Appendix A: Chinese firm data

The paper used data from China's stock market extracted from the RESSET Financial Research Database, provided by Beijing Gildata RESSET Data Tech Co., Ltd. (http://www.resset.cn), a leading provider of economic and financial data in China. Our dataset covers the period 1990–2015 and provides some basic registration information of publicly listed firms on China's stock exchanges (Shanghai and Shenzhen), such as listing date, delisting date, registered address, industry category, yearly total revenue, and yearly number of employees. Although the numbers of newly listed and delisted firms in each year fluctuate, the overall number of firms increases almost linearly over time (Figure A1). The registered addresses of firms cover 31 provinces in mainland China.

The industries of firms are aggregated into two levels, 18 categories and 70 subcategories, according to the "Guidelines for the Industry Classification of Listed Companies" issued in 2011 by the China Securities Regulatory Commission (CSRC) (http://www.csrc.gov.cn). CSRC category and CSRC subcategory codes as well as their associated industries are shown in Figure A2. Letters of the alphabet represent the sectoral level, and two-digit numbers represent the subsectoral level. To reduce noise, we consider only subsectors that have more than three firms. After the aggregation, there are 2,690 firms that operate in 18 industries at sectoral level and 70 industries at subsectoral level.

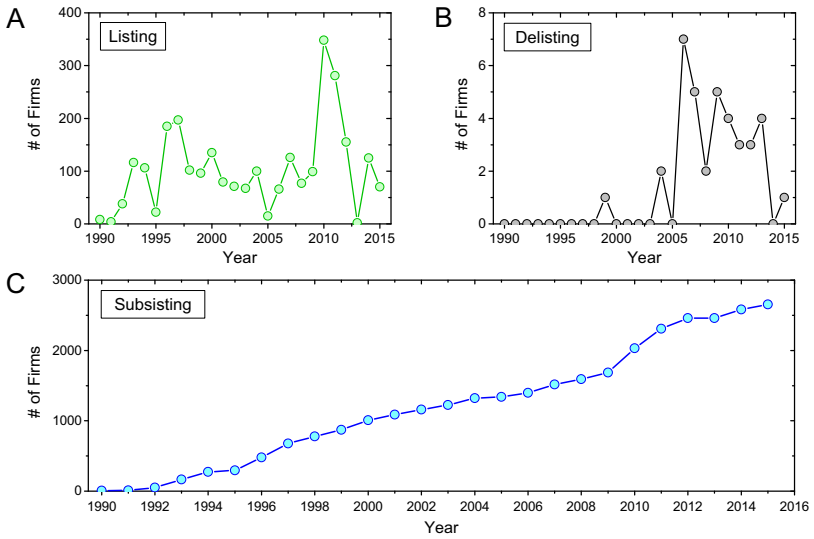

Figure A1: Number of firms that are listed, delisted, and retained in each year. (A) Number of newly listed firms. (B) Number of newly delisted firms. (C) Number of retained firms, that is, the cumulative number of listed firms that have not yet delisted. Source: Authors' calculations.

