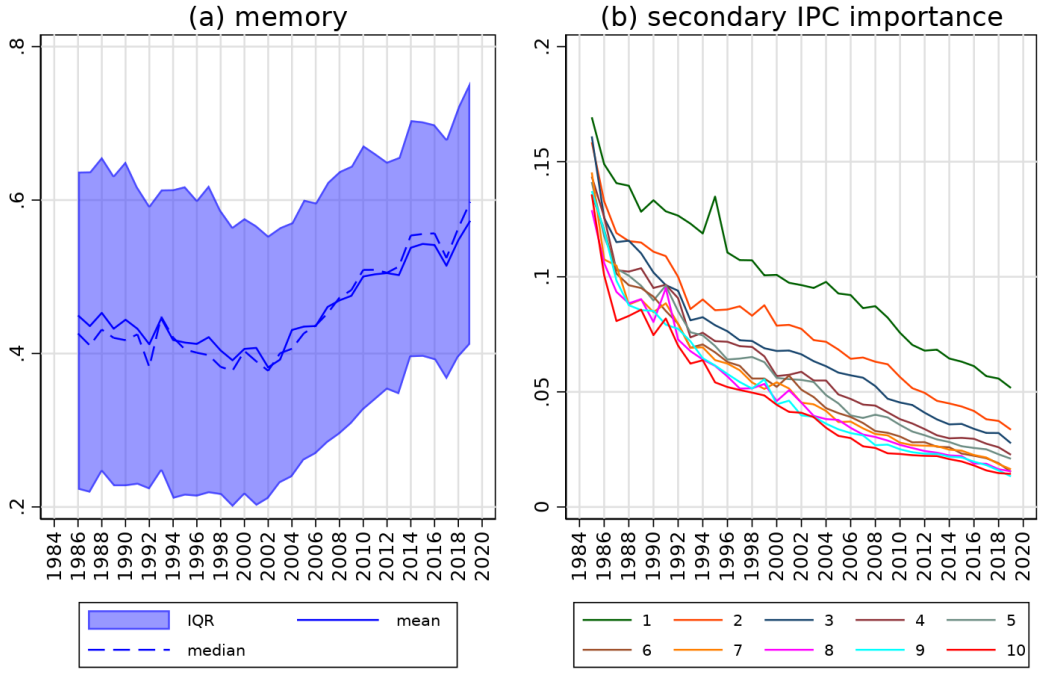
Figure 7.9: Overall representation among breakthrough patents by patentee type



7.4 Memory and the importance of secondary IPCs

Finally, we examine how the persistence of the innovation process and the importance of interactions across IPCs have evolved over time. Panel (a) of Figure 7.10 shows our estimates of the memory coefficient ρ_t^g, while panel (b) shows our estimates of secondary IPC importance $\gamma_t^{gg'}$. Note that in panel (b), secondary IPCs g' are ordered based on their rank of $S_{sec}^{gg'}$ for each main IPC g, where recall that $S_{sec}^{gg'}$ is the share of patents with main IPC g that list g' as a secondary IPC. We observe that memory is roughly constant in the initial years of our sample, but trends steadily upward beginning in the early 2000s. Furthermore, the average importance of secondary IPCs is not only lower than own-IPC importance as expected, but also trends steadily downard throughout our sample. This is suggestive that patenting in China is becoming "narrower", where knowledge generated within an IPC is more important for innovation in the IPC as opposed to knowledge generated in other related IPCs. We see similar patterns at the USPTO (see Appendix Figure A.IV), although the upward trend in memory begins about a decade earlier (in the 1990s). Lastly, note from panel (b) that the ranking of average importance across secondary IPCs is almost always perfectly correlated with the ranking of secondary IPCs by $S_{sec}^{gg'}$, i.e., secondary IPCs that are reported more frequently by patents in an IPC are estimated to have higher importance for innovation in that IPC. This once again provides validation that

Figure 7.10: Memory and importance of secondary IPCs



our importance measure is correlated with other patent characteristics in exactly the way that one would expect.

8 Conclusion

In this paper, we have leveraged frontier methods in machine learning, a new theory of innovation, and novel data on patent ownership to quantify which sources of knowledge have been important for driving the direction of innovation in China. We have developed a new measure of patent importance based on the embeddings of patent text and have shown that this measure is correlated with traditional measures of quality and outperforms these measures in terms of predicting TFP and output at the firm level.

We conclude our analysis with three key takeaways. First, patenting in China has become less innovative and more crowded over time, so that important patents have become less important on average. Second, knowledge within China has become more important than knowledge outside China. Finally, knowledge produced by Chinese entities within China (PIEs, SOEs, universities, and research institutes) has become more important than knowledge produced by foreign entities (FIEs and overseas patentees), although FIEs are still over-represented among the most important patents.

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

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.

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.

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.

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

A Identifying entity and patentee types

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

1. If *appname* contains 解放军, assign *enttype* "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, 公室, 工司, 公社, 商店, 工业}, contains 厂 and has length greater than three characters, or ends in 站, assign *enttype* "enterprise".
3. If *appname* contains 部, 厅, 局 and has length greater than three characters, assign *enttype* "government".
4. If *appname* contains {大学, 学校}, assign *enttype* "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 *enttype* "institute".
6. If *appname* contains 学院, assign *enttype* "university".
7. If *appname* contains 研究, assign *enttype* "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 *enttype* "organization".
9. If *appname* has length no greater than three characters, assign *enttype* "individual".
10. If *appname* does not satisfy any of the above, assign *enttype* "other".

Note that the keyword search is sequential in the sense that if an applicant name has already been assigned an entity 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 entity type *penttype* based on the applicant names listed in the patent application. Note that entity types at the patent level (*penttype*) and applicant level (*enttype*) can differ because some patents have multiple applicants. The assignment of *penttype* is based on the following procedure.

1. If at least one entity type is "enterprise" and no entity types are in {"university", "institute", "organization", "government", "military"}, assign *penttype* "enterprise".

2. If at least one entity type is "university" and no entity types are in {"company", "institute", "organization", "government", "military"}, assign *penttype* "university".

3. If at least one entity type is "institute" and no entity types are in {"company", "university", "organization", "government", "military"}, assign *penttype* "institute".

4. If at least one entity type is "organization" and no entity types are in {"company", "university", "institute", "government", "military"}, assign *penttype* "organization".

5. If at least one entity type is "government" and no entity types are in {"company", "university", "institute", "organization", "military"}, assign *penttype* "government".

6. If at least one entity type is "military" and no entity types are in {"company", "university", "institute", "organization", "government"}, assign *penttype* "military".

7. If at least one entity type is "individual" and no entity types are in {"company", "university", "institute", "organization", "government", "military"}, assign *penttype* "individual".

8. If no entity types are in {"company", "university", "institute", "organization", "government", "military", "individual"}, assign *penttype* "other".

