The Impact of the Trade War: Divergence in Chinese and U.S. Innovations in the Post-Conflict Era*

Xiao Ma Yueyuan Ma Hanyi Tao Yiran Zhang

This Version: Oct. 2024

Abstract

This paper examines the impact of the US-China trade war on China's innovation intensity and direction. Employing a textual analysis method to measure the direction of Chinese firms' innovation compared to that in the US by assessing the similarity between the text of Chinese and US patents, this study finds that increased exposure to US export tariffs reduces the similarity, especially with recent US patents. China's innovation similarity with other developed countries also declines, though to varying extents. Additionally, export tariffs are shown to decrease patent filings in China. To explain these effects, we incorporate the textual analysis algorithm into a quantitative economic model where firms endogenously choose their innovation efforts in various directions within each product. Quantification analysis shows that the demand channel accounts for 48% of the decline in China-US innovation similarity due to tariff shocks. Innovation intensity and direction choices contribute to a 6% decline in Chinese firms' export sales, with direction alone accounting for 1.68%.

JEL Code: F13, F14, O31, O34

Keywords: trade war, innovation direction, textual analysis.

*Xiao Ma: Peking University. Email: xiaoma@phbs.pku.edu.cn, Yueyuan Ma: UC Santa Barbara. Email: yueyuanma@ucsb.edu. Hanyi Tao: ShanghaiTech University. Email: taohy@shanghaitech.edu.cn. Yiran Zhang: Fudan University. Email: yiran_zhang@fudan.edu.cn. We thank Jie Cai, Rafael Dix-Carneiro, Gordon Hanson, Guangzhou Hu, Yang Jiao, Bingjing Li, Ivan Png, Natalia Ramondo, Micheal Song, Gonzalo Vazquez-Bare, and seminar participants at Tsinghua, CICM, and CEIBS for helpful comments. All potential errors are our own.

1 Introduction

A significant headwind to decades of globalization is the US-China trade war, which began in early 2018 when the Trump administration implemented tariffs on imported steel and aluminum (Bown, 2021). Subsequent measures included additional tariffs on Chinese goods and export sanctions on specific Chinese firms, particularly in high-tech and industrial sectors. The primary aim of this trade conflict was to address what the US considered unfair trade practices by China, such as intellectual property theft, forced technology transfer, and trade imbalances, while also considering national security concerns about China's technological advancements. In retaliation, China imposed higher tariffs on US products, particularly agricultural products.

Given that technology is a key point of contention in the conflict and innovation acts as a pivotal catalyst for technological progress, this paper aims to investigate the following two questions: What impact does the trade war have on the innovation intensity and trajectory of Chinese firms? How does the trade war shape the performance of Chinese firms through the innovation channel?

To address these two questions, this paper first constructs a matched dataset that contains comprehensive details on the operational activities, patent filings, and export and import volumes of all publicly listed Chinese firms from 2000 to 2021. The number of patent applications is utilized as an indicator of firms' innovation intensity. In order to assess the technological trajectory of Chinese firms' innovations, we adopt a novel text-based metric that evaluates the similarity between Chinese patents and patents from other major patenting regions worldwide, such as the US, Europe, Japan, and South Korea. Specifically, we employ the Term Frequency-Inverse Document Frequency (TF-IDF) method, a widely recognized statistical technique in the field of natural language processing, to transform patent abstracts into vectors, with a focus on the frequency distribution of informative terms. Subsequently, we calculate the cosine similarity between patents filed by Chinese firms and those originating from other countries. This text-based metric sheds light on the technological alignment between Chinese and foreign patents, providing valuable insights into the trajectory of Chinese innovation progress.

This study empirically analyzes the effects of the trade war on firms' innovation intensity and technological trajectories using a Difference-in-Differences approach. Firm-level exposure to the trade war is captured by changes in US export tariffs, US import tariffs, and a dummy variable indicating whether a firm was subject to sanctions imposed by the US government. The export and import tariff changes are set to zero for the period 2014–2017 and are calculated as the difference between the tariff rates faced by each firm after the trade war (2018–2021) and their average tariff rate during 2014–2017. To address potential endogeneity concerns arising from the trade conflict's effect on trade volumes, we measure firms' exposure to export and import tariffs based on their pre-trade-war trade composition using customs data, combined with actual tariff rates for each product category post-trade-war. In our analysis, we control for firms' operational characteristics and include multiple fixed effects.

The regression analysis shows that increases in US export tariffs lead to a significant reduction in the innovation intensity of Chinese listed firms and contribute to a technological divergence from US patents. Specifically, a 10 percent rise in export tariffs results in a 3.01 percent decrease in firms' patent filings, while reducing the similarity between Chinese and US patents filed in the past five years by 2.04 percent from the historical average. This divergence is especially pronounced for recent US patents, suggesting a stronger impact on cutting-edge technology. These findings indicate that the demand shock caused by export tariffs diminishes Chinese firms' incentives to remain competitive in the US market. Robustness checks show that the regression results are not driven by pre-existing trends, strategic changes of wordings of patent abstracts after the trade war, or the specific choice of the similarity measure. Over time, the effects on both patent filings and technological similarity intensify, highlighting a growing divergence in the technological paths of the two countries. In contrast, changes in US import tariffs do not significantly affect patent applications, yet they result in a comparable negative impact on the similarity to both past and recent US patents. This suggests that import tariff changes do not operate through the demand channel. Sanctions, while positively associated with patent filings, have a weak effect on patent similarity during the sample period. The increase in patent filings could reflect the tendency of sanctions to target more innovative Chinese firms, while the limited effect on similarity may be due to the time lag required for the impact of sanctions to materialize.

This study further examines the impact of rising U.S. export tariffs on the similarities between Chinese patents and those originating from Europe, Japan, and South Korea, while controlling for changes in export and import tariffs with these regions. The results show a decline in patent similarity with varying magnitudes, indicating the multidimensional nature of innovation. Specifically, a 10 percent increase in U.S. export tariffs leads to a 2.03 percent decrease in similarity to European patents filed in the past five years, a 1.85 percent decrease in similarity to Japanese patents, and an insignificant 1.28 percent decrease in similarity to South Korean patents. However, when controlling for the firm's innovation similarity to the U.S., the effect on patent similarity with other regions disappears. This suggests that the divergence between Chinese and other regions' innovations is primarily fueled by the growing distance between Chinese and U.S. innovations.

Which types of firms are more sensitive to export shocks? The analysis shows that Chinese firms' patent similarity to the most innovative U.S. firms—defined as the top 20 firms by patent application volume in each technological class—experiences a larger decline in response to rising U.S. export tariffs. Moreover, high-TFP Chinese firms witness a greater reduction in patent applications, and their innovation activities shift further away from those of the U.S. firms, especially the most innovative ones. Thus, the widening innovation gap between the two countries is primarily driven by the most innovative and productive firms on both sides.

A notable example of a shift in innovation direction can be seen in Xiaomi, the third-largest smartphone company in China. After facing higher tariff rates on exports to the US market, Xiaomi has made significant strides in camera system innovation to appeal to both Asian and European markets, where high-quality imaging is particularly valued by photography enthusiasts.[1] This case illustrates how a firm's innovation strategy is heavily influenced by the preferences of the markets it serves. To systematically analyze the mechanism by which export tariff shocks impact

[1]Xiaomi partnered with Leica, a renowned camera manufacturer, to enhance its smartphones' photographic capabilities.

firms' innovation activities, we develop a partial equilibrium model that focuses on multi-product, multi-destination firms with heterogeneous preferences across export markets. In the model, firms' products are represented by a vector of features, which correspond to the vectorized patent texts in the TF-IDF method. The productivity of each feature is contingent upon both the firm's overall innovation intensity and the innovation direction on the vector space. Firms make decisions on both the intensity and direction of innovation and their participation in the export market for each product. Unlike conventional trade models that assume symmetry preferences across destination countries, our model integrates distinct country-specific tastes for each product variety. Consequently, changes in export tariffs directed at specific destinations not only influence the overall level of innovation intensity through shifts in the total market size but also redistribute innovation efforts across different product features. Specifically, our model predicts that an increase in tariffs on Chinese exports to the US would reduce innovation intensity among Chinese firms and prompt a reorientation of Chinese innovations away from US preferences.

To evaluate the effectiveness of the model in explaining the impact of the export shock and to assess the role of innovation decisions in shaping firms' performance in response to the trade shock, we conducted a quantitative analysis based on the model. The model is calibrated to the period before the trade war, with heterogeneous preferences inferred from the initial TF-IDF vectors of firms' patent applications and firms' product sales across various destination markets. Simulation exercises are performed to capture firms' decision-making with and without the unexpected changes in export tariffs during the trade-war period. The model predicts a 0.98% decline in Chinese patent similarity to U.S. innovations, following a 10% increase in export tariffs, compared to a 2.04% decline observed empirically. Thus, the model explains 48% of the actual divergence in innovation direction between China and the U.S.

Further counterfactual analysis highlights the significant role of firms' innovation decisions in determining their export performance in response to tariff shocks. Changes in innovation intensity and direction result in a 6% decline in Chinese firms' export sales to the U.S. by 2021. Of this decline, 28% is attributable to shifts in R&D direction, leading to a 1.68% reduction in Chinese firms' export sales to the U.S.

Related Literature. Our paper is related to several strands of the literature. First, this paper closely connects with the growing literature on understanding the effects of the trade war (e.g., Fajgelbaum, Goldberg, Kennedy and Khandelwal, 2019; Amiti, Redding and Weinstein, 2019; Fajgelbaum, Goldberg, Kennedy, Khandelwal and Taglioni, 2023; Bonadio, Huo, Kang, Levchenko, Pandalai-Nayar, Toma and Topalova, 2024), especially from the perspective of Chinese firms (e.g., Benguria, Choi, Swenson and Xu, 2022; Jiao, Liu, Tian and Wang, 2022). Most of the existing studies on the trade war primarily examine its impact on global trade patterns and welfare,[2] typically assuming firms' productivity as given. On the contrary, our paper concentrates on the dynamic influence of the trade war on firms' productivity by altering the quantity and direction of inventions, leading to the potential prolongation of the trade war's impact.

By centering on innovation, our paper maintains a strong linkage to an extensive literature on trade and innovation. Extensive empirical evidence highlights the influence of changes in trade exposure on R&D expenditures and productivity, as demonstrated in studies such as Autor, Dorn, Hanson, Pisano and Shu (2020) for US firms, Lileeva and Trefler (2010) for Canadian firms and Bloom, Draca and Van Reenen (2015) and Aghion, Bergeaud, Lequien and Melitz (2018) for European firms. Additionally, empirical analysis has been conducted on how Chinese firms' innovation quantities respond to trade liberalization, as evidenced by studies like Liu and Qiu (2016), Bombardini, Li and Wang (2017), and Liu, Lu, Lu and Luong (2021). This literature typically focuses on the impact of trade liberalization, whereas we study an era where trade became more extravagant. We also contribute to this literature by providing a comprehensive analysis of the impact of trade on both quantity (amount of inventions) and direction (the similarity of inventions with foreign countries' technology) of innovation. The multi-dimensional nature of innovation found in this paper shows that the content of patents is also significantly affected by trade shocks, and is

[2]One exception is Benguria et al. (2022), who show that higher trade policy uncertainty induced by the trade war affects firm investments.

therefore, worth more attention in future research.

Finally, our paper contributes to the literature on patent content measurement by integrating contextual analysis into a quantitative economic model. Previous studies on patents have largely focused on structural information, such as patent counts, citations, and technology classifications (Hu and Jefferson, 2009; Lerner and Seru, 2017). Correspondingly, innovation models have primarily emphasized the quantity or quality of innovation. However, a wealth of information is embedded in the unstructured data of patent texts. Recent research has begun to develop new measures by analyzing text similarity between patents and product files, as well as between earlier and later patents, to assess their scientific or commercial value and direction (Comin and Hobijn, 2010; Hoberg and Phillips, 2016; Gentzkow, Kelly and Taddy, 2019; Bloom, Hassan, Kalyani, Lerner and Tahoun, 2021; Kelly, Papanikolaou, Seru and Taddy, 2021). To our knowledge, no study has yet mapped textual analysis algorithms onto economic models. Our paper bridges this gap by linking word frequencies from the TF-IDF method to the preferences of the markets where firms sell their goods. This allows for a direct quantification and evaluation of how shifts in market preferences, driven by tariff shocks, influence the direction of innovation. Additionally, this paper expands the literature by incorporating patents from multiple patent offices, including those in China, Japan, Korea, and Europe, and develops a text-based metric to assess technological similarities between patents across countries beyond the US.

The rest of the paper is organized as follows. Section 2 describes the background of the US-China trade war, the data sources, and the methods to construct key variables in the empirical analysis. Section 3 introduces the empirical strategy and presents the impact of the trade war on Chinese firms' innovation intensity and direction. Section 4 lays out a quantitative model to unveil the mechanisms in the empirical analysis. Using the calibrated model, we quantify the impact of traded war in Section 5. Section 6 concludes.

2 Context and Data

In this section, we begin by outlining the background of the US-China trade war. Next, we discuss the data sources and detail how we construct the key measures used in our empirical analysis.

2.1 A Brief History of the US-China Trade War

The China–United States trade war began in January 2018 when then U.S. President Donald Trump started imposing tariffs and trade barriers on China. The main objectives were to address what it considered to be unfair trade practices by China, including intellectual property theft, forced transfer of American technology to Chinese companies, and imbalances in the U.S.-China trade relationship. Despite a phase one agreement reached in January 2020, the conflict continued throughout Trump's presidency. President Joe Biden has maintained the tariffs, and as of early 2024, Trump's campaign considered a 60% tariff on Chinese goods.

The United States imposed tariffs on a wide range of Chinese goods, starting with solar panels and washing machines in January 2018, and eventually extending to various other products including steel, aluminum, and a variety of other goods across different sectors. The list expanded to cover technological and industrial goods, particularly focusing on products related to China's ”Made in China 2025” initiative, which aims to make China dominant in global high-tech industries. By July 2018, the U.S. began imposing tariffs on 34 billion worth of Chinese products, extending to 200 billion by September 2018 and eventually covering 250 billion worth of goods by May 2019. The tariffs targeted a broad spectrum of products, from consumer electronics to textiles and agriculture products, aiming to pressure China on trade practices the U.S. deemed unfair.