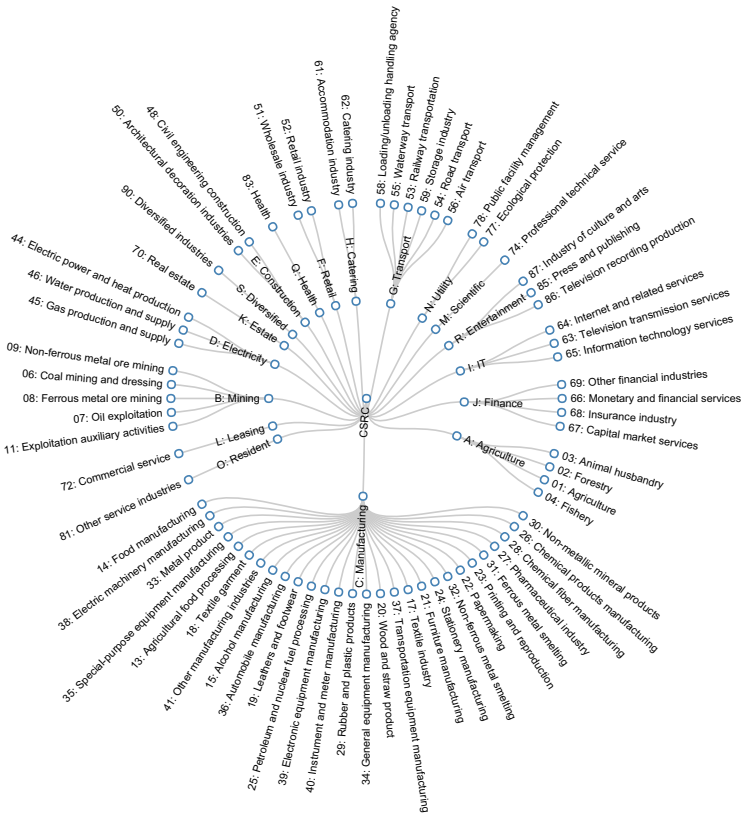

Figure A2: The hierarchical structure of industries after aggregating firms into two levels: sectoral and subsectoral. The inside layer corresponds to the sectoral level coded by letters of the alphabet, and the outside layer shows the subsectoral level coded by two-digit numbers. Source: Authors' calculations.

Appendix B: Distance, travel time, and macroeconomic indicators

We use three distance metrics to explore regional spillovers: geographic, driving, and neighboring distance. We define the geographic distance ($D_{i,j}$) between provinces i and j as a geodesic distance between the capital cities of the two provinces. The driving distance ($V_{i,j}$) is the shortest route between the capital cities of the two provinces, according to Google Maps API in 2015. The neighboring distance $B_{i,j}$ is defined as the least number of provinces an individual in one province has to cross in order to reach another province. For example, the neighboring distance between Beijing and Shandong is two ($B_{i,j} = 2$), because an individual from one province has to cross at least two provinces to reach the other province.

Table B1: Summary statistics of related economic indicators.

Variable	Description	Unit	Obs	Min	Max	Mean	Std. Dev.
A. Province Level							
Population	Resident population in year-end	10k person	31	3.18×10^2	1.07×10^4	4.40×10^3	2.80×10^3
GDP per capita	Per capita GDP	1 CNY/person	31	2.64×10^4	1.05×10^5	5.07×10^4	2.21×10^4
Urban Area	Total urban area in a region	1 sq.km	31	3.62×10^2	2.13×10^4	5.94×10^3	5.17×10^3
Land Area	Total land area in a region	1 sq.km	31	6.34×10^3	1.66×10^6	3.11×10^5	3.87×10^5
Trade	Total value of imports&exports	1,000 USD	31	6.20×10^5	1.24×10^9	1.39×10^8	2.51×10^8
B. Province-pair Level							
Geographical Distance	Between two capital cities	1,000 km	465	114	3,559	1,369	723
Driving Distance	Between two capital cities	1,000 km	465	139	4,883	1,741	962
Neighboring Distance	Number of regions crossed	/	465	1	6	2.9	1.3
Transit Time	Shortest travel time by transit	1 h	465	0.6	71	19.8	14.2
Normal-train Time	Shortest travel time by normal train	1 h	465	1.6	71	25.5	14.3
Driving Time	Shortest travel time by driving	1 h	465	1.9	59	19.4	11.5
Δ Population (log)	Difference in resident population	/	465	0.0071	3.5182	0.9330	0.7650
Δ GDP per capita (log)	Difference in GDP per capita	/	465	0.0000	1.3815	0.4502	0.3316
Δ Urbanization	Difference in urban area/land area	/	465	0.0053	262.31	7.4503	20.973
Δ Trade (log)	Difference in imports&exports	/	465	0.0058	7.6024	1.9055	1.4401

Notes: The summary statistics of macroeconomic data, distance metrics, and travel time measures are for 2014, 2015, and 2015, respectively.
Source: Authors' calculations.