9. Assign all remaining patents *penttype* "mixed".

Finally, we combine information about the location, entity type, and ownership type of each patent to define eight mutually exclusive patentee types:

Table A.I: Classification of patentee types

	location	entity type	ownership
private-invested enterprise (PIE)	domestic	enterprise	private
state-owned enterprise (SOE)	domestic	enterprise	state
foreign-invested enterprise (FIE)	domestic	enterprise	foreign
university	domestic	university	-
institute	domestic	institute	-
individual	domestic	individual	-
other domestic	domestic	other	-
overseas	overseas	-	-

B Supplementary tables for firm TFP and output regressions

This section provides supplementary information about data and results for the firm-level regressions discussed in section 6.2. Table A.I provides summary statistics about the regression sample. Table A.II provides results from estimating equation (6.1) where patent stocks at t include only patents applied for between $t-4$ and t ("new patents"). Table A.III provides results from estimating equation (6.1) where firm fixed effects are added to each regression.

Table A.I: Summary statistics for the firm regression sample

	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
no. of firms	110,982	112,421	116,779	128,742	141,451	159,778	235,210	231,200	257,313	290,280
no. of firms with:										
patents	1,605	1,789	2,089	2,444	3,308	4,338	6,303	7,492	9,790	12,875
active patents	188	198	260	381	495	738	1,391	1,961	2,692	3,562
top-1% important patents	29	30	36	52	77	127	197	278	420	612
top-2% important patents	51	52	69	102	163	247	415	554	795	1,149
top-5% important patents	117	138	175	233	379	571	951	1,278	1,789	2,492
top-10% important patents	200	241	321	442	681	1,025	1,637	2,137	2,984	4,105
top-25% important patents	429	525	692	913	1,348	1,919	3,002	3,823	5,217	7,112
top-50% important patents	737	863	1,073	1,355	1,958	2,684	4,172	5,221	7,021	9,417
top-1% cited patents	28	40	53	73	106	147	235	304	400	516
top-2% cited patents	51	69	90	123	185	259	432	542	717	921
top-5% cited patents	118	150	187	244	353	501	819	1,032	1,390	1,883
top-10% cited patents	207	246	319	421	630	895	1,343	1,627	2,081	2,780
top-25% cited patents	383	454	614	809	1,221	1,697	2,577	3,112	4,073	5,416
top-50% cited patents	663	759	967	1,212	1,748	2,368	3,573	4,319	5,653	7,490
mean (conditional on positive) of:										
patents	2.12	2.29	2.52	3.19	4.15	5.50	6.47	7.89	8.82	9.57
active patents	1.94	1.93	2.01	1.85	1.87	2.71	3.47	3.84	4.26	4.72
top-1% important patents	1.10	1.17	1.14	1.12	1.34	1.34	1.33	1.41	1.44	1.91
top-2% important patents	1.24	1.27	1.29	1.30	1.40	1.49	1.50	1.56	1.61	1.96
top-5% important patents	1.36	1.37	1.39	1.54	1.67	1.81	1.88	2.00	2.12	2.34
top-10% important patents	1.42	1.48	1.52	1.71	1.93	2.18	2.33	2.54	2.69	2.91
top-25% important patents	1.66	1.75	1.87	2.24	2.64	3.16	3.45	3.87	4.15	4.41
top-50% important patents	1.80	1.95	2.17	2.87	3.67	4.70	5.22	6.04	6.55	6.94
top-1% cited patents	1.32	1.33	1.51	2.34	2.91	3.28	3.11	3.35	3.32	3.25
top-2% cited patents	1.37	1.41	1.61	2.45	3.17	3.49	3.40	3.72	3.80	3.71
top-5% cited patents	1.54	1.63	1.82	2.58	3.21	3.57	3.61	3.99	4.11	4.18
top-10% cited patents	1.67	1.76	2.00	2.72	3.33	3.85	3.94	4.29	4.54	4.60
top-25% cited patents	1.75	1.85	2.12	2.86	3.51	4.25	4.42	4.99	5.36	5.57
top-50% cited patents	1.78	1.91	2.16	2.87	3.62	4.52	4.85	5.57	6.08	6.41

Notes: This table provides summary statistics for the sample of firms in the NBS data that have information on both TFP and output from 1998-2007.

Table A.II: Regressions of firm TFP and output on new patent stocks

(a) outcome: log TFP (×100); fixed effects: industry-year					
$p =$	2	5	10	25	50
i. stock of new top-p% important patents, $\beta^{topimp,p}$	0.47	0.37	0.18	0.05	0.03
	(4.55)	(5.85)	(5.65)	(4.61)	(3.82)
ii. stock of new top-p% cited patents, $\beta^{topcit,p}$	0.15	0.08	0.05	0.02	-0.00
	(2.64)	(2.72)	(2.77)	(1.59)	(-0.43)
iii. stock of new active patents, β^{active}	-0.05	-0.11	-0.15	-0.13	-0.12
	(-1.16)	(-2.52)	(-3.13)	(-2.68)	(-2.41)
iv. has patents, γ	3.66	3.63	3.64	3.66	3.68
	(22.31)	(22.13)	(22.19)	(22.36)	(22.49)
observations (m)	1.78	1.78	1.78	1.78	1.78
R^2	0.74	0.74	0.74	0.74	0.74
adjusted R^2	0.74	0.74	0.74	0.74	0.74
(b) outcome: log output (×100); fixed effects: industry-year					
$p =$	2	5	10	25	50
i. stock of new top-p% important patents, $\beta^{topimp,p}$	5.03	4.86	2.54	0.72	0.30
	(13.72)	(21.87)	(22.12)	(17.34)	(11.04)
ii. stock of new top-p% cited patents, $\beta^{topcit,p}$	0.31	0.18	0.13	-0.12	-0.21
	(1.57)	(1.74)	(0.97)	(-3.44)	(-6.18)
iii. stock of new active patents, β^{active}	1.01	0.16	-0.33	0.268	0.67
	(7.34)	(1.03)	(-2.00)	(1.62)	(3.84)
iv. has patents, γ	110.66	110.20	110.29	110.64	110.94
	(191.51)	(190.54)	(190.77)	(191.54)	(192.18)
observations (m)	1.78	1.78	1.78	1.78	1.78
R^2	0.15	0.15	0.15	0.15	0.15
adjusted R^2	0.15	0.15	0.15	0.15	0.15

Notes: This table reports the results of estimating equation (6.1) by OLS where the dependent variable is log TFP in panel (a) and log output in panel (b). Patent stocks at t include only patents applied for between $t-4$ and t. t-statistics are shown in parentheses.

Table A.III: Regressions of firm TFP and output on patent stocks with firm fixed effects

a) outcome: log TFP (×100); fixed effects: industry-year, firm					
$p =$	2	5	10	25	50
i. stock of top-p% important patents, $\beta^{topimp,p}$	0.09	0.07	0.02	0.01	0.01
	(0.96)	(1.26)	(0.73)	(0.59)	(0.93)
ii. stock of top-p% cited patents, $\beta^{topcit,p}$	-0.02	-0.01	-0.00	-0.00	-0.01
	(-0.33)	(-0.32)	(-0.27)	(-0.35)	(-0.77)
iii. stock of active patents, β^{active}	0.01	0.00	0.00	0.00	0.00
	(0.42)	(0.12)	(0.10)	(0.13)	(0.06)
iv. has patents, γ	0.78	0.78	0.78	0.79	0.79
	(3.32)	(3.28)	(3.31)	(3.33)	(3.33)
observations (m)	1.69	1.69	1.69	1.69	1.69
R^2	0.89	0.89	0.89	0.89	0.89
adjusted R^2	0.86	0.86	0.86	0.86	0.86
(b) outcome: log output (×100); fixed effects: industry-year, firm					
$p =$	2	5	10	25	50
i. stock of top-p% important patents, $\beta^{topimp,p}$	0.71	0.83	0.48	0.14	0.07
	(3.57)	(6.39)	(6.74)	(5.45)	(3.89)
ii. stock of top-p% cited patents, $\beta^{topcit,p}$	0.22	0.15	0.10	0.03	-0.01
	(1.78)	(2.38)	(2.49)	(1.41)	(-0.51)
iii. stock of active patents, β^{active}	-0.05	-0.15	-0.20	-0.21	-0.19
	(-1.11)	(-3.01)	(-3.99)	(-3.88)	(-3.52)
iv. has patents, γ	22.94	22.82	22.80	22.87	22.92
	(43.87)	(43.60)	(43.57)	(43.72)	(43.85)
observations (m)	1.69	1.69	1.69	1.69	1.69
R^2	0.86	0.86	0.86	0.86	0.86
adjusted R^2	0.83	0.83	0.83	0.83	0.83