China retaliated by imposing tariffs on U.S. goods in several rounds, affecting a wide array of products, including agricultural products, automobiles, and seafood. The Chinese government's response was strategically targeted to impact key U.S. industries, particularly those in states with significant political importance. China's tariffs were seen as a direct countermeasure to the U.S. tariffs, aiming to hurt the U.S. economy in areas where it could potentially influence political pressure on the U.S. administration to change its policies.

Table 1 reports the products that are affected by the tariff escalation the most. The products are defined by the Harmonized System (HS) codes, a standardized numerical method of classifying traded goods. We calculate the difference between the average tariff in 2018-2021 and the tariff in 2017 for each 8-digit HS code, then list the products with the most significant positive changes. The exercise is done for both exports to the US and exports to China. As shown in Table 1, among Chinese goods exported to the US, manufacturing products, especially electrical and power equipment, experienced the most significant increase in export tariff. Among US goods exported to China, agricultural products were imposed the highest tariffs.

Table 1: Products with Largest Increases in Tariff, in Percentage Points

Export to the US		Export to China	
HS Product	Tariff Change	HS Product	Tariff Change
Generators	45.0%	Meat, of swine	55.0%
Electric accumulators	45.0%	Offal, edible	55.0%
Electrical apparatus	35.6%	Aluminum	50.0%
Iron	32.5%	Nuts, edible	45.0%
Steel	32.5%	Fruit, edible	45.0%

Notes: This table shows the products (measured by the 8-digit HS code) that experienced the largest increase in tariff due to the trade war. The left panel lists the exporting goods from China to the US and the corresponding percent increase in tariff; the right panel lists the exporting goods from the US to China and the corresponding percent increase in tariff.

In addition to tariffs, the U.S. government implemented export controls and sanctions on Chinese firms. These measures were designed to restrict Chinese companies' access to U.S. technology, especially those technologies that could have military applications or contribute to the enhancement of China's surveillance capabilities.

2.2 Data Sources

We begin our analysis by constructing a matched dataset with information on Chinese listed firms' operations, patents, and trade from 2000 to 2021. This dataset is compiled from four different sources, enabling us to conduct a comprehensive study on the effect of the trade war on Chinese listed firms.

The first database, the China Stock Market & Accounting Research Database (CSMAR henceforth), provides financial reports of all firms listed in the Chinese stock market. We collect firm name, industry type, ownership, sales, employment, capital stock, R&D spending, and export destinations. We clean firms' basic information and financial reports in the CSMAR data following Tan, Tian, Zhang and Zhao (2020).

The second source is the China Customs Trade Data (CCTD henceforth). This dataset offers detailed information about firm-level trade transactions from 2014 to 2016, including information on firms' names, trade destination countries (for exports) and origin countries (for imports), eight-digit HS product codes, and the value of their exports and imports in U.S. dollars. We merge the CCTD data with the listed firm data using consolidated firm names (He, Tong, Zhang and He, 2018) to construct listed firms' exposure to tariff changes during the trade war period.

The third dataset consists of Chinese patent data from the China National Intellectual Property Administration (CNIPA) and US patent data from the United States Patent and Trademark Office (USPTO). The Chinese patent data covers all the invention patent filings between 1985 and 2023, detailing an applicant's bibliography information, filing date, grant date, abstract, and references (patents) cited. It also provides the English translations of the abstracts, which are used to construct similarities with US patents. The USPTO provides records of patent grants from 1976 and patent filings since 2000. We collect the same indicators as for Chinese patent data. Moreover, we collect records of patent filings for European countries, Japan, and Korea from the PATSTAT Global 2023 Autumn Edition for supplementary analysis. We describe the data-cleaning process in detail in the following section.

For our analysis, we utilize tariff data from Bown (2021) to construct the US tariff rates on imports from China and China's tariff rates on imports from the US during the trade war period between 2017 and 2021. The raw data is based on 10-digit Harmonized System (HS) products for the US and 8-digit HS products for China. To determine the tariff rate for each year, we calculate the tariff rate on December 31 of that year, taking into account all tariff changes throughout the year. In order to measure tariff rates prior to the trade war, we rely on reported tariff data from the World Integrated Trade Solution (WITS) between 2014 and 2016, which is based on 6-digit HS products. To ensure consistency in product classification across different data sets, we aggregate the tariff data from Bown (2021) into 6-digit HS products using trade volume as weights. Furthermore, we converted the 6-digit HS codes in 2017 – 2021 into the version used during 2014 – 2016 by employing the concordances provided by WITS.

2.3 Text-based Patent Similarity

In order to properly measure the similarities between Chinese patents and patents in foreign countries, we first define the scope of Chinese patents and U.S. patents. In USPTO and CNIPA, both domestic residents and foreigners can apply for patents. Simply treating the patents filed in CNIPA as the Chinese patents and patents filed in USPTO as the US patents is misleading. Thus, we define Chinese patents as those that are filed in CNIPA by Chinese domestic residents, including firms, individuals, universities, and research institutes. We apply a similar rule for patents filed in USPTO to identify patents filed by U.S. domestic residents.

Next, we clean patent abstract data following standard procedures in literature (Bloom et al., 2021). We first remove symbols and numbers and only keep English letters in abstracts. Then, we lemmatize all nouns and verbs with Standard CoreNLP 4.5.4 (Manning, Surdeanu, Bauer, Finkel, Bethard and McClosky, 2014), which converts nouns from plural to singular and converts verbs to bare infinitives. These procedures turn each piece of patent abstract into a list of tokens, where each token is a lemmatized word.

With all patent abstracts cleaned, we adopt the TF-IDF algorithm to vectorize each piece of abstract. Compared to other methods of textual analysis, the TF-IDF algorithm can generate economic meaning through a quantitative model as this paper illustrates in Section 5.[3]. Before vector-

[3]Besides, the TF-IDF method is widely used in the literature (Acemoglu, Yang and Zhou, 2022; Kelly et al., 2021; Autor, Chin, Salomons and Seegmiller, 2024). We also adopt methods based on the Large Language Model to vectorize patent abstracts as robustness checks and get similar results.

ization, we first calculate the document frequency of each word and remove the too-frequent words and too-infrequent words. The document frequency is the count of a word's appearance in different pieces of abstracts. If a word appears in too many patents, it means that this word is not informative in representing the technical features of patents. If a word only appears in a few patents, it is likely to be a typo or man-made word, which is also not informative in representing the technical features of patents. In this paper, we put Chinese and U.S. patents together and drop words with a document frequency larger than 100,000 and lower than 20 (Bloom et al., 2021). Then, we apply the TF-IDF method to vectorize all patent abstracts. The size of the vector is the total number of unique tokens (words). Each point in the vector is the term frequency of a word, which is the count of the appearance of a word within a patent abstract divided by its document frequency. Intuitively, each vector represents the technical features of a patent by highlighting words of relatively high term frequency to document frequency. Thus, the text similarity between the two patents is a good indicator of the technical similarity between them.

Then, we construct the similarity measures between Chinese firms' patents and U.S. patents. We begin our analysis by calculating the cosine similarity between each Chinese and U.S. patent filed in the same technology class, measured at the 3-digit IPC level. We only compare patents within the same technology class since it is meaningless to compare the technical features between biological patents and semiconductor patents. Moreover, the text similarity between patents in the same field precisely captures the differences in technological trajectory between Chinese and US patents. The similarity between a Chinese patent p filed in year t and technology class x and a U.S. patent p' filed in year τ and technology class x is defined by Equation (1). As a result, we obtained a list of similarities between each Chinese patent filed in each technology class x and in each year t to the U.S. patents filed in the same technology class x from 2000 to 2021.

$$\mathrm{Sim}_{p,p',t,\tau,x} = Cos\{V_{p,t,x}, V_{p',\tau,x}\} \tag{1}$$

We obtain the country-level and firm-level similarities between Chinese and U.S. patents by aggregating the patent-level similarities. For instance, in order to compare the similarity between a Chinese listed firm i and the U.S. in year t and technology class x, we calculate the pairwise similarity between the patents filed by firm i filed in year t and technology class x and all U.S. patents filed in year τ and technology class x with Equation (2). Here, $N_{i,t,x}$ represents the number of patents firm i applied in technology class x and year t, and $M_{US,\tau,x}$ represents the number of U.S. patents in class x and year τ.

$$\text{Sim}_{i,t,\tau,x} = \frac{1}{N_{i,t,x}} \frac{1}{M_{US,\tau,x}} \sum_{p \in i} \sum_{p' \in US} Cos\{V_{p,t,x}, V_{p',\tau,x}\} \tag{2}$$

Then, the firm-level similarity to U.S. patents of firm i is defined as the weighted average of similarities across technology classes $x \in X$. More details of the construction process are presented in the Appendix.

2.4 Exposure to Trade Shocks

To measure the extent to which Chinese listed firms were affected by tariffs during the trade war period, we utilize tariff and customs data. More specifically, for each firm i, we calculate its exposure to the US tariffs by relying on its export composition during the pre-trade-war period (2014-2016, based on available data). This calculation is performed using the following formula:

$$\text{exposure to US tariff}_{i,t} = \sum_j \frac{\text{export}_{i,j,14-16}}{\sum_j \text{export}_{i,j,14-16}} \text{tariff}_{j,t}^{US}. \tag{3}$$

Our firm-level customs data allows us to designate $export_{i,j,14-16}$ as the value of exports for firm i regarding 6-digit HS product j between 2014 and 2016. $\text{tariff}_{j,t}^{US}$ indicates the tariff rate that the US imposed on the import of product j from China during year t. By this formula, we gauge the extent to which firm i is exposed to US tariffs, which portrays the average tariff rates faced by the firm when exporting to the US in year t based on its pre-trade-war export structure.

As a response to the deteriorating trade environment in the US, China implemented retaliatory measures by increasing tariffs on imports from the US. This escalation in import tariffs could potentially impact Chinese listed firms, particularly through changing the competition in firms' output market and the prices of imported inputs (Brandt, Van Biesebroeck, Wang and Zhang, 2017). Given that we consistently account for industry fixed effects in our regression analyses, any changes in competition within firms' output market are already captured. To evaluate the impact of China's import tariffs imposed on the US on the listed firms, specifically through the prices of imported inputs, we rely on their import composition prior to the commencement of the trade war:

$$\text{exposure to China's tariff}_{i,t} = \sum_j \frac{\text{import}_{i,j,14-16}}{\sum_j \text{import}_{i,j,14-16}} \text{tariff}_{j,t}^{CHN} \tag{4}$$

where $import_{i,j,14-16}$ is the amount of imports for firm i regarding 6-digit HS product j between 2014 and 2016. $\text{tariff}_{j,t}^{CHN}$ indicates the tariff rate that China imposed on the import of product j from the US during year t. By this formula, we measure the average tariff rates faced by the firm when importing from the US in year t based on its pre-trade-war import structure.

3 Empirical Analysis

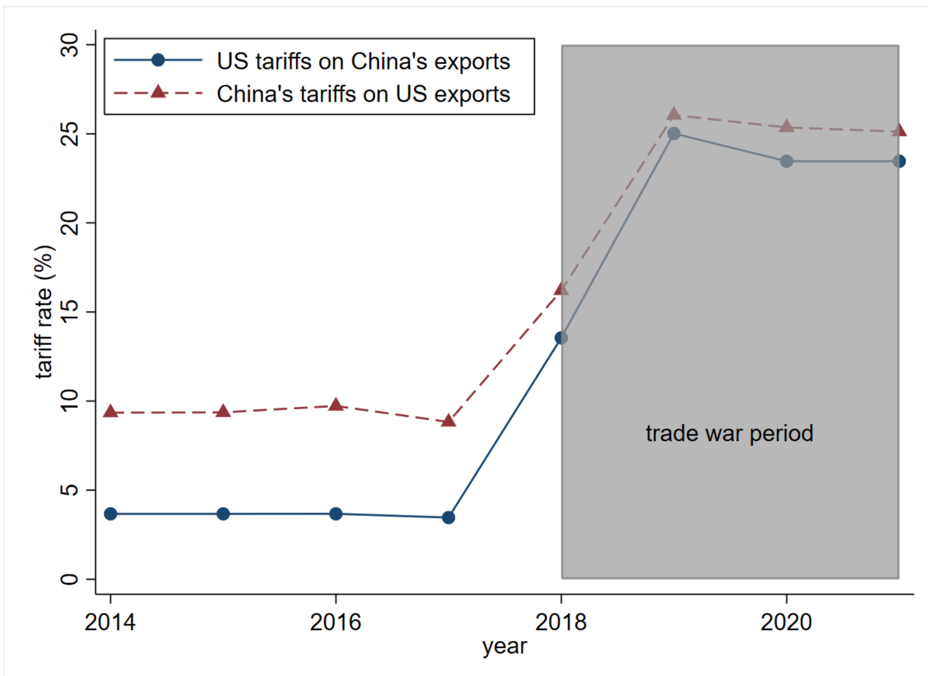
In this section, we explore how the trade war affected Chinese firms' innovation. We start by describing the evidence on tariff changes and aggregate patent similarity changes between Chinese and US patents. Then, we provide a formal empirical analysis.

3.1 First Glance at Data

Figure 1 displays the overall patterns in tariff rates during the trade war period, in line with previous studies (Fajgelbaum et al., 2019; Amiti et al., 2019; Fajgelbaum et al., 2023). Specifically, the US imposed a tariff increase of approximately 20 percentage points (averaged across 6-digit HS products) on China's exports, while China raised tariffs on US exports by around 15 percentage points.

Figure 2 illustrates the distribution of Chinese listed firms' exposure to tariffs imposed by the US and China, calculated by using Equations (3) and Equation (4). The data reveals a significant rise and substantial diversity in changes experienced during the trade war timeframe. This variability suggests the potential for varying impacts of the trade war on firms, a factor that will be explored in our empirical analysis.