We use three measures to estimate travel time: transit, normal-train and driving time. The transit time is defined as the shortest time by HSR passenger trains. If there is no HSR train on the whole route even by transfer, the shortest travel time by normal train is used. In this study, HSR passenger trains have a number starting with "G," "C," and "D," while normal trains have a number staring with "Z," "T," or "K," or just a number. The driving time is the shortest commuting time between the capital cities of two provinces by driving, which is estimated using Google Maps API in 2015. The introduction of HSR between two provinces is the accessibility between capital cities through HSR passenger trains, which is also identified using Google Maps API in 2015.

We collect macroeconomic data at the province-level, including GDP per capita, population, total value of imports and exports, urban area, and total area (see Table B1). The level of urbanization is defined as the share of urban area in a province. All of these macroeconomic indicators are from "China's Statistical Yearbooks," published by the National Bureau of Statistics of China (http://www.stats.gov.cn). These macroeconomic indicators cover the 1990–2015 period and 31 provinces.

Table B1 shows the brief descriptions and summary statistics of distance metrics, travel time measures, and macroeconomic indicators. At the province level, we use data of population, GDP per capita, urban area, land area, and trade in 2014. At the province-pair level, we use data of geographical distance, driving distance, neighboring distance, transit time, normal-train time, and driving time in 2015, and the data of difference measures, including Δpopulation (log), ΔGDP per capita (log), Δurbanization, and Δtrade (log), in 2014.

Appendix C: Representation of the industry space

We build a "province-industry" bipartite network $G = \{P, I, E\}$ to connect provinces and industries (Figure C1), where P is the set of provinces, I is the set of industries at subsectoral level, and E is the set of links. The weight of link $x_{i,\alpha}$ is the number of firms in province i that operate in industry α. In the following, i and α indicate province-related and industry-related indexes, respectively.

To visualize the network of industries, we build an industry space based on the proximity matrix Φ, that is, similarities between pairs of industries. There are three steps to build the industry space. (i) Build a maximum spanning network, as shown in Figure C2A. We calculate the maximum spanning tree to reach all nodes using a minimum number of links. This network includes 69 links that ensures the connectivity and maximizes the total proximity. (ii) Build a maximum weighted network, as shown in Figure C2B. We keep the links with weight exceeding threshold $\phi' = 0.81$. The network includes 116 links and provides a distinguishable visualization. (iii) Combine the maximum spanning network and the maximum weighted network, as shown in Figure C2C. The superposed network contains 145 links and 70 nodes, which represent 70 industries at subsectoral level.

To present a better network visualization, we use the ForceAtlas2 algorithm in Gephi (http://gephi.github.io) to lay out the superposed network. ForceAtlas2 is a force directed layout, which places each node with consideration of the other nodes and enables us to avoid overlapping links and to untangle dense clusters. Figure C2D shows the layout of the industry space. After preparing the skeleton, we adjust the size of nodes by the number of firms in that industry at the subsectoral level and color the nodes by industries at the sectoral level. Likewise, we adjust the thickness and color of the links by the proximity values. The final industry space based on the data of 2015 is shown in Figure C2E.

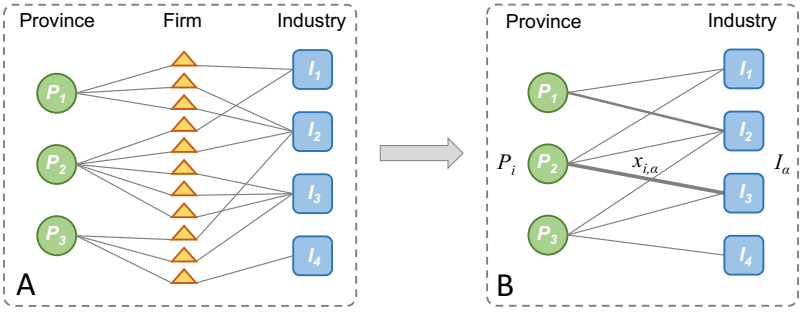

Figure C1: The illustration of "province-industry" bipartite network. P and I represent provinces and industries at the sub-sectoral level, respectively. The weight of link $x_{i,\alpha}$ corresponds to the number of firms in province i that belong to industry α. Source: Authors' calculations.