Notes: This table reports the results of estimating equation (6.1) by OLS where the dependent variable is log TFP in panel (a) and log output in panel (b), with firm fixed effects added to each regression. t-statistics are shown in parentheses.

Table A.IV: Regressions of firm TFP and output on patent stocks, baseline vs. naive importance

(a) outcome: log TFP (×100); fixed effects: industry-year year					
$p =$	2	5	10	25	50
i. stock of top-p% important patents, $\beta^{topimp,p}$	0.27	0.26	0.16	0.06	0.08
	(2.55)	(3.96)	(4.58)	(3.88)	(4.32)
ii. stock of top-p% important patents (naive), $\tilde{\beta}^{topimp,p}$	0.62	0.15	0.03	-0.01	-0.05
	(3.94)	(2.51)	(0.86)	(-0.60)	(-2.95)
iii. stock of top-p% cited patents, $\beta^{topcit,p}$	-0.27	-0.08	0.02	0.03	0.03
	(-2.35)	(-1.14)	(0.47)	(1.43)	(2.13)
iv. stock of active patents, β^{active}	-0.05	-0.08	-0.10	-0.09	-0.08
	(-2.56)	(-3.33)	(-3.64)	(-3.25)	(-2.89)
v. has patents, γ	3.65	3.64	3.63	3.64	3.60
	(22.31)	(22.17)	(22.07)	(22.06)	(21.85)
observations (m)	1.78	1.78	1.78	1.78	1.78
R^2	0.74	0.74	0.74	0.74	0.74
adjusted R^2	0.74	0.74	0.74	0.74	0.73
(b) outcome: log output (×100); fixed effects: industry-year					
$p =$	2	5	10	25	50
i. stock of top-p% important patents, $\beta^{topimp,p}$	5.31	5.94	3.72	1.80	2.41
	(14.07)	(26.05)	(30.78)	(33.21)	(35.64)
ii. stock of top-p% important patents (naive), $\tilde{\beta}^{topimp,p}$	0.59	-2.22	-2.34	-1.73	-2.17
	(1.07)	(-10.29)	(-21.26)	(-28.90)	(-33.95)
iii. stock of top-p% cited patents, $\beta^{topcit,p}$	0.20	2.81	3.24	2.11	1.17
	(0.49)	(11.47)	(21.83)	(26.46)	(23.55)
iv. stock of active patents, β^{active}	0.32	-0.12	0.04	0.29	0.37
	(4.37)	(-1.39)	(0.44)	(2.98)	(3.76)
v. has patents, γ	110.61	109.72	109.02	108.65	108.39
	(191.40)	(189.39)	(187.82)	(186.92)	(186.29)
observations (m)	1.78	1.78	1.78	1.78	1.78
R^2	0.15	0.15	0.15	0.15	0.15
adjusted R^2	0.15	0.15	0.15	0.15	0.15

Notes: This table reports the results of estimating equation (6.1) by OLS where the dependent variable is log TFP in panel (a) and log output in panel (b), with the stock of top patents based on the naive measure of importance in equation (4.1) added to each regression. t-statistics are shown in parentheses.

C Results for granted USPTO utility patents

This section presents results for granted USPTO utility patents. Figure A.I shows the share of patents by patentee type, which is identified based on the address of the main applicant. Patents identified as belonging to "US applicants" have only US applicants, those identified as belonging to "overseas non-CN applicants" have at least one non-China overseas applicant and no applicants from China, and those identified as belonging to "CN applicants" have at least one applicant from China. Figure A.II shows how the innovativeness and crowdedness of patenting has changed over time, as discussed in section 7.1. Figure A.III shows how relative patent importance within IPC-years by patentee type has changed over time, as discussed in section 7.3. Figure A.IV shows our estimates of memory and secondary IPC importance, as discussed in section 7.4.