Figure 1: Average Tariff Rates across 6-digit HS Products

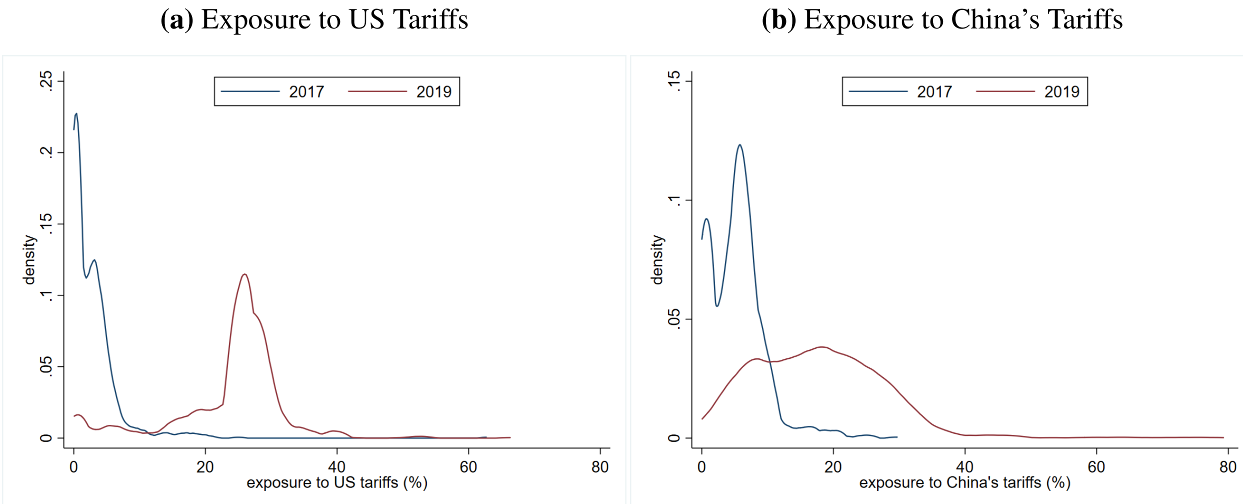


Notes: The figure displays the average US tariffs on China's export (solid blue curve) and the average China's tariff on US export (dashed red curve) from 2014 to 2021. Post-2016 data points are based upon the trading composition between 2014-2016 and the actual tariff rates across the 6-digit HS products. The shaded area is the trade-war period.

Figure 3 displays the aggregate similarity between Chinese and US patents in the ICT and Electrical Computers and Digital Processing Systems industries. We first separately identify the ICT patents and the Electrical Computers and Digital Processing Systems patents filed by the Chinese and US patent offices and then calculate the similarities between them.[4] For instance, the left panel of Figure 3 shows the similarities between Chinese and US ICT patents from 2000 to

[4]The ICT sector includes four subfields: telecommunications, consumer electronics, computers, office machinery, and other ICTs (OECD, 2007). The definition of Electrical Computers and Digital Processing Systems can be found at https://www.uspto.gov/web/offices/ac/ido/oeip/taf/reports.htm.

Figure 2: Distribution of Listed Firms' Exposure to US Tariffs



Notes: The figure shows the distribution of Chinese listed firms' exposure to tariffs on export to the US (Panel (a)) and tariffs on import from the US (Panel (b)). 2017 and 2019 are respective one year before and after the start of the trade war.

2021. Each pixel represents the similarity between Chinese patents filed in year t to US patents filed in year τ. Under the assumption that the US is the leading country in technology, we only visualize the similarity between Chinese patents that are filed in year t and the US patents that are filed before year t. Here, we sort the Chinese patents according to filing year in the rows and the US patents in the columns. The pixel in the southwest corner represents the similarity between Chinese patents filed in 2021 and US patents filed in 2000. The pixel in the southeast corner represents the similarity between Chinese patents filed in 2021 and US patents filed in 2021. The pixel in the diagonal represents the similarity between Chinese and US patents filed in the same year. Both panels depicted in Figure 3 demonstrate a decrease in the similarities between Chinese and US patents during the post-trade war period.

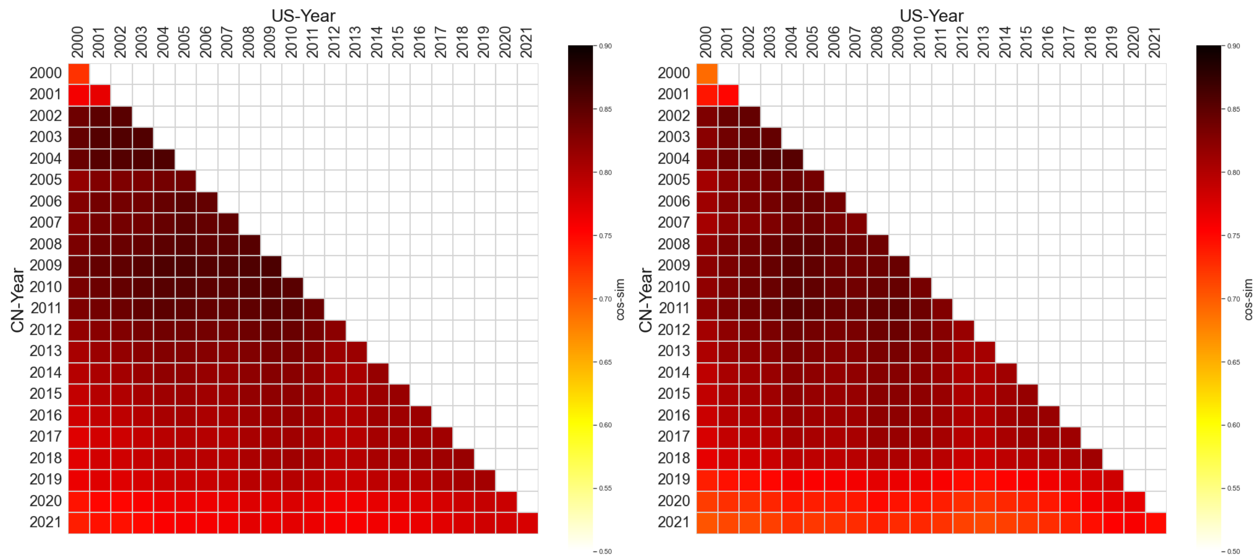
3.2 Identification Strategy

As shown in the previous section, the trade war largely increased the tariff imposed on China's exports to the US and China's imports from the US. In the same period, the industry-level similarity between Chinese and US patents stopped its previous rising trends. To identify the impact of the

Figure 3: Similarity between Chinese and US Patents

(a) ICT

(b) Electrical Computers and Digital Processing Systems



Notes: The figure visually presents the aggregate similarity between Chinese patents filed in a given year along the x-axis and corresponding US patents filed in the year along the y-axis. The degree of similarity is indicated by the darkness of each square, with darker shades denoting higher levels of similarity.

export and import tariff on China's innovation activities, we start with the following regression,

$$
\begin{aligned}
Y_{ist} =& \beta_1 \Delta ln(1+\text{tariff}_{it}^{E,US}) + \beta_2 \Delta ln(1+\text{tariff}_{it}^{I,US}) + \beta_3 \text{sanctioned}_{it} \\
&+ \gamma X_{it-1} + \phi Z_{it} + \alpha_i + \mu_t + \theta_{st} + \epsilon_{ist}.
\end{aligned}
\tag{5}
$$

where i indexes firms, s indexes the firms' 3-digit industry code, and t indexes years. The dependent variable, Y_{ist}, includes measures of the intensity and the direction of China's innovation. The natural logarithm of the one plus the number of patents applied by the firm is a proxy of the firm's innovation intensity. The firm-level similarity measure developed by this paper is a proxy for the firm's innovation direction. We use the demeaned average similarity to US patents granted in the recent 0-5 years (i.e., $\frac{1}{5}\sum_{\tau=t}^{t+5} \text{Sim}_{i,t,\tau}$) and assign value 0 to firms without any patent applications in the baseline regression. The trade-war related independent variables include the firm-specific export tariff change, $\Delta ln(1+\text{tariff}_{it}^{US})$, the firm-specific import tariff change, $\Delta ln(1+\text{tariff}_{it}^{CHN})$, and a dummy variable, sanction_{it}, indicating whether the firm is sanctioned at year t. The export and import tariff change is assigned the value 0 in 2014-2017 and is equal to the difference between the tariff rate firm i faces in year t and the firm's average tariff rate in 2014-2017. We control for firm-level characteristics in X_{it-1}, which include the natural logarithm of the firm's employment, total assets, and the profit share as a percentage of total revenue in year $t-1$. In the regression on innovation intensity, we include a dummy variable indicating whether the number of patent applications is positive to control for potential differences between firms that have patents and those that do not in a given year. In the regression on innovation direction, we add the number of patent applications as a control variable to ensure that the effect of the trade war on patent similarity is not driven by the quantity of patents. These contemporaneous variables are included in Z_{it}. The term α_i accounts for firm fixed effects, capturing unobserved firm-level heterogeneity. The term μ_t represents year fixed effects, reflecting time variation in the aggregate economy. The term θ_{st} denotes the industry-by-year fixed effects, capturing variations in industry-level characteristics over time.

The sample used for regression analysis includes all the publicly listed firms in China that has applied for at least one patent in 2000-2021. This restriction on patent applications ensures that the sample represents innovating firms. 75% of Chinese listed firms satisfies this restriction. The sample period is from 2014 to 2021, respectively four years before and after the trade war. We winsorize the top and bottom 1% of the innovation intensity, similarity, and the firm-level characteristics by year to avoid the impact of extreme values. The summary statistics of the major variables are displayed in Table 2. Exporting and non-exporting firms are presented separately to highlight their differing patterns. A firm is classified as an exporter if it has a positive export value in 2014-2017. On average, exporters have more patent applications, and their patents are more similar to those from other advanced economies. While non-exporters show a significant increase in patent similarity to other economies, the patent similarity of exporters to foreign countries either remains relatively unchanged or decreases. During this time, both export and import tariffs between China and the U.S. increased significantly. Approximately 1.2% of observations are sanctioned in 2018-2021.

Table 2: Summary Statistics

	2014-2017			2018-2021		
	mean	sd	count	mean	sd	count
Exporting Firms						
Patent Application Number	17.12095	40.25324	3795	20.93145	51.7699	5193
Similarity to US Patents (0-5 Years)	.7297993	.5811307	3795	.7548028	.5967714	5192
Similarity to EU Patents (0-5 Years)	.7632472	.6101537	3795	.7593173	.5940287	5193
Similarity to JP Patents (0-5 Years)	.7720983	.6105756	3795	.7650645	.6026168	5191
Similarity to KR Patents (0-5 Years)	.7426545	.5946121	3795	.7597006	.5806005	5192
Non-exporting Firms						
Patent Application Number	8.354087	29.92004	3903	13.89704	45.48336	5041
Similarity to US Patents (0-5 Years)	.4458641	.5943695	3902	.6289777	.6562887	5041
Similarity to EU Patents (0-5 Years)	.4562983	.6055783	3901	.6192423	.6308126	5040
Similarity to JP Patents (0-5 Years)	.4564408	.6053658	3902	.6280834	.6411932	5040
Similarity to KR Patents (0-5 Years)	.4577629	.6082342	3902	.6629596	.6691643	5041
Trade Shock						
Change in Export Tariff (China-US)	0	0	7698	.0536953	.088711	10234
Change in Import Tariff (China-US)	0	0	7698	.0287934	.0570207	10234
Sanctioned	0	0	7698	.0117256	.1076534	10234

Notes: This table reports the summary statistics of the main dependent and independent variables in the "pre" and "post" periods of the regression sample.

3.3 Impact of the Trade War on China's Innovation

The impact of the trade war on Chinese firms' innovation intensity and direction is reported in Table 3. Firm and year fixed effects are controlled in all columns. Columns (2) and (5) further control industry-specific year trends. Columns (3) and (6) instead further control industry-by-year fixed effects. The first three columns show that a 10 percent increase in the export tariff on Chinese goods leads to a decrease in the number of patent applications of Chinese firms by around 3.01 percent. The last three columns indicate that a 10 percent increase in the export tariff on Chinese goods decreases the similarity of Chinese patents to US patents by 2.04 percent of their historical average. The coefficient values do not vary significantly across different regression specifications, showing robustness.

The import tariff has an insignificant effect on firms' patent applications, while it significantly reduces patent similarity to the US. A 10 percent increase in the import tariff from the US decreases the similarity of Chinese patents to US patents by 3.040 percent of their historical average. The sanction dummy has a significant positive relationship with firms' patent applications, but weakly correlates to patent similarity to the US. The positive relationship can be explained by the targeting behavior of US sanctions—Chinese firms with a high growth potential in innovation activities are more likely to be sanctioned. The insignificant effect on innovation direction may be due to the limited time span of our observations, as sanctions mostly came with a buffer period. They may have larger influences in the long run. Since changes in export tariff have a significant impact on both China's innovation intensity and direction, this paper will focus on the mechanism behind the export shock in the following sections.

Table 3: Impact of the Trade War on Chinese Firms' Innovation Intensity and Direction

	Patent Application Number			Similarity to US Patent (0-5 Years)		
	(1)	(2)	(3)	(4)	(5)	(6)
Δ Export Tariff	-0.188	-0.298*	-0.301*	-0.236**	-0.186*	-0.204*
	(0.147)	(0.158)	(0.166)	(0.0973)	(0.102)	(0.106)
Δ Import Tariff	0.117	0.208	0.206	-0.309**	-0.324**	-0.340**
	(0.221)	(0.221)	(0.227)	(0.137)	(0.141)	(0.145)
Sanctioned	0.259***	0.221***	0.223***	-0.0338	-0.0218	-0.0174
	(0.0832)	(0.0794)	(0.0827)	(0.0533)	(0.0562)	(0.0586)
Firm Characteristics	Y	Y	Y	Y	Y	Y
Firm Fixed Effect	Y	Y	Y	Y	Y	Y
Year Fixed Effect	Y	Y	Y	Y	Y	Y
Industry × Year Trend	N	Y	N	N	Y	N
Industry × Year Fixed Effect	N	N	Y	N	N	Y
Observations	17,932	17,932	17,932	17,930	17,930	17,930
R-squared	0.877	0.880	0.882	0.544	0.548	0.560

Notes: Standard errors are clustered at the firm level. Δ denotes the difference between the tariff rates in a given year to the average rates in 2014-2017. Firm-level controls include the natural logarithm of the firm's employment, total assets, and the share of profit as a proportion of total revenue in the previous year. For regressions on patent applications, a dummy variable indicating whether the firm files for positive number of patents is controlled. For regressions on patent similarity, patent application number is controlled. Industries are defined at the 3-digit level. *** Significant at the 1 percent level; ** Significant at the 5 percent level; * Significant at the 10 percent level.

The significant impact of export tariffs on China's innovation activities underscores the effect of changes in demand. Weakened demand from the U.S. market may reduce the incentive for Chinese firms to imitate U.S. patents, as the benefits of gaining a cost advantage or improving quality relative to U.S. goods diminish. This effect is expected to be more pronounced for patents similar to recent U.S. patents because competition is most intense at the cutting edge. To test this hypothesis, we calculate the similarity between Chinese patents filed each year and U.S. patents filed in different time periods. Specifically, we compute the demeaned similarity of a Chinese firm's patents filed respectively in year t with all US patents filed in years t and $t-1$, $t-2$ and $t-3$, as well as $t-4$ and $t-5$. Similarity to more recent US. patents captures the extent to which a firm's innovation is aligned with the competitive fringe of other goods providers for the U.S. market. While similarity to older US patents is less influenced by competition for market demand.