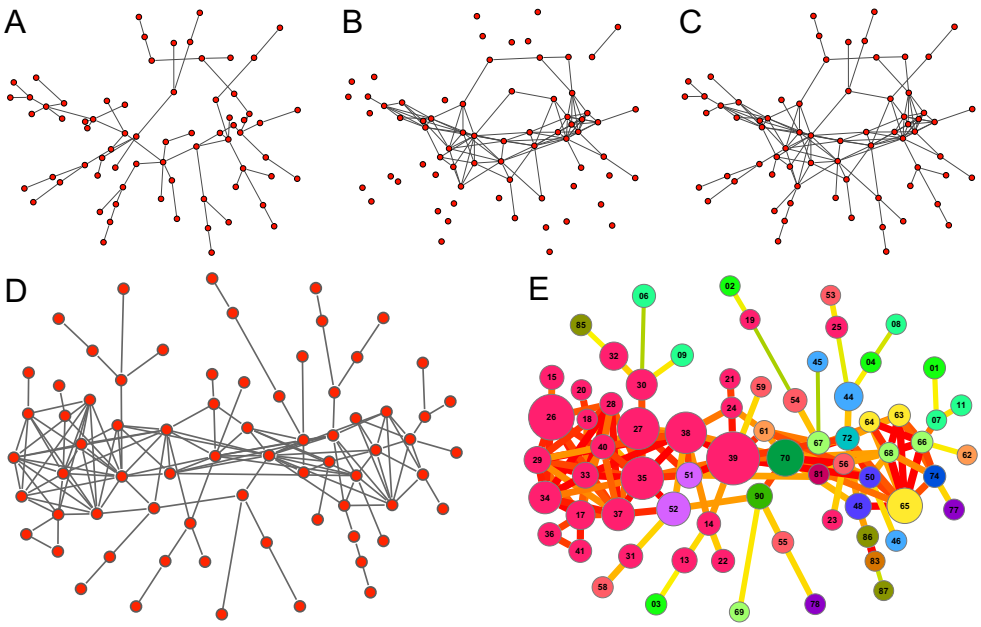

Figure C2: How to construct the industry space.(A) The first step: build a maximum spanning network. (B) The second step: build a maximum weighted network with $\phi > 0.81$. (C) The last step: build a superposed network by combining the maximum spanning network and the maximum weighted network. (D) Layout of the product space, using a ForceAtlas2 algorithm in Gephi. (E) The final outcome: the industry space. The color of nodes corresponds to 18 industries at sectoral level. The size of nodes is proportional to the number of listed firms in that industry. The color and weight of links are associated with the ϕ value between two industries. Source: Authors' calculations.

We further explore the distributions of proximity values among industries. Figure C3A represents the proximity matrix Φ in the way of a clustered structure. The matrix shows two big modules and some small modules, supporting the existence of two density cores in the industry space. Figure C3B describes the density distribution of the proximity values in the matrix Φ, which is approximated by a normal distribution with an average value of around 0.5.