Figure A.I: Patent shares by patentee type — granted USPTO patents

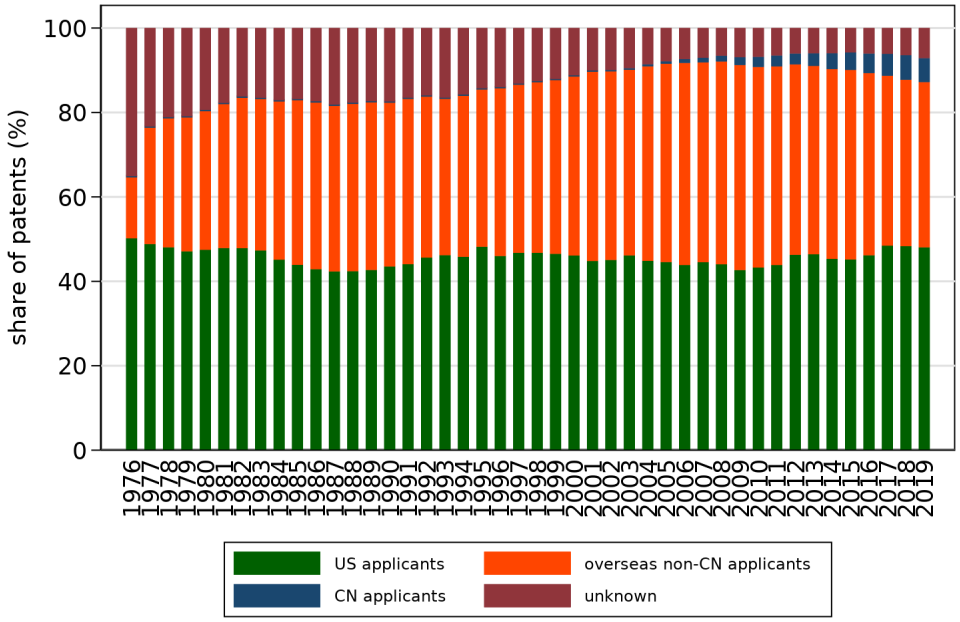


Figure A.II: Innovativeness and crowdedness of patenting – granted USPTO patents

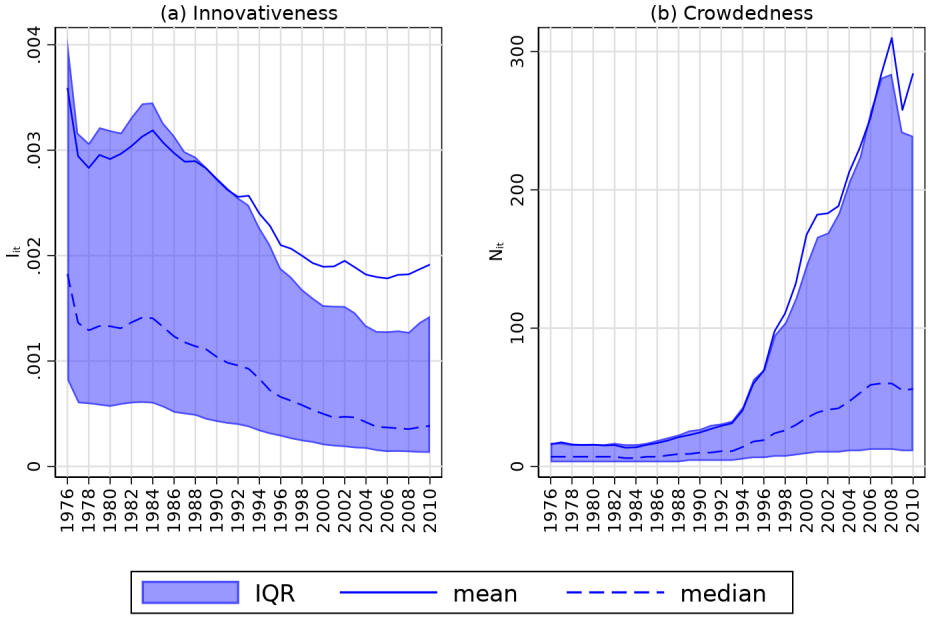


Figure A.III: Relative patent importance by patentee type – granted USPTO patents

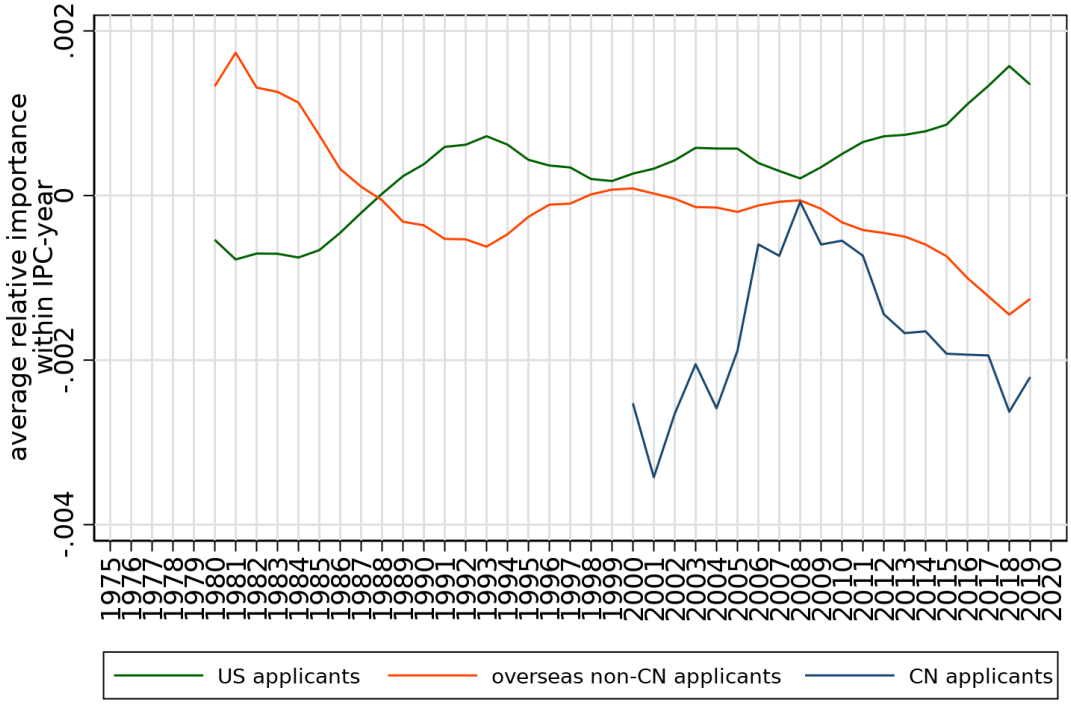
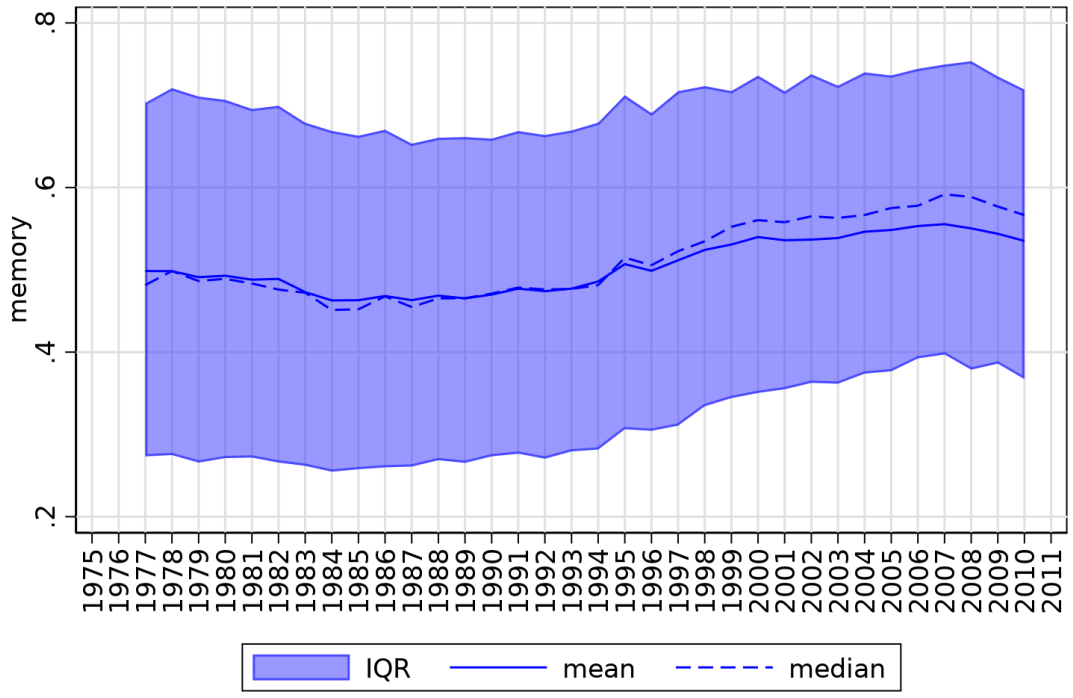