The effect of the trade war on patent similarity to U.S. patents filed in different time periods is presented in Table 4. Chinese patents show a more significant divergence from US patents filed within the most recent three years. A 10 percent increase in export tariffs decreases the similarity to U.S. patents filed in the current and previous year by 2.13 percent and reduces the similarity to U.S. patents filed two to three years ago by 2.31 percent. In contrast, the decrease in similarity to older U.S. patents is smaller and nearly insignificant. This pattern is robust across different regression specifications and confirms that the demand shock has a more pronounced effect on China's patent similarity with the most recent technologies in the U.S.

The impact of changes in import tariffs does not significantly differ for similarity to U.S. patents filed in various periods. This suggests that the demand channel may not be a driving factor for the effects of import tariffs.

3.4 The Gradual Impact of the Export Tariff

The impact of the export tariff may not be unleashed immediately since it often takes time to make adjustments to innovation activities. To explore the persistence of the export tariff influence, we allow the coefficient of the tariff changes to vary over years using the following regression,

$$\begin{aligned} Y_{ist} = & \sum_{\tau=2018}^{2021} \beta_{1\tau} \mathbb{1}(t=\tau)\Delta ln(1+\text{tariff}_{it}^{E,US}) + \sum_{\tau=2018}^{2021} \beta_{2\tau} \mathbb{1}(t=\tau)\Delta ln(1+\text{tariff}_{it}^{I,US}) \\ & + \beta_3 \text{sanctioned}_{it} + \gamma X_{it-1} + \phi Z_{it} + \alpha_i + \mu_t + \theta_{st} + \epsilon_{ist}, \end{aligned} \tag{6}$$

Table 4: Impact of the Trade War on Chinese Patents' Similarity to US Patents by Filing Period

	Similarity to US Patent								
	Recent 0-1 Years			Recent 2-3 Years			Recent 4-5 Years		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Δ Export Tariff	-0.250***	-0.198**	-0.213**	-0.275***	-0.220**	-0.231**	-0.193*	-0.150	-0.171
	(0.0958)	(0.100)	(0.104)	(0.0987)	(0.104)	(0.108)	(0.101)	(0.105)	(0.109)
Δ Import Tariff	-0.319**	-0.326**	-0.352**	-0.284**	-0.302**	-0.316**	-0.306**	-0.324**	-0.334**
	(0.136)	(0.141)	(0.144)	(0.139)	(0.143)	(0.147)	(0.141)	(0.144)	(0.148)
Sanctioned	-0.0404	-0.0262	-0.0223	-0.0344	-0.0237	-0.0184	-0.0316	-0.0203	-0.0171
	(0.0536)	(0.0566)	(0.0593)	(0.0545)	(0.0571)	(0.0594)	(0.0546)	(0.0576)	(0.0598)
Firm Characteristics	Y	Y	Y	Y	Y	Y	Y	Y	Y
Firm Fixed Effect	Y	Y	Y	Y	Y	Y	Y	Y	Y
Year Fixed Effect	Y	Y	Y	Y	Y	Y	Y	Y	Y
Industry $\times$ Year Trend	N	Y	N	N	Y	N	N	Y	N
Industry $\times$ Year Fixed Effect	N	N	Y	N	N	Y	N	N	Y
Observations	17,924	17,924	17,924	17,925	17,925	17,925	17,918	17,918	17,918
R-squared	0.544	0.549	0.560	0.539	0.544	0.556	0.534	0.539	0.550

Notes: Standard errors are clustered at the firm level. Δ denotes the difference between the tariff rates in a given year and the average rates in 2014-2017. Firm-level controls include the natural logarithm of the firm's employment, total assets, and the share of profit as a proportion of total revenue in the previous year. For regressions on patent applications, a dummy variable indicating whether the firm files for positive number of patents is controlled. For regressions on patent similarity, patent application number is controlled. Industries are defined at the 3-digit level.
*** Significant at the 1 percent level; ** Significant at the 5 percent level; * Significant at the 10 percent level.

where $\mathbb{1}(t = \tau)$ is an indicator function of whether year t equals to τ. Figure 4 plots the coefficient, $\beta_{1\tau}$, and its 90 percent confidence interval in the regression above. The dependent variables are respectively the patent application number and the similarity to US patents filed in different time periods. In addition to the firm and year fixed effects, the industry-by-year fixed effects are controlled. Therefore, the aggregate effect over years correspond to the last column in Table 3 and columns (3) and (6) in Table 4.

Figure 4 shows that the change in export tariff on Chinese goods has a gradual and persistent impact on both the innovation intensity and direction in China. Panel (a) indicates the adverse consequences of the export tariff on patent applications of Chinese firms have intensified over time. Panel (b)-(d) demonstrates how the export tariff progressively influenced the alignment of Chinese firms' innovative activities with those in the US. The initial small impact grows larger in absolute magnitude and more significant over time, indicating the time required to realign research priorities. The long-lasting negative effect reflects an increasing divergence in the innovative endeavors between China and the US.

Figure 4: Effect on Innovation Intensity and Similarity to US Patents

(a) Effect on Patent Application Number

(b) Similarity to US in recent 0-5 Years

(c) Similarity to US in recent 0-1 Years

(d) Similarity to US in recent 2-3 Years

Notes: The figure shows the heterogeneous effect of the export tariff change on Chinese firms' innovation intensity and similarity to the US in each year after the trade war. Both the point estimate and the 90 percent confidence interval are presented. Standard errors are clustered at the firm level. Firm-level controls include the natural logarithm of the firm's employment, total assets, and the share of profit as a proportion of total revenue in the previous year. For regressions on patent applications, a dummy variable indicating whether the firm files for positive number of patents is controlled. For regressions on patent similarity, patent application number is controlled. Firm fixed effects, year fixed years, and industry-by-year fixed effects are controlled.

3.5 Robustness Checks

Three major concerns arise regarding the regression results presented in the previous sections. First, firms with higher growth potential in innovation intensity and greater similarity to US innovations may be more exposed to the trade shock. In other words, there may be heterogeneous pre-existing trends that could lead to endogeneity. Second, the decrease in similarity between Chinese and US patents might not indicate a divergence in innovation direction; instead, it could reflect strategic changes in the wording of patent abstracts to avoid tariff increases. Third, the observed impact on patent similarity may be influenced by the choice of the similarity measure.

To address the first concern, we conduct placebo tests using a sample from 2013 to 2017. The year 2013 is considered a "pre" period without any tariff changes. We assign the actual export and import tariff changes from 2018 to 2021 to the years 2014 to 2017, respectively, as counterfactual changes, and then replicate regression equation 6 using this sample. The point estimate and the 90 percentage intervals of the impact of the export tariff change on firms' innovation intensity and direction in each year are plotted in Appendix Figure A-3. The results show that the impact of the counterfactual export shocks is not significant, suggesting that there are no pre-existing trends.

The second concern regarding the potential strategic wording of patent abstracts can be evaluated by focusing exclusively on patents without Patent Cooperation Treaty (PCT) applications. A PCT application allows an applicant to seek patent protection in multiple PCT member countries through a single international filing. Both China and the US are members of the PCT. Applicants who file under the PCT have the option of requesting an International Preliminary Examination, which provides an early indication of the patentability of the invention in certain member countries before their patent offices conduct their own examinations. This preliminary examination can help guide strategic decisions about where to pursue patent protection. Given the cost efficiency and additional guidance provided by the PCT application process, it is often the preferred option for applicants seeking protection outside their home country. Consequently, patents with PCT applications are more likely to be intended for international publication and are more inclined to strategically adjust their wording to align with the requirements of patent offices in other countries. By recalculating patent similarity using only a firm's patents without PCT applications and running the regression equation 4, we can derive a measure of patent similarity change that is less influenced by strategic wording behaviors. Columns (1)–(4) in Appendix Table A-1 show that the impact of increased export and import tariffs is very similar to the baseline results, confirming the divergence of China's innovation activities from those of the U.S.

Regarding the third concern about the robustness of the baseline results to different measures of patent similarity, we use a state-of-the-art natural language processing (NLP) model, Embedding Model E5, to calculate similarity. Unlike the TF-IDF method, which captures the importance of words based on their frequency in a document, Embedding Model E5 captures deep semantic information, including the meaning of words, phrases, and sentences in context, utilizing the transformer architecture.[5] Columns (5)-(8) in Appendix Table A-1 display the impact of the trade shock on patent similarity as calculated by Embedding Model E5. Although this new similarity measure is based on a very different algorithm, the magnitudes of the trade shock coefficients are very close to those of the baseline results. This indicates that the baseline results are not solely driven by the choice of similarity measurement.

3.6 China's Innovation Similarity to Other Countries

Has the similarity in innovation between China and other countries changed following the increase in export tariff on Chinese goods by the US government? How does the similarity of Chinese patents with those of other countries relate to similarity between China and the U.S.? To answer these questions, our analysis investigates the impact of the US-China trade war on the similarity of Chinese patents with those from other high-innovation economies. Notably, in addition to the US and China, Europe, Japan, and South Korea are prominent regions with significant patent activities. Therefore, this study evaluates the similarity of China's patent output with these regions through the following regression analysis,

$$\begin{aligned} Y^{*}_{ist} =& \beta_1 \Delta ln(1 + \text{tariff}_i^{E,US}) + \beta_2 \Delta ln(1 + \text{tariff}_i^{I,US}) + \beta_3 \text{sanctioned}_i \\ &+ \beta_4 \Delta ln(1 + \text{tariff}_i^{E,*}) + \beta_5 \Delta ln(1 + \text{tariff}_i^{I,*}) + \gamma X_{ist-1} + \phi Z_{it} + \alpha_i + \mu_t + \theta_{st} + \epsilon_{is}. \end{aligned} \quad (7)$$

where $* \in \{EU, JP, KR\}$. The dependent variable (Y^{*}_{ist}) represents the similarity of Chinese firms' patents to those in Europe, Japan, and South Korea filed in different periods (namely, the past 0-5 years, 0-1 years, 2-3 years, and 4-5 years). In addition to the control variables specified in the baseline regression (Equation 5), this analysis includes changes in export and import tariffs in

[5]A more detailed description of Embedding Model E5 is presented in Appendix A.

the respective regions. The effect of the increasing export tariff on Chinese goods is captured by the value of β_1. To further investigate how the effect depends on the patent similarity between China and the U.S., we compare the value of β_1 when the regression excludes and includes similarity between Chinese firms' patents and U.S. patents filed in the corresponding periods in Z_{it}.

Table 5 presents the results for the three other economies. Columns (1)-(4) do not control for the similarity between Chinese patents and US patents filed in the corresponding periods, while columns (5)-(8) do. The results indicate that the similarity of Chinese patents to those from other regions decreases to varying extents. A 10 percent increase in the US export tariff leads to a 2.03 percent decrease from the historical average in the similarity of Chinese patents to European patents filed over the past five years, as shown in column (1). This effect is comparable to that observed for U.S. patents, which see a decrease of 2.04 percent. The negative impact on the similarity to Japanese patents is slightly smaller at 1.85 percent, while the effect on South Korean patents is even smaller and statistically insignificant. Columns (5)-(8) reveal that nearly all the impact of the export shock disappears after controlling for similarity to US patents, suggesting that the divergence in China's innovation activities from the U.S. is the primary reason for the divergence from other regions. In other words, Chinese patents do not further deviate from other regions once the distance between Chinese and U.S. patents is accounted for. Additionally, the magnitude of the export shock's effect on patent similarity with other regions is positively correlated with the extent to which this similarity can be explained by the degree of similarity to U.S. patents.

The varying degrees of change in patent similarity across different regions highlight the multifaceted nature of innovation activities. While patent quantity offers a basic measure of innovation output, a deeper understanding is gained through textual analysis of patent contents, which provides richer insights into the specific directions of innovation.

3.7 Heterogeneity in Trade Shock Impact

Do Chinese firms' innovations deviate more from the most innovative firms in the US? Are more productive firms in China more sensitive to the export tariff shocks?

Table 5: Impact of the Export Tariff on Chinese Patents' Similarity to other Regions

VARIABLES	Similarity to Patents in other Regions							
	0-5 Years	0-1 Years	2-3 Years	4-5 Years	0-5 Years	0-1 Years	2-3 Years	4-5 Years
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Europe								
Δ Export Tariff	-0.203*	-0.192*	-0.212*	-0.222**	-0.0283	-0.00694	-0.0180	-0.0892*
	(0.106)	(0.108)	(0.109)	(0.106)	(0.0496)	(0.0537)	(0.0544)	(0.0528)
Similarity to US					0.868***	0.861***	0.849***	0.838***
					(0.00873)	(0.00929)	(0.00888)	(0.00896)
Japan								
Δ Export Tariff	-0.185*	-0.166	-0.209*	-0.194*	-0.00666	0.0219	-0.0143	-0.0502
	(0.106)	(0.107)	(0.106)	(0.108)	(0.0556)	(0.0575)	(0.0585)	(0.0609)
Similarity to US					0.848***	0.838***	0.826***	0.819***
					(0.00948)	(0.00979)	(0.00972)	(0.00986)
South Korea								
Δ Export Tariff	-0.128	-0.130	-0.138	-0.126	0.0286	0.0352	0.0322	0.0427
	(0.105)	(0.106)	(0.106)	(0.108)	(0.0530)	(0.0570)	(0.0566)	(0.0610)
Similarity to US					0.831***	0.829***	0.806***	0.800***
					(0.00928)	(0.00949)	(0.00963)	(0.0100)
Firm Characteristics	Y	Y	Y	Y	Y	Y	Y	Y
Firm Fixed Effect	Y	Y	Y	Y	Y	Y	Y	Y
Year Fixed Effect	Y	Y	Y	Y	Y	Y	Y	Y
Industry × Year Fixed Effect	Y	Y	Y	Y	Y	Y	Y	Y

Notes: Standard errors are clustered at the firm level. Δ denotes the difference between the tariff rates in a given year and the average rates in 2014-2017. Firm-level controls include the natural logarithm of the firm's employment, total assets, and the share of profit as a proportion of total revenue in the previous year. The number of Patent applications is also controlled. Columns (1)-(4) do not control similarity between the firm's patents and the US patents as a whole filed in the corresponding periods, while columns (5)-(8) do. Industries are defined at the 3-digit level.
*** Significant at the 1 percent level; ** Significant at the 5 percent level; * Significant at the 10 percent level.