Figure A.IV: Memory and importance of seconary IPCs – granted USPTO patents



D Decomposition of patent importance by patentee type pair

Consider the OLS estimate of the coefficient vector $\vec{P}_t^g$ for the model specified in equation (4.9) with the modifications described in section (5.1). This is given by:

$$\left(\vec{P}_t^g\right)' = \left(\bar{X}_t^{g\prime}\bar{X}_t^g\right)^{-1}\bar{X}_t^{g\prime}\vec{\mu}_{Ft}^g \tag{D.1}$$

where $\bar{X^g}_t$ contains the embeddings of all right-hand side regressors. The average forward embedding $\vec{\mu}_{Ft}^g$ can be decomposed into the average forward embedding for patents owned by each patentee type s, $\vec{\mu}_{Ft}^{gs}$:

$$\vec{\mu}_{Ft}^g = \sum_s S_{t+1}^{gs}\vec{\mu}_{Ft}^{gs} \tag{D.2}$$

where S_{t+1}^s is the share of future patents attributed to patentee type s. Hence, the OLS coefficient estimate can be decomposed as:

$$\left(\vec{P}_t^g\right)' = \sum_s S_{Ft}^{gs}\left(\vec{P}_t^{gs}\right)' \tag{D.3}$$

where $\left(\vec{P}_t^{gs}\right)'$ contains measures of the importance for patenting by patentee type s:

$$\left(\vec{P}_t^{gs}\right)' = \left(\bar{X}_t^{g\prime}\bar{X}_t^g\right)^{-1}\bar{X}_t^{g\prime}\vec{\mu}_{Ft}^{gs} \tag{D.4}$$

Online Appendix to

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

by

Philipp Boeing, Loren Brandt, Ruochen Dai, Kevin Lim, and Bettina Peters

A Basic patent statistics

Technology classes. Panel (i) of Table 1 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 1 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 2 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.[16] 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 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.

[16]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 1: 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.

Table 2: 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.

B 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 1 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 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.

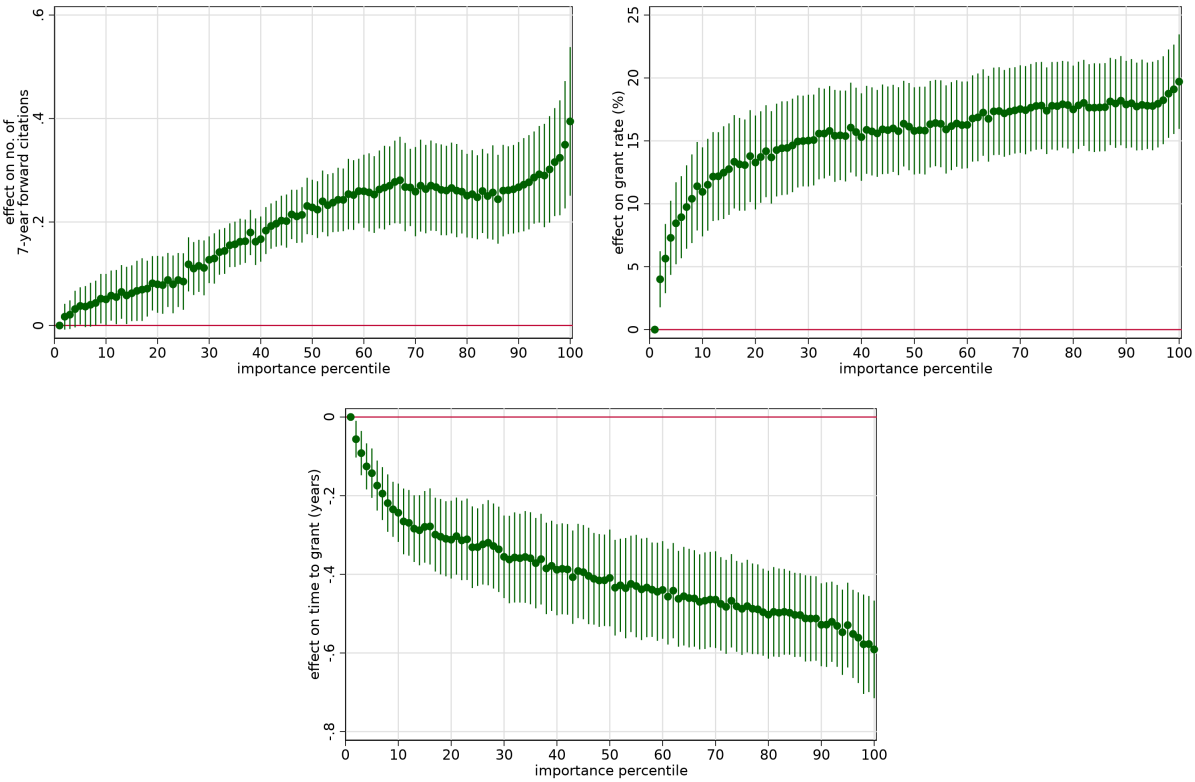
C Trends in patent shares and importance by patentee type

Figure 1 shows the same information as panel (a) of Figure **??** 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 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 2 shows the same information as Figure **??** 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.

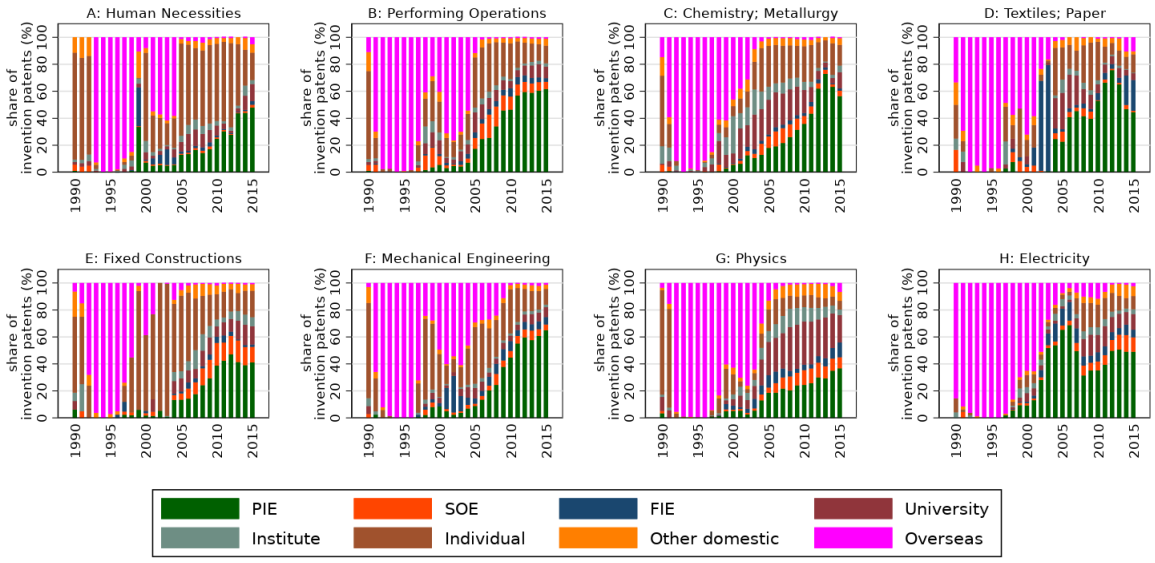
In panel (a), we observe that the timing of the decline in average importance in the early 2000s differs slightly across sections. For example, the declines occurred earlier in sections E (Fixed Constructions) and F (Mechanical Engineering). Nonetheless, in all IPC sections, average patent importance 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

Figure 1: 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 (**??**). 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.

Figure 1: Patent shares by patentee type for top 5% most important patents



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.

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.

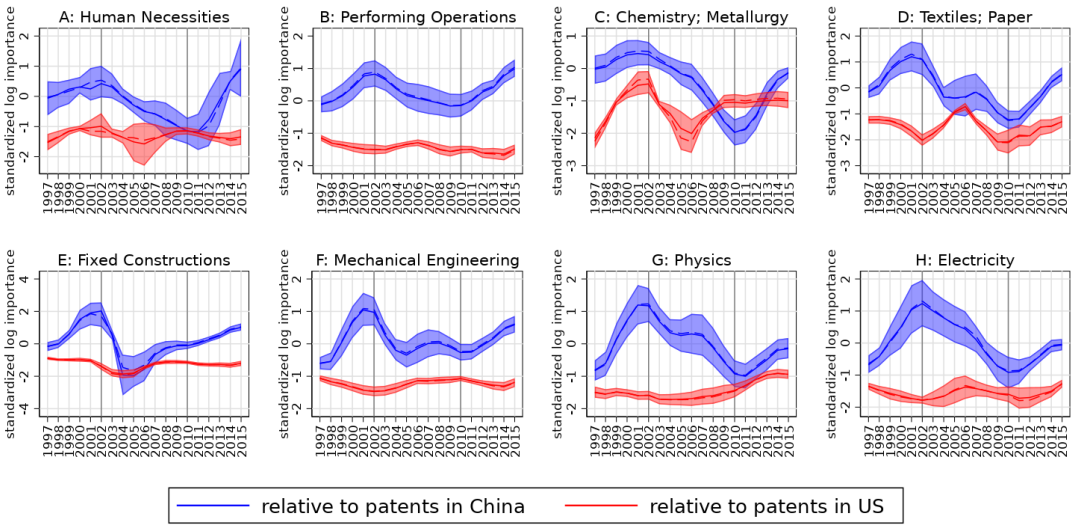
D Traditional measures of patent quality in China

Number of forward citations. Figure 1 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.[17]

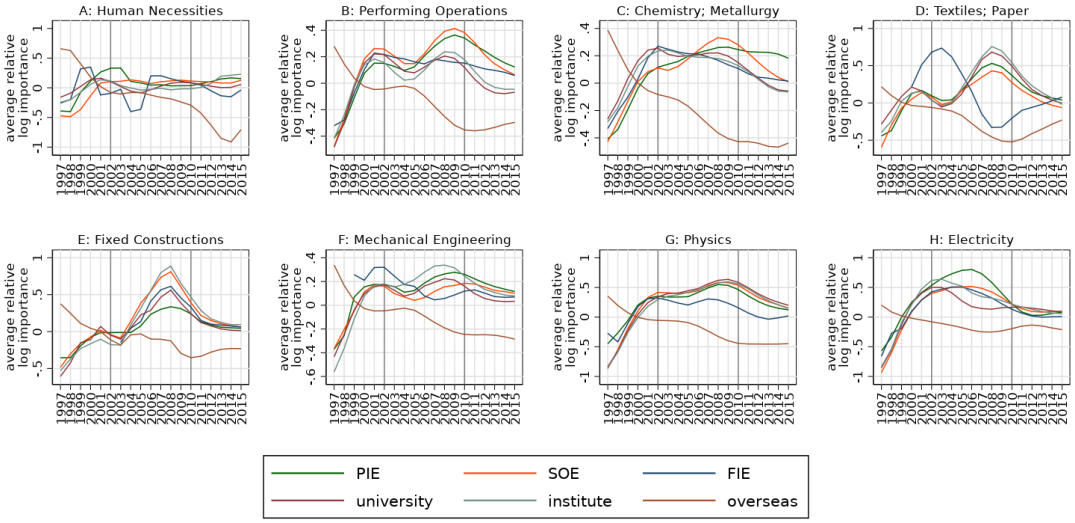
[17]There is also censoring in the earlier years of our sample, as we observe fewer backward citations per patent application for patents that are applied for before 2000. Hence, patents applied for in earlier years only receive

Figure 2: The evolution of patent importance by patentee type

(a) Average patent importance

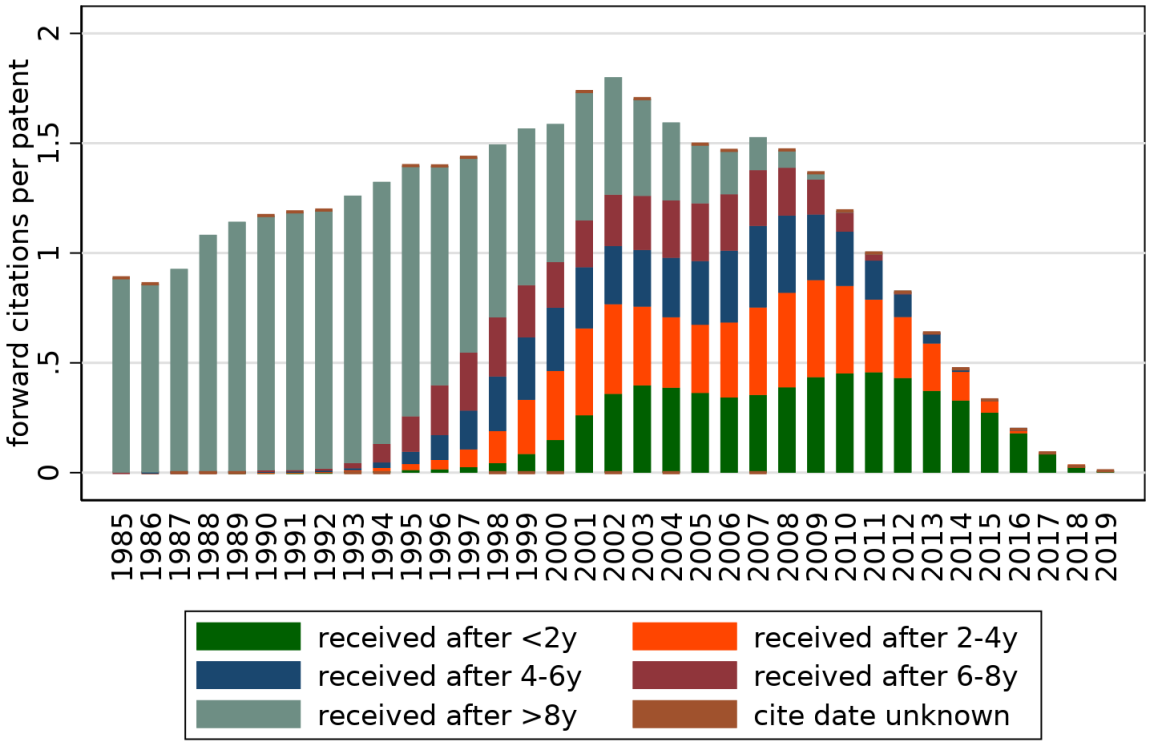


(b) Relative patent importance by patentee type



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 1: 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.

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.[18] Figure 2 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 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.