To answer the first question, we calculate the patent similarity between Chinese firms and the most innovative US firms in each IPC, which is defined as the top 20 firms with the highest number of patent applications in that IPC. Firm-level similarity is still the weighted average of the similarity in each IPC. The results are presented in columns (1)-(4). Compared to the baseline coefficient values in Table 3 and Table 4, changes in the export tariff have a larger impact on patent similarity, showing that innovation activities of Chinese firms deviate more from the most innovative firms in the US.

To answer the second question, we classify Chinese firms into two groups based on their average TFP from 2014 to 2017, the period before the trade war. High-TFP firms are defined as those with above-median TFP within their industry. Columns (5)-(10) illustrate the trade war's impact on innovation intensity and direction among these TFP groups. High-TFP firms experience

Table 6: Heterogeneous Impact of the Trade War by Innovativeness and TFP

	Similarity to Top US Firms				Innovation by TFP of Chinese Firms					
	0-5 Years	0-1 Years	2-3 Years	4-5 Years	Patent Number		All US Firms (0-5 Years)		Top US Firms (0-5 Years)	
					High TFP	Low TFP	High TFP	Low TFP	High TFP	Low TFP
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Δ Export Tariff	-0.290**	-0.322***	-0.338***	-0.205	-0.298	-0.271	-0.434***	-0.0542	-0.562***	-0.0998
	(0.118)	(0.120)	(0.122)	(0.126)	(0.235)	(0.244)	(0.148)	(0.156)	(0.156)	(0.183)
Δ Import Tariff	-0.346**	-0.254	-0.336**	-0.361**	0.599*	-0.203	-0.422**	-0.300	-0.501**	-0.241
	(0.164)	(0.168)	(0.169)	(0.175)	(0.318)	(0.336)	(0.190)	(0.228)	(0.221)	(0.258)
Sanctioned	-0.00419	-0.0129	-0.00778	-0.00567	0.201*	0.269**	-0.0658	0.0571	-0.0857	0.111
	(0.0802)	(0.0816)	(0.0824)	(0.0795)	(0.111)	(0.120)	(0.0612)	(0.116)	(0.0732)	(0.176)
Firm Characteristics	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Firm Fixed Effect	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Year Fixed Effect	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Industry × Year	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Observations	17,885	17,800	17,822	17,808	8,625	9,159	8,624	9,158	8,606	9,131
R-squared	0.527	0.517	0.523	0.515	0.891	0.872	0.574	0.569	0.548	0.531

Notes: Standard errors are clustered at the firm level. Δ denotes the difference between the tariff rates in a given year to the average rates in 2014-2017. Columns (1)-(4) explores the impact of the trade war on patent similarity between Chinese firms and the most innovative US firms in each IPC, defined as the top 20 firms with the highest number of patent applications in that IPC. Columns (5)-(10) explore the impact of the trade war on Chinese firms' innovation intensity and directions by their average TFP level in 2014-2017. The dependent variables are, respectively, patent numbers in columns (5)-(6), similarity to all US firms in columns (7)-(8), and similarity to the most innovative US firms in each IPC in columns (9)-(10). Firm-level controls include the natural logarithm of the firm's employment, total assets, and the share of profit as a proportion of total revenue in the previous year. Industries are defined at the 3-digit level.
*** Significant at the 1 percent level; ** Significant at the 5 percent level; * Significant at the 10 percent level.

a slightly larger decrease in the number of patent applications due to export tariff shocks, as shown in columns (5) and (6). However, their patent similarity to US firms declines significantly more than that of low-TFP firms, as seen in columns (7) and (8). Moreover, columns (9) and (10) demonstrate that innovation activities in high-TFP Chinese firms deviate even further from those of the most innovative US firms, indicating that the growing divergence in innovation direction between the US and China is largely driven by the most innovative or productive firms in both countries.

4 Quantitative Model

We now develop a dynamic model to illustrate the mechanisms behind the impact of tariff rates on firms' innovation in China. We focus on understanding Chinese listed firms' decisions regarding innovation and exporting. We take the wage rate and destination market conditions as given.[6]

4.1 Model Setup

Preferences and Market Demand. We consider a world with many destination markets indexed by $n = 0, 1, 2, .., N$, where $n = 0$ represents the domestic market. There are a set of $\mathcal{I}$ products that can be potentially produced, and each firm ω can produce a subset of products $\mathcal{I}(\omega) \subset \mathcal{I}$. Each product i has a set of features $\mathcal{K}_i$ (e.g., engines and air-conditioning for a car). Firms producing the same product can differ in the product's features (e.g., distinct car models), and we treat each firm's product as a variety. We assume that within a product market, different varieties are engaged in monopolistic competition.

Consumers in each destination n have the following preferences:

$$U_t^n = \prod_{i \in \mathcal{I}} (Q_{it}^n)^{\gamma_i^n},$$

$$Q_{it}^n = \left[\sum_{\omega} (q_{it}^n(\omega))^{\frac{\sigma-1}{\sigma}} \right]^{\frac{\sigma}{\sigma-1}}, \quad \text{where } q_{it}^n(\omega) = \left(\sum_{k \in \mathcal{K}_i} (\gamma_{ik}^n)^{\frac{1}{\epsilon}} q_{ikt}^n(\omega)^{\frac{\epsilon-1}{\epsilon}} \right)^{\frac{\epsilon}{\epsilon-1}}.$$

where the upper-level preferences are Cobb-Douglas preferences over product-level composite goods Q_{it}^n, with γ_i^n governing the expenditure share and $\sum_i \gamma_i^n = 1$. Within each product, consumers have a nested CES preference over different varieties with the elasticity of substitution $\sigma > 1$. Under monopolistic competition, we compute the demand for a variety produced by firm ω as given by $q_{it}^n(\omega) = (p_{it}^n(\omega))^{-\sigma} (P_{it}^n)^{\sigma-1} \gamma_i^n E_t^n$, where $p_{it}^n(\omega)$ is the price charged by firm ω for each unit of $q_{it}^n(\omega)$. P_{it}^n is the aggregate price index of the composite good of product i in country n, and E_t^n is country n's total expenditure. Each firm's product is a bundle of different features,

[6]Despite China being viewed as a "world factory," the share of foreign manufacturing expenses spent on Chinese goods was only around 5% in 2015 (which reflects cross-border trade barriers), according to OECD Inter-Country Input-Output Table. Thus, we assume that changes in China have minimal impacts on aggregate conditions in other countries.

with variable $q^n_{ikt}(\omega)$ denoting the quantity level of feature k offered by firm ω. Parameter γ^n_{ik} captures the taste of consumers from country n for feature k of product i: for example, US consumers usually prefer SUVs to sedans, while the opposite is true for Chinese consumers. The parameter $\epsilon > 1$ is the elasticity of substitution between different features of a variety.

Firms' Production and Trade Costs. If a firm is endowed with the technology for product $i \in \mathcal{I}(\omega)$, it produces different features of product i using the following equation:

$$q_{ikt}(\omega) = z_{ikt}(\omega)^{\frac{1}{\epsilon-1}} l_{ikt}(\omega), \quad k \in \mathcal{K}_i.$$

$z_{ikt}(\omega)$ is feature-specific productivity level, and $l_{ikt}(\omega)$ captures the amount of labor hired in producing feature k. We introduce the exponent $\frac{1}{\epsilon-1}$ on $z_{ikt}(\omega)$ to simplify the derivation that follows. Additionally, it is noteworthy that in the special case where there is only one feature ($\#M_i = 1$) with elasticity $\epsilon = \sigma$, revenue becomes proportional to productivity, $p^n_{it}q^n_{it} \propto z^n_{it}$, which aligns with a common assumption in the growth literature (e.g., Akcigit, Celik and Greenwood, 2016).[7]

Given the production function, the firm will minimize the cost of producing each unit of $q^n_{it}(\omega)$ (after accounting for consumers' preferences toward different features in market n), and thus we can compute the marginal cost of $q^n_{it}(\omega)$ as:

$$c^n_{it}(\omega) = \left[\sum_{k \in \mathcal{K}_i} \gamma^n_{ik} z_{ikt}(\omega)\right]^{\frac{1}{1-\epsilon}} w_t.$$

To serve market n, a firm needs to pay iceberg costs $\tau^n_i \geq 1$. Moreover, exporting a certain product incurs fixed export costs f^n_i (Melitz, 2003) in units of labor, with the costs of the local market being $\tau^0_i = 1$ and $f^0_i = 0$.

[7] In the following derivation, we assume that innovation enhances the level of $z_{ikt}(\omega)$. If, instead, we assume that innovation directly improves the level of $z_{ikt}(\omega)^{\frac{1}{\epsilon-1}}$, this would result in excessively high innovation returns (since revenues are elastic with respect to $z_{ikt}(\omega)^{\frac{1}{\epsilon-1}}$) and could lead to the risk of explosive solutions.

Firms' Productivity and Innovation. We assume that firms' productivity levels evolve over time as follows:

$$\underbrace{z_{ik,t+1}(\omega)}_{\text{next-period feature-specific productivity}} = \underbrace{(1-\delta)z_{ikt}(\omega)}_{\text{current-period productivity}} + \underbrace{[s_{ikt}(\omega)a_{it}(\omega)]^{\phi}}_{\text{increment from innovation}} .$$

The term $a_{it}(\omega)$ represents the number of inventions by firm ω related to product i at time t, while $s_{ikt}(\omega)$ denotes the share of innovation directed towards feature k, with the condition that $\sum_{k\in\mathcal{K}_i} s_{ikt}(\omega) = 1$. The existing literature on directed technology change, such as the works by Acemoglu (2010) and Acemoglu, Aghion, Bursztyn and Hemous (2012), examines factor-biased or sector-biased technological change. Our study broadens this scope by exploring how firms can allocate their innovation efforts across various features of a product. We follow the literature (e.g., Bloom, Romer, Terry and Van Reenen, 2020) to assume that innovation efforts exhibit the diminishing returns, with $0 < \phi < 1$. Incurring an invention would cost ψ units of labor.

4.2 Solving Firm's Problem

Static Problem: Choosing Export Price and Status. Given productivity levels, we first solve a firm ω's optimal prices and export status in each time t. As production function exhibits constant returns to scale, the export decisions are independent across destinations. The firm chooses the price to maximize variable profits for each market n:

$$\max_{\{\mathbf{1}^n_{it}(\omega),p\}} \pi^n_{it}(\omega) = \mathbf{1}^n_{it}(\omega)\left[\left(p - \tau^n_i c^n_{it}(\omega)\right) p^{-\sigma} \left(P^n_{it}\right)^{\sigma-1}\gamma^n_i E^n_t - w_t f^n_i\right] .$$

where $\mathbf{1}^n_{it}(\omega) \in \{0,1\}$ is a dummy variable indicating whether to export product i to market n. We can solve price $p^n_{it}(\omega) = \tilde{\sigma}\tau^n_i c^n_{it}(\omega)$ if the firm exports to market n, where $\tilde{\sigma} = \frac{\sigma}{\sigma-1}$. The corresponding profits are given by:

$$\pi_{it}^n(\omega) = \mathbf{1}_{it}^n(\omega) \left[\frac{1}{\sigma} \left(\tilde{\sigma} \tau_i^n c_{it}^n(\omega) \right)^{1-\sigma} \left(P_{it}^n \right)^{\sigma-1} \gamma_i^n E_t^n - w_t f_i^n \right] = \mathbf{1}_{it}^n(\omega) \left[\zeta_{it}^n c_{it}^n(\omega)^{1-\sigma} - w_t f_i^n \right]. \tag{8}$$

where $\zeta_{it}^n = \frac{1}{\sigma} \left(\tilde{\sigma} \tau_i^n \right)^{1-\sigma} \left(P_{it}^n \right)^{\sigma-1} \gamma_i^n E_t^n$ represents aggregate demand factor from market n for product i, which we take as given in our quantitative analysis. The firm exports to destination n ($\mathbf{1}_{it}^n(\omega) = 1$) if and only if $\zeta_{it}^n c_{it}^n(\omega)^{1-\sigma} \geq w_t f_i^n$.

Dynamic Problem: Innovation Choices. We can then solve the firm's innovation choices to maximize the value for product i:

$$\begin{aligned} V_{it}(\mathbf{z}_{it}(\omega)) = \max_{\{s_{ikt}(\omega), a_{it}(\omega)\}} & \sum_{n=0}^{N} \pi_{it}^n(\omega) - \psi w_t a_{it}(\omega) + \frac{1}{1+r} V_{it+1}(\mathbf{z}_{it+1}(\omega)) \\ \text{s.t. } & z_{ik,t+1}(\omega) = (1-\delta) z_{ikt}(\omega) + [s_{ikt}(\omega) a_{it}(\omega)]^{\phi} \\ & \sum_{k \in \mathcal{K}_i} s_{ikt}(\omega) = 1. \end{aligned}$$

The first-order conditions regarding the share of inventions devoted to feature k, $s_{ikt}(\omega)$, imply:

$$\frac{s_{ik't}(\omega)}{s_{ikt}(\omega)} = \left[\frac{\partial V_{it}(\mathbf{z}_{it}(\omega)) / \partial z_{ik't+1}(\omega)}{\partial V_{it}(\mathbf{z}_{it}(\omega)) / \partial z_{ikt+1}(\omega)} \right]^{\frac{1}{1-\phi}}. \tag{9}$$

Here, $\partial V_{it}(\mathbf{z}_{it}(\omega)) / \partial z_{ik't+1}(\omega)$ captures the marginal return from improving the productivity of feature k:

$$\frac{\partial V_{it}(\mathbf{z}_{it}(\omega))}{\partial z_{ikt+1}(\omega)} = \sum_{x=t+1}^{\infty} \left(\frac{1-\delta}{1+r} \right)^{x-t} \frac{1-\sigma}{(1-\epsilon)(1-\delta)} w_x^{1-\epsilon} \sum_{n=0}^{N} \mathbf{1}_{ix}^n(\omega) \zeta_{ix}^n (c_{ix}^n(\omega))^{\epsilon-\sigma} \gamma_{ik}^n.$$

The reliance of marginal benefits on $\sum_{n=0}^{N} \mathbf{1}_{ix}^n(\omega) \zeta_{ix}^n (c_{ix}^n(\omega))^{\epsilon-\sigma} \gamma_{ik}^n$ indicates that demand for a particular feature k, as captured by the weighted average destination's taste for a certain feature γ_{ik}^n, would affect the firm's proportion of innovation focused on that feature. Consequently, if foreign consumers prefer feature k and the firm serve these consumers (higher γ_{ik}^n for some foreign market

n with $\mathbf{1}^n_{ix}(\omega) = 1$), this optimal scenario suggests that the firm would allocate more effort towards feature k across all its innovations.