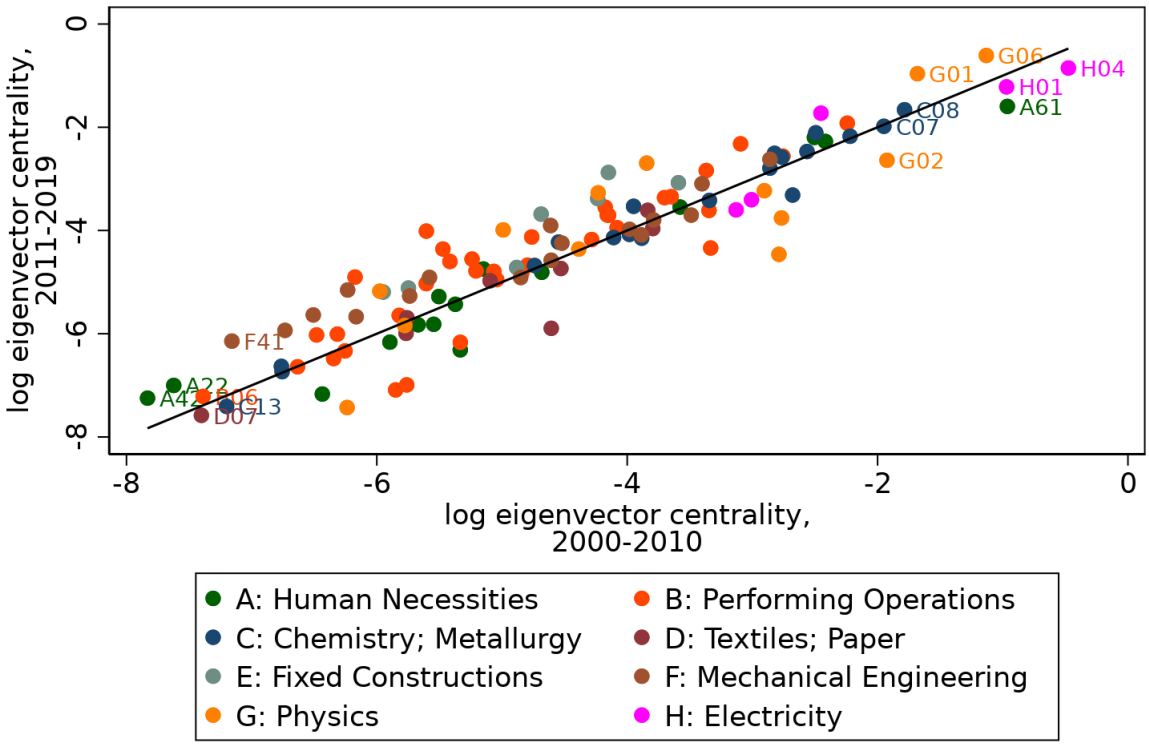
Figure 3 shows the share of patents that fall into each of these eight categories in each year, by the main IPC section of each patent. Patent grant rates are between 40-60% in most cases. We observe increasing grant rates in all IPC sections throughout the 1990s, followed by falling grant rates from 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 around ten years after the application date of a patent.

Timing of legal status change events. Figure 4 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

citations much further into the future.

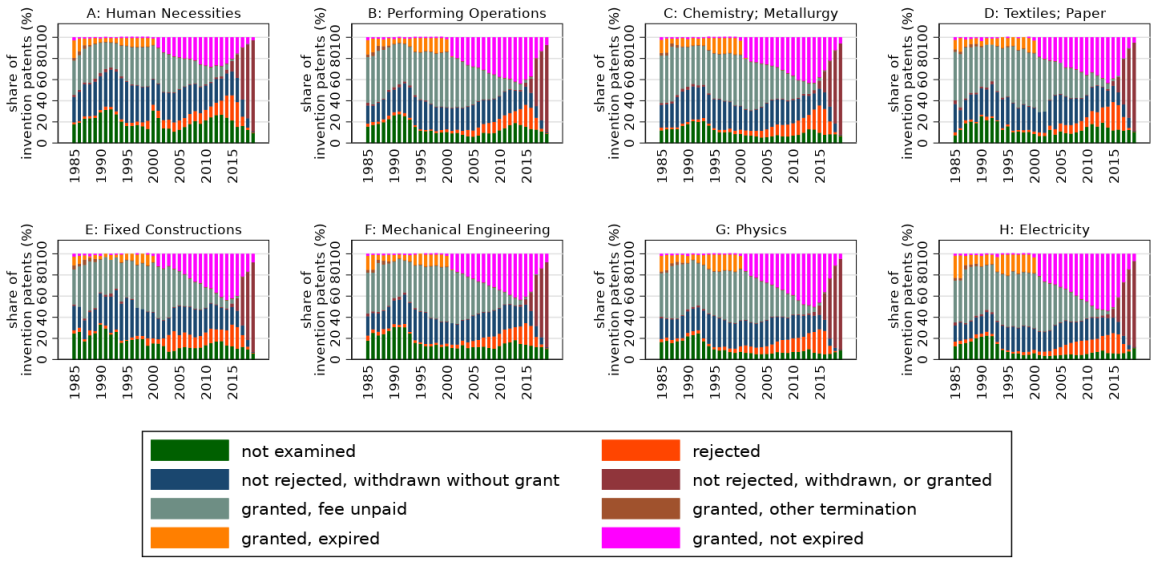
[18] The eigenvector centrality of a node i is defined recursively as $\lambda c_i = \sum_{j=1}^{N} w_{ij} c_j$, where N is the number of nodes in the graph, w_{ij} is the weight of the link from node i to j, and λ is the largest eigenvalue of the weighted adjacency matrix (a matrix with ij-element equal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

Figure 2: 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 3: Patent shares by legal status



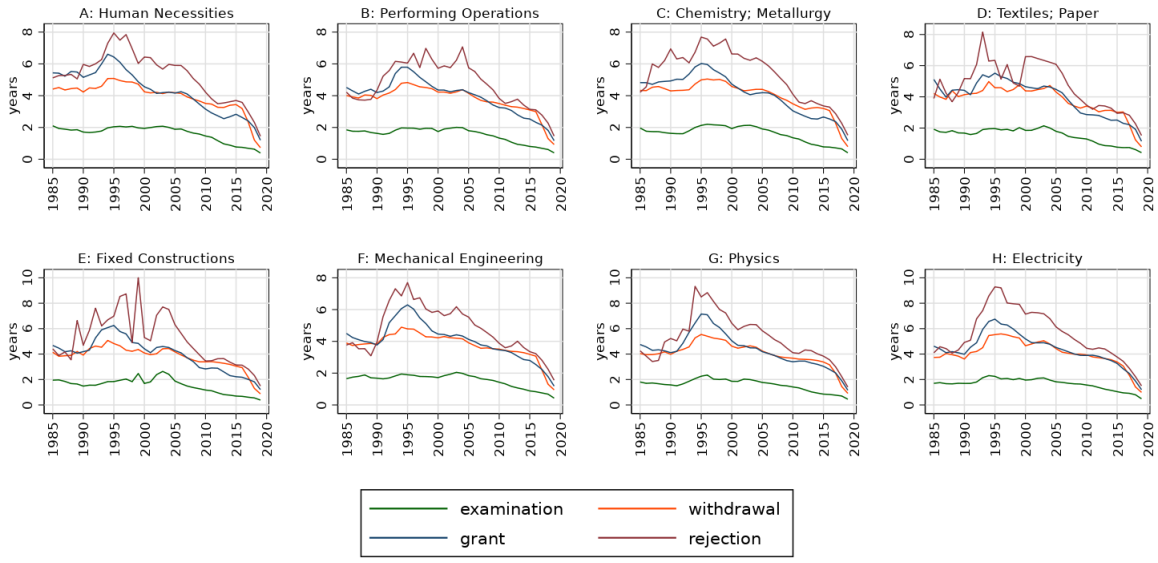
Notes: This figure shows the shares of invention patent applications within each IPC section and year based on eight mutually-exclusive legal status events.