Using the first-order condition with regard to innovation quantity $a_{it}(\omega)$, we can also obtain the following solutions for the number of inventions:

$$a_{it}(\omega) = \left(\frac{\sum_{k \in \mathcal{K}_i} \phi(s_{ikt}(\omega))^{\phi} \partial V_{it}(\mathbf{z}_{it}) / \partial z_{ikt+1}}{w_t \psi} \right)^{\frac{1}{1-\phi}}. \tag{10}$$

Now, consider the impact of permanently higher tariff rates in market n, which raise iceberg costs τ_n, thereby reducing export revenues ζ^n_{ix} ($x \geq t$) for all future periods. According to equation (9), if the firm is actively producing good i and exporting to market n, the decline in export revenues ζ^n_{ix} will shift the focus of innovations in product i away from the preferences of consumers in country n (γ^n_{ik}). Additionally, equation (10) suggests that the lower export revenues ζ^n_{ix} will also decrease the total quantity of innovation, $a_{it}(\omega)$, assuming that the firm exports to market n.

Finally, suppose that different features correspond to different words in the text. We can thus compute the similarity between firm ω's innovation vector $\mathbf{a}_{it}(\omega) = [a_{ikt}]$ and another vector of innovations characterized by the vector $\mathbf{b}_{it} = [b_{ikt}]$ in product i across features:

$$Sim(\mathbf{a}_{it}(\omega), \mathbf{b}_{it}) = \frac{\sum_{k \in \mathcal{K}_i} a_{ikt}(\omega) b_{ikt}}{\left[\sum_{k \in \mathcal{K}_i} (a_{ikt}(\omega))^2\right]^{1/2} \left[\sum_{k \in \mathcal{K}_i} (b_{ikt})^2\right]^{1/2}}. \tag{11}$$

5 Quantitative Analysis

To evaluate the significance of innovation intensity and trajectories in assessing the impact of trade shocks on firm performance, we conduct a quantitative analysis using the U.S.-China trade war as a case study. In this section, we first calibrate the model using 2016 data, and then simulate the effects of the trade war.

Because our primary focus is on the impact of trade shocks on firms' innovation direction choices and their aggregate implications, we concentrate on innovative listed firms in this section.

These firms have filed at least one patent between 2000 and 2016 and have operations both before and after the trade war. This restriction allows us to link the model moments to the data, resulting in a sample of 2,185 innovative firms. Given our focus on this specific subset of firms, we will abstract from general equilibrium responses of aggregate prices and quantities, and the trade war decreases the export demand for Chinese firms due to increased iceberg costs.

5.1 Calibration

We now describe the calibration of the model to the data. The countries in the following analysis are restricted to China and the US. Thus, $n \in \{0, 1\}$, where $n = 0$ refers to China, and $n = 1$ refers to the US.

5.1.1 Methodology

Parameters in the model are categorized into three groups. The first group is calibrated by a prior information from the aggregate statistics or the literature. The second group is calibrated by direct estimation from the micro-level data. The third group is jointly calibrated through the model.

Table 7 reports the parameters in the first group, $(\mathcal{I}, \mathcal{K}, \sigma, \epsilon, \delta, \phi)$. The number of products, $\mathcal{I}$, is set to 120 to match the total number of 3-digit IPC codes in our sample. The number of features, $\mathcal{I}$, is set to 256 to match the total number of features. In our empirical analysis, the similarity measure is constructed based on word-level features, weighted by the TF-IDF index. However, the high dimensionality of the word-level features (exceeding 10,000) due to the vast number of words in patent texts poses challenges for simulation. To simplify the analysis, we apply Non-negative Matrix Factorization (NMF) to reduce the dimensionality of the data. This involves clustering words with similar meanings into 256 features.[8] The elasticity of substitution over different varieties, σ, and the elasticity of substitution over different features both take the

[8] A detailed description of the NMF method is provided in Appendix E. Running the regression of trade shocks on patent similarity (Equation 5) using these low-dimensional embeddings yields very similar results comparable to those in Tables 3 and 4.

value of 5, following Head and Mayer (2014). The depreciation rate of technology, σ, is 0.08, consistent with Holmes, McGrattan and Prescott (2015). The elasticity of innovation output with regard to cost, ϕ, is set to be 0.5, which is commonly used by the growth literature (e.g., Bloom, Jones, Van Reenen and Webb (2020)) and based on empirical findings.

Table 7: Parameter Values from a Priori Information

Notation	Definition	Value	Source
$\mathcal{I}$	Number of products	120	Data
$\mathcal{K}$	Number of features	256	Data
σ	Elasticity of substitution over different varieties	5	Head and Mayer (2014)
ϵ	Elasticity of substitution over different features	5	Head and Mayer (2014)
δ	Depriation rate of technology	0.08	Holmes et al. (2015)
ϕ	Decreasing return rate of innovation	0.5	Bloom et al.(2020)

Notes: This table shows parameter values derived from the aggregate statistics or the literature.

Parameters in the second group, $(\gamma_{ik}^n, z_{ik,2016})$, are directly estimated from the data. Since γ_{ik}^n captures the taste of consumers from country n for feature k of product i, we directly use each country's IPC-feature NMF matrix index between 2000 and 2016, which is unaffected by the trade war, to calibrate γ_{ik}^n.[9] $z_{ik,2016}$ is firm i' initial productivity of feature k in product i. It is pinned down by the accumulated patents number of each firm in 2016 at the IPC-feature level.

The third group of parameters, (ζ_i^n, f_i^n, ϕ), is jointly calibrated within the model by minimizing the distance between moments in the model and the data. The parameter ζ_i^n represents the aggregate demand from market n for product i. Since it influences firms' sales in market n, we use the average sales of firms selling product i to market n in 2016 to calibrate ζ_i^n. There are a total of $N \times \mathcal{K} = 240$ moments. The fixed export cost, $f_i^n, n = 1$, determines the extensive margin of exporting, and is therefore calibrated using the share of exporters among firms producing product i in 2016. The parameter ϕ, which governs average innovation costs, is calibrated based on the average number of patent applications between 2016 and 2021. Table 8 compares the moments

[9] For each IPC, we concatenate the text of patents applied between 2000 and 2016. Then, a TF-IDF matrix is constructed based on the words in this concatenated text. Finally, we apply Non-negative Matrix Factorization (NMF) to reduce the dimensionality of the TF-IDF matrix to 256 features.

generated by the model with their empirical counterparts. Due to space limitations, we report only the average moments across all products for firm sales and exporter shares. The model aligns well with the data.

Table 8: Parameters from the Minimum Distance Estimation

Parameter	Moment	Data	Model
ζ_i^0	Avg domestic sales for each product (billion)	544.78	544.78
ζ_i^1	Avg exporting sales for each product (billion)	24.78	24.61
f_i^1	Avg share of exporters among all firms producing each product	0.42	0.42
ψ	Avg number of patent applications for each firm	26.3	26.3

Notes: Parameters in this table are jointly calibrated to minimize the distance between the model and data moments. The value of the moments in both the data and model are presented.

5.2 Simulation

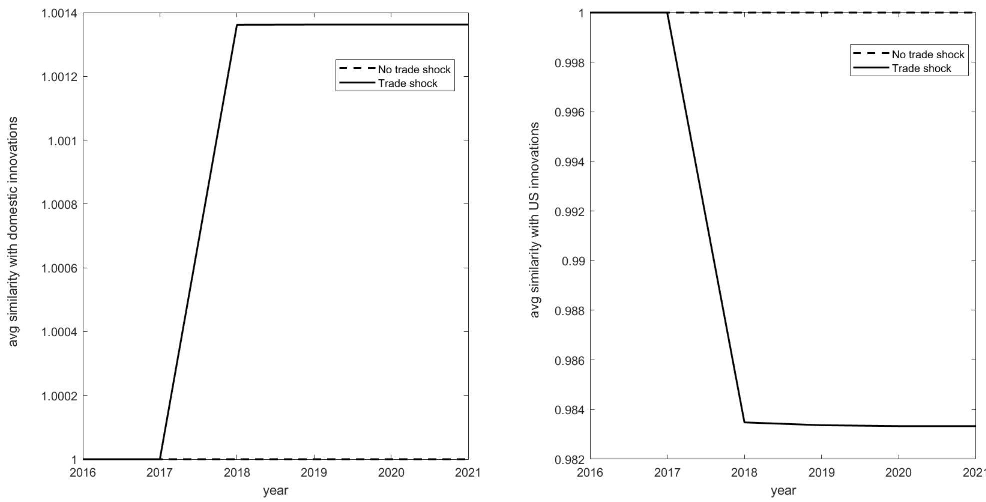
How effectively does the demand-side channel in the model capture changes in innovation trajectories resulting from the trade war? To what extent do export tariffs impact firm performance, especially export sales, through the innovation channel?

To answer these questions, we simulate the calibrated model from 2016 to 2021. First, the simulation is conducted without any tariff shocks to get the baseline trajectory of firms' decisions. Second, unexpected increases in the export tariff rate as in the data from 2018 to 2021 are introduced in the simulation process. Then we compare the two simulation results.

It is worth noting that firms' innovation choices, $a_{it}(\omega)$ and $s_{ikt}(\omega)$, depend on their future export status, $\mathbf{1}_{ix}^{n}(\omega)$, for $x = t+1, t+2, \ldots$ (see Equations 10 and 9). In turn, the decision regarding export status is influenced by innovation choices through their impact on productivity (Equation 8). Therefore, we iterate over the time paths of innovation and export decisions for each firm until the results converge. Additionally, we assume that export tariff rates after 2022 remain constant at their 2021 levels.

Figure 5 shows the similarity of patents filed by Chinese firms to domestic and U.S. innovations over time. The solid and dashed lines represent the simulation results with and without export

Figure 5: Patent Similarity to Domestic and US Innovations



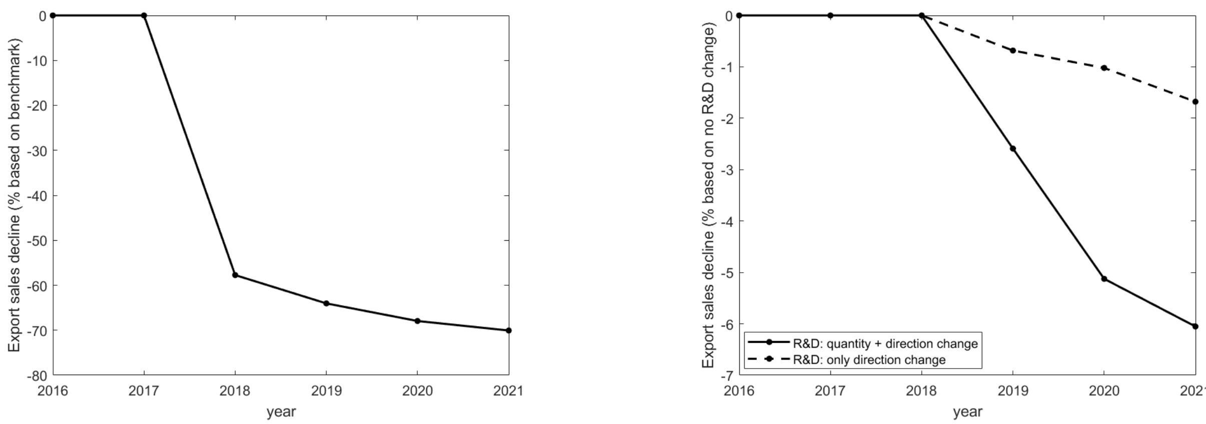
Notes: The figure illustrates the similarity of Chinese patents to domestic innovations (the left panel) and to U.S. innovations (the right panel), using 2016 as the base year. The dashed curve represents the simulation results without unexpected export tariff shocks, while the solid curve represents the results with the shocks.

shocks, respectively. The simulations reveal an increase in similarity to domestic innovations and a decline in similarity to U.S. innovations. Specifically, a 10 percent increase in export tariffs reduces the similarity to U.S. innovations by 0.98 percent relative to the historical average. Compared to the 2.04 percent decline observed in the empirical results (Table 3), the demand channel captured by the model explains 48 percent of the decrease in similarity to U.S. innovations.

The simulation with the actual export tariff shocks also reveal that the export tariff increase results in a 70% decline in export sales to the US up until 2021, as shown by the left panel of Figure 6. To study the role of innovation in this impact, we conduct two counterfactual analyses. First, we shut down the change in both innovation intensity ($a_{it}(\omega)$) and direction ($s_{ikt}, k \in \mathcal{K}$) and simulate the economy with export shocks to isolate the effect of R&D choices. Second, we shut down the change in only innovation direction ($s_{ikt}, k \in \mathcal{K}$) and simulate the economy with export shocks to isolate the effect of R&D directions. Comparing the results of these counterfactual simulations with the original simulation under export tariff shocks allows us to quantify the effect of innovation choices, as shown in the right panel of Figure 6.

Compared to a scenario where firms maintain their R&D behavior in both quantity and direction during the trade war happen, Changes in R&D result in a 6% decline in export sales to the U.S. by 2021. Of this decline, 28% is attributable to changes in R&D direction, leading to a 1.68% reduction in Chinese firms' export sales. Therefore, innovation trajectory has a non-trivial effect on firms' performance.

Figure 6: Changes in Export Sales due to Innovation Decisions



Notes: The figure shows the change in Chinese firms' export sales due to export tariff shocks (the left panel) and the impact of firms' innovation choices (the right panel), with 2016 as the base year. In the right panel, the solid curve represents the combined contribution of both innovation intensity and direction, while the dashed curve isolates the contribution of innovation direction alone.

6 Conclusion

This paper delves into the impact of the US-China trade war on the innovation strategies of Chinese firms. Given that China's technological progress was one of the primary catalysts for the initiation of the trade conflict by the Trump administration, this study aims to ascertain whether the conflict has influenced the trajectory of China's innovation efforts. Leveraging natural language processing on patent abstracts, we develop a novel metric for measuring patent similarity between China and the US, complementing the citation-based metrics commonly utilized in the literature (e.g., Han, Jiang and Mei (2021)). Our findings reveal that a reduction in export tariffs leads to diminished R&D intensity among publicly listed Chinese firms, alongside a divergence in innovation patterns between China and the US.

This divergence highlights the multi-dimensional nature of innovation activities, as evidenced by the varying degrees of similarity decline between China and other advanced economies, such as Europe, Japan, and South Korea. This suggests that innovation embodies elements of both Romer and Ricardo, as discussed in Hsieh, Klenow and Shimizu (2022). We develop a model incorporating heterogeneous preferences among destination countries to elucidate the potential mechanisms underlying the export shock. An escalation in export tariffs to a particular country diminishes the exporter's incentive to innovate in line with that country's demand. The departure from a symmetric destination setting has received scant attention in prior literature, yet our empirical results underscore its significant role in shaping the nexus between trade dynamics and innovation.

References

Acemoglu, D. (2010), 'When does labor scarcity encourage innovation?', *Journal of Political Economy* **118**(6), 1037–1078.

Acemoglu, D., Aghion, P., Bursztyn, L. and Hemous, D. (2012), 'The environment and directed technical change', *American Economic Review* **102**(1), 131–66.

Acemoglu, Yang, D. Y. and Zhou, J. (2022), Power and the direction of research: Evidence from china's academia.

Aghion, P., Bergeaud, A., Lequien, M. and Melitz, M. J. (2018), 'The Impact of Exports on Innovation: Theory and Evidence', *Working Paper* .

Akcigit, U., Celik, M. A. and Greenwood, J. (2016), 'Buy, keep, or sell: Economic growth and the market for ideas', *Econometrica* **84**(3), 943–984.

Amiti, M., Redding, S. J. and Weinstein, D. E. (2019), 'The impact of the 2018 tariffs on prices and welfare', *Journal of Economic Perspectives* **33**(4), 187–210.

Autor, D., Chin, C., Salomons, A. and Seegmiller, B. (2024), 'New Frontiers: The Origins and Content of New Work, 1940–2018*', *The Quarterly Journal of Economics* **139**(3), 1399–1465.

Autor, D., Dorn, D., Hanson, G. H., Pisano, G. and Shu, P. (2020), 'Foreign competition and domestic innovation: Evidence from us patents', *American Economic Review: Insights* **2**(3), 357–74.

Benguria, F., Choi, J., Swenson, D. L. and Xu, M. J. (2022), 'Anxiety or pain? the impact of tariffs and uncertainty on chinese firms in the trade war', *Journal of International Economics* **137**, 103608.

Bloom, N., Draca, M. and Van Reenen, J. (2015), 'Trade induced technical change? the impact of chinese imports on innovation, it and productivity', *The Review of Economic Studies* **83**(1), 87–117.

Bloom, N., Hassan, T. A., Kalyani, A., Lerner, J. and Tahoun, A. (2021), The diffusion of disruptive technologies, Technical report, National Bureau of Economic Research.

Bloom, N., Jones, C. I., Van Reenen, J. and Webb, M. (2020), 'Are ideas getting harder to find?', *American Economic Review* **110**(4), 1104–1144.

Bloom, N., Romer, P., Terry, S. J. and Van Reenen, J. (2020), 'Trapped Factors and China's Impact on Global Growth', *The Economic Journal* **131**(633), 156–191.

Bombardini, M., Li, B. and Wang, R. (2017), 'Import Competition and Innovation: Evidence from China', *Working Paper* .

Bonadio, B., Huo, Z., Kang, E., Levchenko, A. A., Pandalai-Nayar, N., Toma, H. and Topalova, P. (2024), 'Playing with blocs: Quantifying decoupling', *Working Paper* .

Bown, C. P. (2021), 'The US-China Trade War and Phase One Agreement', *PIIE Working Paper* .

Brandt, L., Van Biesebroeck, J., Wang, L. and Zhang, Y. (2017), 'Wto accession and performance of chinese manufacturing firms', *American Economic Review* **107**(9), 2784–2820.

Comin, D. and Hobijn, B. (2010), 'An exploration of technology diffusion', *American economic review* **100**(5), 2031–2059.

Fajgelbaum, P. D., Goldberg, P. K., Kennedy, P. J. and Khandelwal, A. K. (2019), 'The Return to

Protectionism*', *The Quarterly Journal of Economics* **135**(1), 1–55.

Fajgelbaum, P. D., Goldberg, P., Kennedy, P., Khandelwal, A. and Taglioni, D. (2023), 'The us-china trade war and global reallocations', *American Economic Review: Insights* **forthcoming**.

Gentzkow, M., Kelly, B. and Taddy, M. (2019), 'Text as data', *Journal of Economic Literature* **57**(3), 535–74.

Han, P., Jiang, W. and Mei, D. (2021), 'Mapping us-china technology decoupling, innovation, and firm performance', *Innovation, and Firm Performance* **2**.

He, Z.-L., Tong, T., Zhang, Y. and He, W. (2018), 'A Database Linking Chinese Patents to China's Census Firms', *Nature Scientific Data* **5**.

Head, K. and Mayer, T. (2014), Gravity equations: Workhorse, toolkit, and cookbook, *in* 'Handbook of international economics', Vol. 4, Elsevier, pp. 131–195.

Hoberg, G. and Phillips, G. (2016), 'Text-based network industries and endogenous product differentiation', *Journal of Political Economy* **124**(5), 1423–1465.

Holmes, T. J., McGrattan, E. R. and Prescott, E. C. (2015), 'Quid pro quo: Technology capital transfers for market access in china', *The Review of Economic Studies* **82**(3), 1154–1193.

Hsieh, C.-T., Klenow, P. J. and Shimizu, K. (2022), Romer or ricardo?, Technical report, Working paper.

Hu, A. G. and Jefferson, G. H. (2009), 'A great wall of patents: What is behind china's recent patent explosion?', *Journal of Development Economics* **90**(1), 57–68.

Jiao, Y., Liu, Z., Tian, Z. and Wang, X. (2022), 'The Impacts of the U.S. Trade War on Chinese Exporters', *The Review of Economics and Statistics* pp. 1–34.

Kelly, B., Papanikolaou, D., Seru, A. and Taddy, M. (2021), 'Measuring technological innovation over the long run', *American Economic Review: Insights* **3**(3), 303–20.

Lee, D. D. and Seung, H. S. (1999), 'Learning the parts of objects by non-negative matrix factorization', *Nature* **401**(6755), 788–791.

Lerner, J. and Seru, A. (2017), 'The use and misuse of patent data: Issues for corporate finance and beyond', *NBER Working Paper* .

Lileeva, A. and Trefler, D. (2010), 'Improved Access to Foreign Markets Raises Plant-level Productivity... For Some Plants', *Quarterly Journal of Economics* **125**(3), 1051–1099.

Liu, Q., Lu, R., Lu, Y. and Luong, T. A. (2021), 'Import Competition and Firm Innovation: Evidence from China', *Journal of Development Economics* **151**.

Liu, Q. and Qiu, L. D. (2016), 'Intermediate Input Imports and Innovations: Evidence from Chinese Firms' Patent Filings', *Journal of International Economics* **103**, 166–183.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S. and McClosky, D. (2014), The stanford corenlp natural language processing toolkit, *in* 'Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations', pp. 55–60.

Melitz, M. (2003), 'The Impact of Trade on Intra-Industry Reallocations and Aggregate Industry Productivity', *Econometrica* **71**(6), 1695–1725.

OECD (2007), *OECD Science, Technology and Industry: Scoreboard 2007*, Organisation for Economic Co-operation and Development.

Paatero, P. and Tapper, U. (1994), 'Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values†', *Environmetrics* **5**, 111–126.

Tan, Y., Tian, X., Zhang, X. and Zhao, H. (2020), 'The real effect of partial privatization on corporate innovation: Evidence from china's split share structure reform', *Journal of Corporate Finance* **64**, 101661.

Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R. and Wei, F. (2024), 'Multilingual e5 text embeddings: A technical report'.

Appendix

A Alternative Similarity Measurement

A.1 Alternative Firm-level Vector Construction Method

There is an alternative way to construct vectors for firms and countries with the TF-IDF method. We calculate the similarities between the patents of Chinese listed firms and U.S. patents as follows. On the one hand, for each firm, we add up the vectors of patents filed in the same year and within the same technology class at the 3-digit IPC level and construct firm-year-IPC patent vectors for all the Chinese listed firms from 2000 to 2021. On the other hand, we sum up the vectors of U.S. patents filed in the same year and the same technology class and construct year-IPC patent vectors for U.S. patents.

As a result, we can use a vector to represent the patents of firm i filed in year t and in technology class x. Meanwhile, U.S. patents that filed in year t and in technology class x can also be represented by a vector. For each firm, we calculate the cosine similarity of its patent vector of technology class x in year t to the U.S. patent vector of technology class x from 2000 to year t. We only compare patents within the same technology class since it is meaningless to compare the technical features between biological patents and semiconductor patents.

The above procedures generate a list of similarities for each firm's patents of all technology classes and years. To calculate the average technical similarity of a firm's patents to U.S. patents, we compute a weighted average of the similarities within each technology class where the firm holds patents. The weight assigned to each technology class is proportional to the number of patents the firm has in that class. This process is illustrated in Equation (12).

$$\text{Sim}_{i,t,\tau} = \sum_{x \in IPC_{i,t}} \frac{N_{i,x,t}}{N_{i,t}} CosSim_{i,US,x,t,\tau}\{Vector_{i,t,x}, Vector_{US,\tau,x}\} \tag{12}$$

where $IPC_{i,t}$ represents the set of technology classes in which firm i has applied for patents in year t. $N_{i,x,t}$ denotes the number of patents filed by firm i in technology class x during year t, and $N_{i,t}$ represents the total number of patents filed by firm i in that year. We construct patent similarities between Chinese patents and European patents, Japanese patents, and Korean patents using the same method.

Similarly, we define the aggregate-level similarity between Chinese and U.S. patents with Equation (13). For Chinese patents, we sum up the vectors of patents filed in the same year t and in the same technology class x at the 3-digit IPC level. As a result, each vector represents the technical features of patents filed in year t and in technology class x as a whole. We apply the same method to U.S. patents and construct the patent vectors at the year-IPC level. As a result, we obtained a list of similarities of Chinese patents filed in each technology class x and in each year t to the U.S. patents filed in the same technology class x from 2000 to 2021.

$$\text{Sim}_{x,t,\tau} = CosSim_{x,t,\tau}\{Vector_{CN,t,x}, Vector_{US,\tau,x}\} \tag{13}$$

A.2 Alternative Patent Vectorization Method

Although the TF-IDF method has been widely used in previous studies to generate vector representation of text (Acemoglu et al., 2022; Kelly et al., 2021; Autor et al., 2024), people still have concerns about its drawbacks. On the one hand, it does not consider the semantic meanings of words during vectorization. For instance, The TF-IDF representation of 'man bits dog' is the same as the representation of 'dot bits man'. On the other hand, the vectors generated by the TF-IDF method are sparse and have high dimensions, which, in turn, increases the noise and cost of processing.

The development of the Large Language Model enables us to get better embedding for patent abstracts. We adopt an open-source language model, the E5 Model, which was developed by Microsoft, to vectorize patent abstracts. Embedding Model E5 generates low-dimensional vector representations for arbitrary-length texts, which play key roles in many NLP tasks, such as large-scale retrieval. Compared to high-dimensional and sparse representations like TF-IDF, it has the potential to overcome the lexical mismatch issue and facilitate efficient retrieval and matching between texts. We present a brief summary of the technical features of the E5 Model, and please refer Wang, Yang, Huang, Yang, Majumder and Wei (2024) for details.

The key feature of a good embedding model is that the vector representations of text with similar meanings are close in the embedding space. In order to achieve this, the E5 Model relies on high-quality training data. Previous studies have tried ways of methods to generate training data. For instance, OpenAI treats neighboring sentences as positive pairs to generate training data. The inverse cloze task (ICT) randomly selects a sentence in a paragraph and treats it and the remaining sentences in the same graph as positive pairs. In order to improve the training data, the E5 Model utilizes web crawl data, such as Reddit, StackExchange, and Wikipedia, to generate 1.3 billion text pairs. For instance, it treats post and comment pairs from Reddit and title and abstract pairs from Wikipedia as positive pairs. Moreover, after the first round of training, E5 is fine-tuned with human-labeled data to improve its accuracy. As a result, the E5 Model outperforms the existing embedding models with 10 times fewer parameters.

The E5 Model returns a 1×1024 vector for each piece of patent abstract. One thing that must be pointed out is that the vectors generated by the E5 Model cannot be added up to represent the combination of abstracts. So, we calculate the pairwise similarity between patents of Chinese listed firms and U.S. for each technology class with Equation (2), and the firm-level similarity to U.S. patents is the weighted average across x.

The impacts of the trade war on patent similarity to the U.S., as measured by the E5 Model, are shown in Columns (5)–(8) of Table A-1. These results are similar in magnitude and significance to those obtained using the TF-IDF method. This provides strong evidence of robustness to different measurement choices.

B Patent Data Truncation Issue

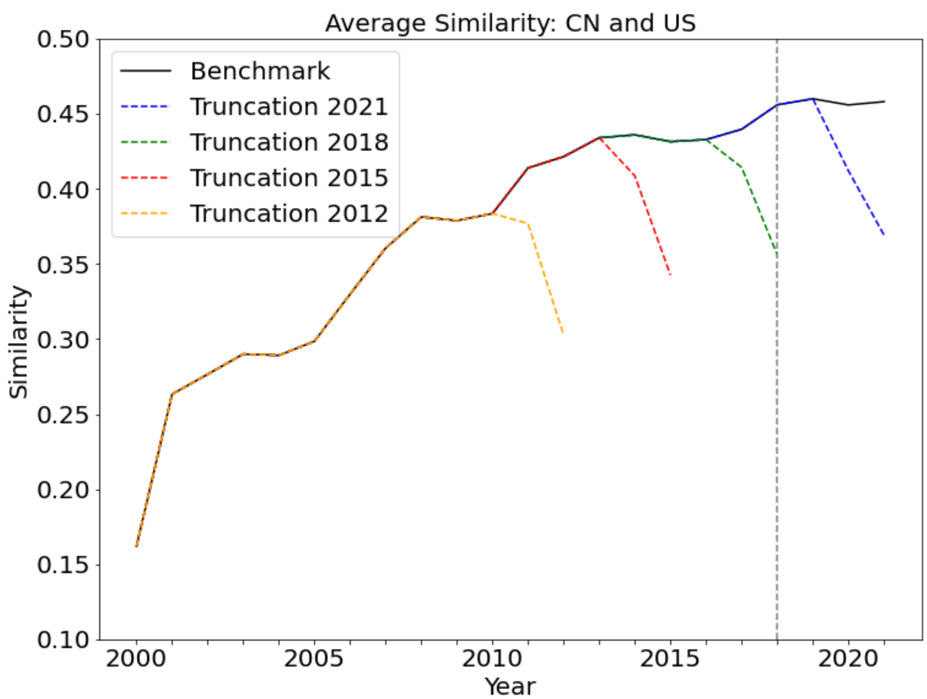
Although we have up-to-date patent data from the Chinese and U.S. patent offices, we only study the patents filed before 2021. The reason is that according to the patent laws in China and the U.S., a patent filing can be kept from the public for at most 18 months. After that, its filing materials, including abstract, claims, reference cited, description, and illustration graphs, should be open to the public. In this project, we collect patent data published till Sept. 2023, which covers all patent filings before 2021.