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 5 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.

Priority filings at the USPTO. Figure 6 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.

Figure 4: 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.

E 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 $i \in \Omega_t$:

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

We can write this equivalently as:

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

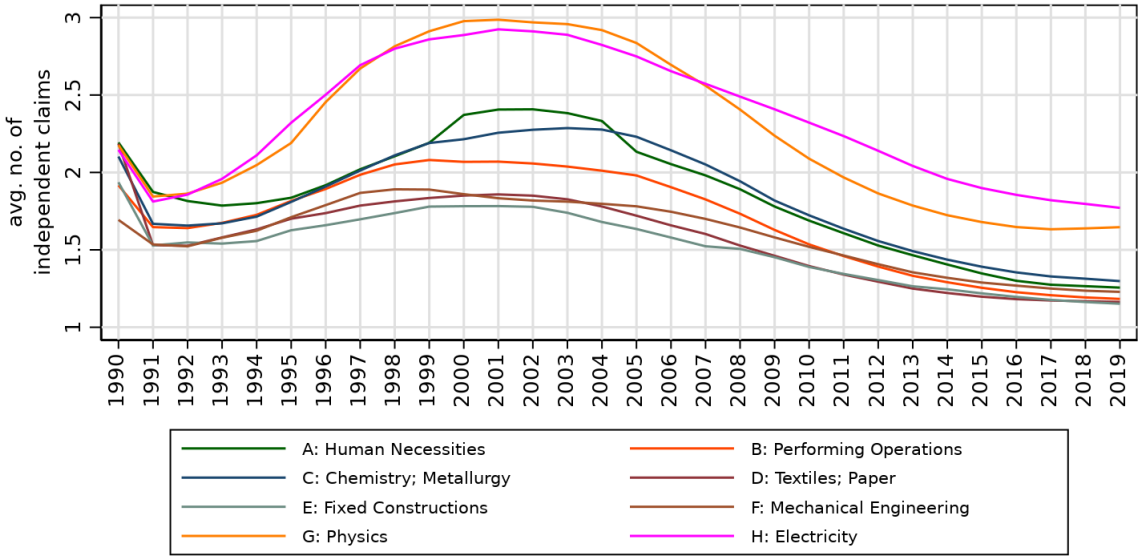
where $s_{jt} \equiv \frac{N_{jt}}{N_t}$ is the share of observations accounted for by class j and $\bar{X}_{jt} \equiv \frac{1}{N_{jt}} \sum_{i \in \Omega_{jt}} X_{it}$ is the average outcome within class j. Totally differentiating, we obtain:

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

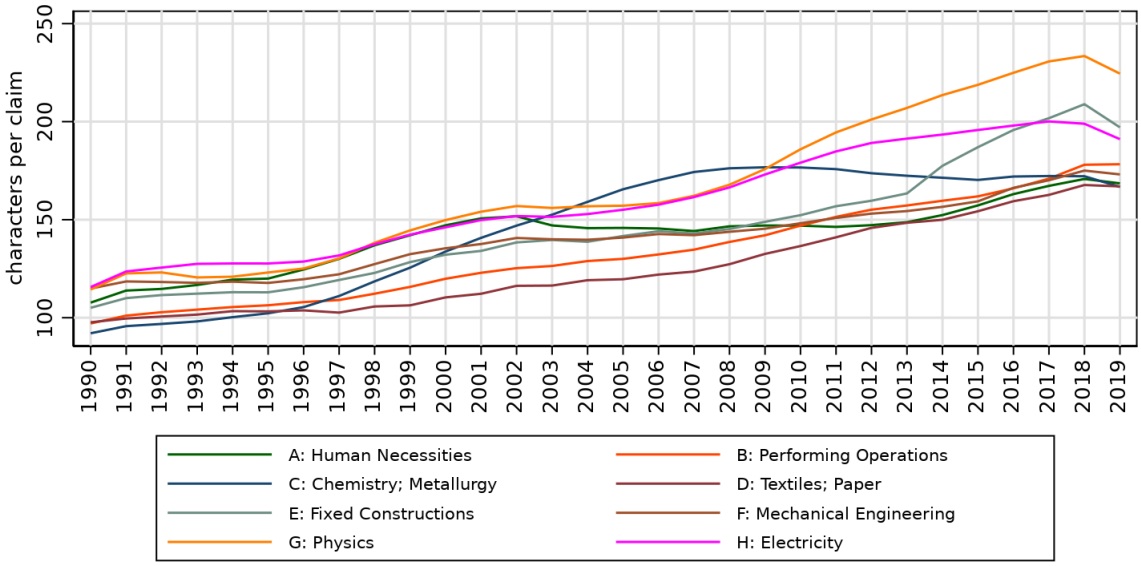
where $\hat{y} \equiv \frac{dy}{y}$ indicates the marginal log change in variable y and $r_{jt} \equiv \frac{\bar{X}_{jt}}{\bar{X}_t}$ is the ratio of the average outcome in class j to the average outcome overall.

Figure 5: Number and length of claims per patent

(a) Number of independent claims

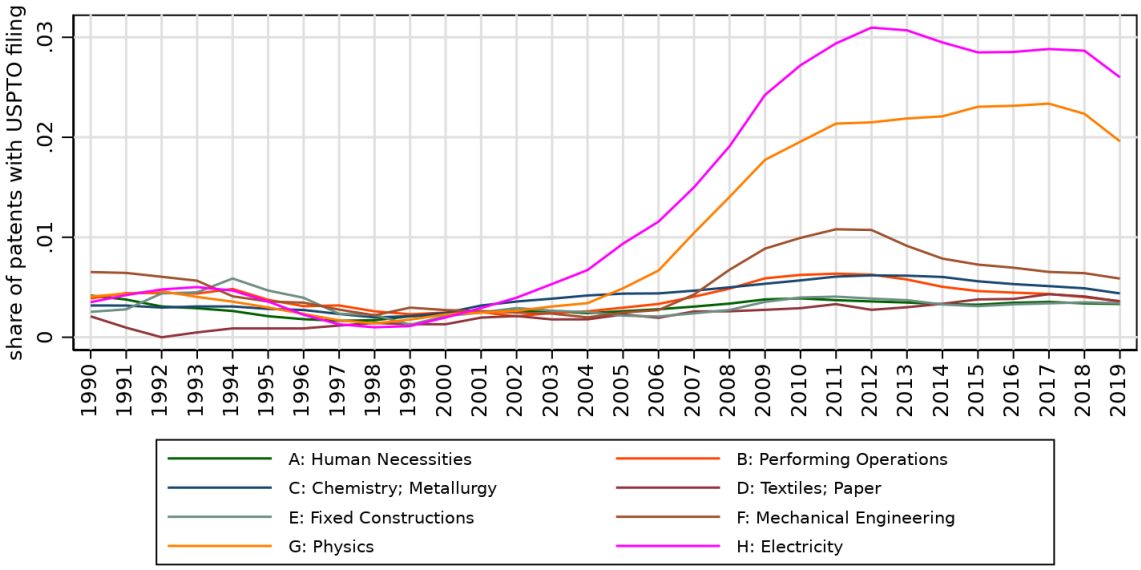


(b) Average length of claims



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 6: 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.