Moreover, we conduct a robustness check by truncating patent data in different publication years and comparing the changes in aggregate similarities. In Figure A-1, the dark line represents the average similarity between Chinese and U.S. patents from 2000 to 2021. The data used in the calculation are the patent filings that were published before Sept. 2023, and it is the benchmark case in our paper. Then, we manipulate the sample by selecting patents according to their publication year. We selected patent filings that were published before 2021, 2018, 2015, and 2012 and calculated the similarities between Chinese and US patents with the same methods.

The blue dashed line represents the sample with publication year before 2021, and patents that were filed in 2020 and 2021 may not be totally included in the sample. Clearly, compared with the benchmark sample, the similarities in 2020 and 2021 in this truncated sample are substantially lower due to the missing data. Similarly, we observe under-estimated similarities for the years around the truncation year in other truncated samples.

Since our patent data includes all patents published before Sept. 2023, the change of similarity between Chinese and U.S. patents after 2018 is not derived by the data truncation issue.

Figure A-1: Robustness Check on Truncation of Publication Year



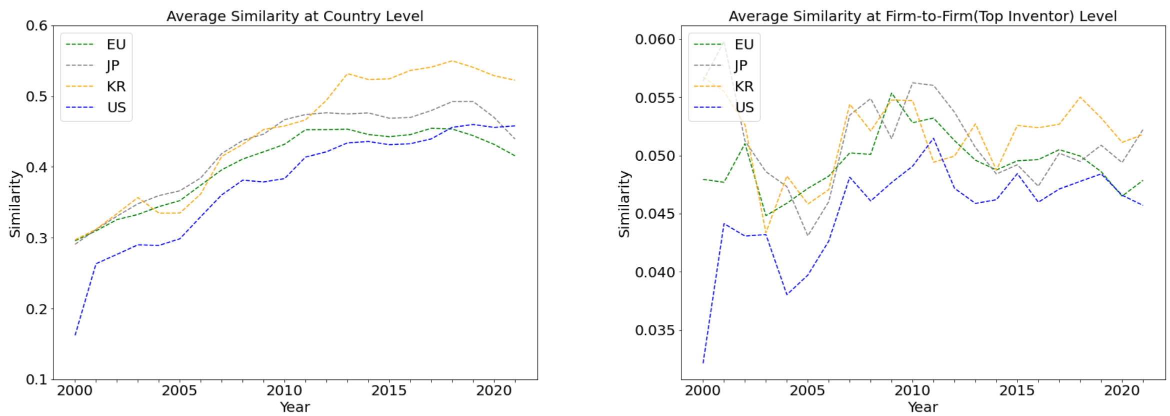
C Similarity for other countries.

In this paper, we study the patents of 16 European countries that had joined the European Patent Convention, including Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, Netherlands, Portugal, Spain, Sweden, Switzerland, and the United Kingdom before 2000. Their patent filings account for almost all of the total patent filings in Europe. We define the EU domestic patents as those filed in these countries and the European Patent Office by the residents in these countries. For Japanese patents and Korean patents, we adopt the same criteria to identify domestic patents. Since the patent office does not always provide an English version of the patent abstract, we look for the patents with non-English abstracts in Google Patents and adopt the English version provided by Google. Only 10.61% of Japanese patents, 20.11% of Korean patents, and 23.21% of European patents need to obtain English abstracts from Google Patents.

We present the aggregate similarity between Chinese patents and foreign patents in the left panel of Figure A-2. The statistics are calculated as follows. We first sum up vectors of Chinese patents by filing year t and three-digit IPC x and construct year-IPC-level patent vectors $V_{t,x,CN}$. Then, we calculate the similarity between the Chinese patent vector $V_{t,x,CN}$ and foreign patent vector $V_{t,x,F}$ for all technology class x from 2000 to 2021. The average similarity in each year is measured as the simple average of the similarities across technology classes. Before 2018, despite a disparity in levels, both Chinese and foreign patents exhibited a comparable upward trajectory, which ceased thereafter. Similarly, we present the aggregate similarity between Chinese listed firms' patents and foreign patents in the right panel of Figure A-2. The statistics are calculated as follows. We first sum up vectors of Chinese listed firms' patents by filing year t and three-digit IPC x and construct year-IPC-level patent vectors $V_{t,x,CN\ List}$. The construction of patent vectors of foreign patents is a bit different. In order to make both sides comparable, we identify the patents filed by top inventors in each technology class and sum up their patent vectors to represent the country-IPC-level patent vector, denoted by $V_{t,x,F\ Top}$. The top inventors in each country and

Figure A-2: Similarity between Chinese and Foreign Patents

(a) Country-Level Similarity **(b)** Firm-Level Similarity



technology class are defined as the twenty applicants with the highest annual average filing activity in each three-digit IPC in each country. Despite being considerably more volatile and lacking a distinct upward trend prior to the trade war, the resemblances between patents held by Chinese listed firms and those of top inventors from foreign nations demonstrate a downward trajectory for most countries post-2018.

D Additional Empirical Results

D.1 Check Pre-existing Trends

To determine whether there were any pre-existing heterogeneous trends among different firms before the trade war, we conduct the following placebo test using a sample from 2013 to 2017, the period before the trade war,

$$Y_{ist}^{placebo} = \sum_{\tau=2014}^{2017} \beta_{1\tau} \mathbb{1}(t=\tau) \Delta ln(1+\text{tariff}_{it}^{E,US}) + \sum_{\tau=2014}^{2017} \beta_{2\tau} \mathbb{1}(t=\tau) \Delta ln(1+\text{tariff}_{it}^{I,US}) \\ + \beta_3 \text{sanctioned}_{it} + \gamma X_{it-1} + \alpha_i + \mu_t + \theta_{st} + \epsilon_{ist}. \tag{14}$$

The dependent variables are firms' patent application number and patent similarity to US patent filed in the recent 5 years. The changes in tariff rates, $\Delta \ln(1+\text{tariff}i^{E,US})$ and $\Delta \ln(1+\text{tariff}i^{I,US})$, are assigned the actual values that the corresponding firms faced four years later. As a result, the year 2013 remains free of any tariff changes, while the years 2014-2017 are assigned the values corresponding to 2018-2021. Figure A-3 presents the point estimates and the 90% confidence intervals for $\beta_{1\tau}$ for each year in the sample. The results indicate that there is no significant difference in the trends between "treated" and "non-treated" firms before the trade war. Therefore, the impact of the export tariff change found in the paper is not due to pre-existing trends.

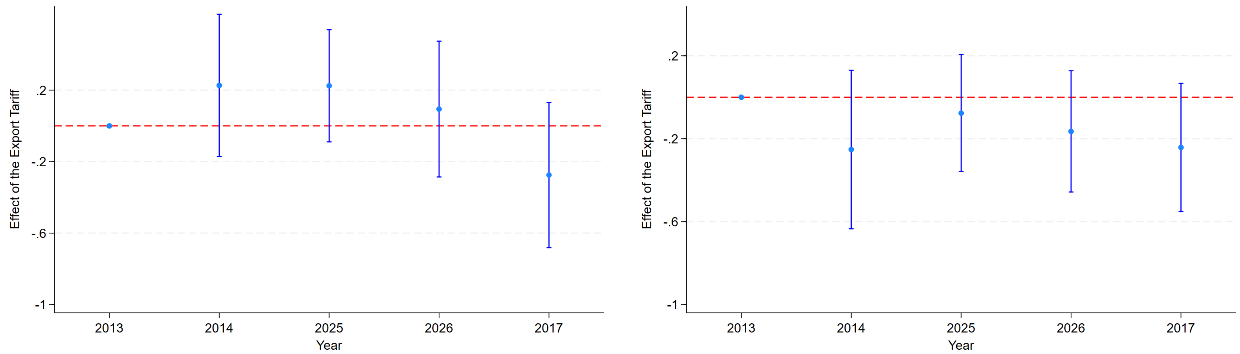
D.2 Strategic Patenting

Patents without the Patent Cooperation Treaty (PCT) applications are less prone to strategic adjustment of their abstract to cater to foreign patent offices. We measure patent similarity based only on patents without PCT applications and rerun the baseline regressions. The results are presented in the first four columns of Table A-1. The effects of changes in export and import tariffs on the similarity of Chinese patents to U.S. patents filed in different periods remain significant and are very close to the baseline results, indicating that the reduction in similarity is not primarily driven

Figure A-3: Effect on Innovation Intensity and Similarity to US Patents

(a) Effect on Patent Application Number

(b) Similarity to US in recent 0-5 Years



Notes: The figure shows the heterogeneous effect of the export tariff change on Chinese firms' innovation intensity and similarity to the US in each year after the trade war. Both the Point estimate and the 90 percent confidence interval are presented. Standard errors are clustered at the firm level. Firm-level controls include the natural logarithm of the firm's employment, total assets, and the share of profit as a proportion of total revenue in the previous year. For regressions on patent applications, a dummy variable indicating whether the firm files for a positive number of patents is controlled. For regressions on patent similarity, the patent application number is controlled. Firm fixed effects, year fixed years, and industry-by-year fixed effects are controlled.

by strategic patenting behaviors.

Table A-1: Impact of the Trade War on Chinese Patents' Similarity to US Patents by Filing Period

	Similarity based on non-PCT Applications				Similarity based on Embedding			
	0-5 Years	0-1 Years	2-3 Years	4-5 Years	0-5 Years	0-1 Years	2-3 Years	4-5 Years
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Δ Export Tariff	-0.225**	-0.241**	-0.251**	-0.185*	-0.228***	-0.229***	-0.230***	-0.228***
	(0.107)	(0.105)	(0.108)	(0.110)	(0.0626)	(0.0625)	(0.0626)	(0.0626)
Δ Import Tariff	-0.379**	-0.386***	-0.352**	-0.370**	-0.271***	-0.267***	-0.267***	-0.267***
	(0.147)	(0.146)	(0.149)	(0.150)	(0.0890)	(0.0888)	(0.0889)	(0.0890)
Sanctioned	-0.0205	-0.0253	-0.0216	-0.0182	-0.0473	-0.0469	-0.0473	-0.0471
	(0.0582)	(0.0590)	(0.0596)	(0.0595)	(0.0323)	(0.0323)	(0.0323)	(0.0323)
Firm Characteristics	Y	Y	Y	Y	Y	Y	Y	Y
Firm Fixed Effect	Y	Y	Y	Y	Y	Y	Y	Y
Year Fixed Effect	Y	Y	Y	Y	Y	Y	Y	Y
Industry $\times$ Year Fixed Effect	Y	Y	Y	Y	Y	Y	Y	Y
Observations	17,869	17,863	17,864	17,857	17,929	17,923	17,924	17,917
R-squared	0.555	0.555	0.551	0.546	0.744	0.744	0.744	0.744

Notes: Standard errors are clustered at the firm level. Δ denotes the difference between the tariff rates in a given year to the average rates in 2014-2017. The similarity measure in columns (1)-(4) are based on patents without PCT applications. The similarity measure in columns (5)-(8) are based on all patents but using the Embedding Model E5 method. Firm-level controls include the natural logarithm of the firm's employment, total assets, and the share of profit as a proportion of total revenue in the previous year. Industries are defined at the 3-digit level. *** Significant at the 1 percent level; ** Significant at the 5 percent level; * Significant at the 10 percent level.

E The NMF Method

In order to generate low-dimensional vector representation for parameter estimation in the quantitative model, we adopt the Non-negative Matrix Factorization (NMF) method to reduce the dimension of vectors generated by the TF-IDF method.

There have been long-lasting interests in transforming texts into low-dimensional dense embeddings. Early works, such as Latent Semantic Indexing (LSA), Latent Dirichlet Allocation (LDA), and Principal Component Analysis (PCA), have been widely used in computational science and natural language processing. However, the low-dimensional matrix generated by these methods contains negative values, which may cause troubles in model estimation. As a result, we adopt the NMF method to lower the dimension of the original TF-IDF matrix, which only returns non-negative values in the low-dimensional matrix.

The NMF method has been widely used in image processing and natural language processing. For instance, in order to do facial recognition quickly, the NMF method can lower the pixels of the original graphs while still keeping the important features, which reduces the time and resource cost in computation. In text mining, the NMF method can reconstruct the original high-dimensional bag-of-words matrix to a low-dimensional topic matrix. Therefore, given a set of documents, the NMF method identifies topics and simultaneously classifies the documents among these different topics. We briefly present the technical features of the NMF method, and please refer to Paatero and Tapper (1994) and Lee and Seung (1999) for technical details.

$$A_{m \times n} = W_{m \times k} H_{k \times n} \tag{15}$$

For a matrix A of dimensions m by n generated by the TF-IDF method, where each element is larger or equal to zero, it can be factorized into two matrices, W and H, as defined in Equation (15). W matrix is usually labeled as the feature matrix, where k is the number of features. H matrix is usually labeled as the coefficient matrix, which serves as the bridge between the original high-dimensional bag of words and the new low-dimensional features. Intuitively, each element of the new low-dimensional vector is a linear combination of elements of the original high-dimensional vector with the coefficients in the H matrix.

In our paper, we adopt the NMF method to identify the technical features of each patent, where we set $k = 256$. As a result, we transform the original $m \times n$ dimensional TF-IDF matrix into a $m \times k$ matrix where m is the number of patents, and n is the count of unique words in the patent abstracts. The coefficient matrix H obtained here can be used to transform any high-dimensional bag-of-words vector to a low-dimensional topic vector.