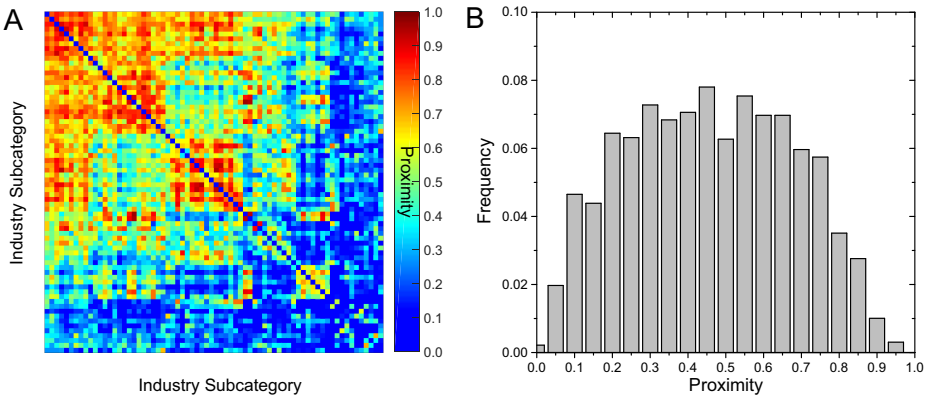

Figure C3: (A) Hierarchically clustered matrix based on the original proximity matrix (Φ). The colors indicate the value of proximity. (B) The distribution of the proximity in matrix Φ. The proximity matrix is calculated based on data in year 2015. Source: Authors' calculations.

Appendix D: Robustness check of inter-industry spillover

We use a visualization method to explore how the structure of the industry space shapes the economic diversification paths of provinces in China. Figure D1 shows the industries that were significantly present in Beijing, Hebei, Shanghai, and Zhejiang, in 1992, 1995, 2000, 2005, 2010, and 2015. In these four illustrative examples, the new industries that are present in each of these provinces tend to be connected to other industries that were already present in that province. For example, Beijing and Shanghai have gradually occupied Internet and financial services industries while Hebei and Zhejiang have gradually occupied manufacturing industries. In particular, the comparative advantage of Shanghai has gradually shifted from manufacturing to services and information activities during this period.

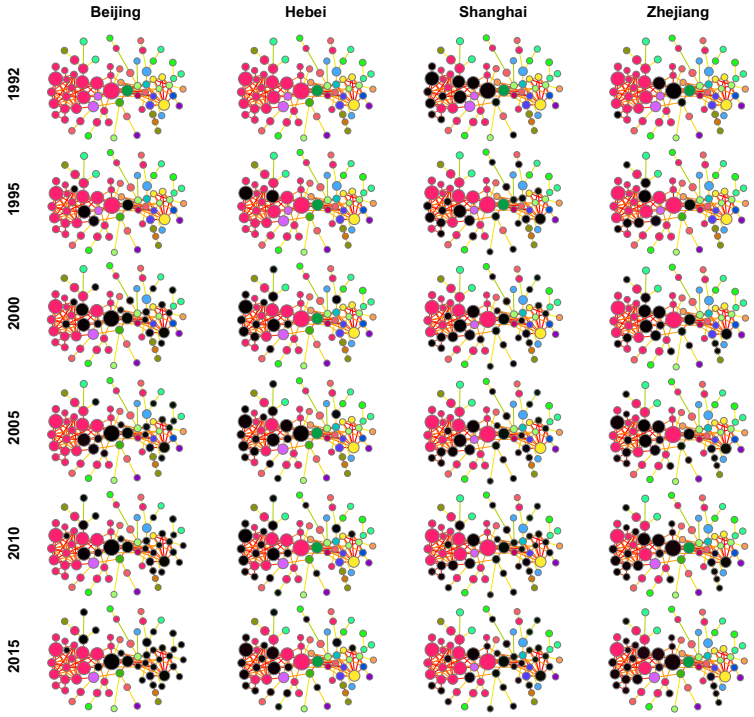

Figure D1: Evolution of China's provincial industrial structure. Four illustrated provinces are Beijing, Hebei, Shanghai, and Zhejiang between 1992 and 2015. Black circles indicate industries in which a province has RCA ($RCA \geq 1$). Source: Authors' calculations.

To check the robustness of inter-industry spillovers, we further explore the relationship between the density of active related industries and the presence of new industries in provinces. Figure D2A presents the relationship between the number of industries, in which provinces have RCA, and the number of new industries, in which provinces have developed RCA in a 5-year period. Using China's firm data, we count the number of industries in 2001 and check if

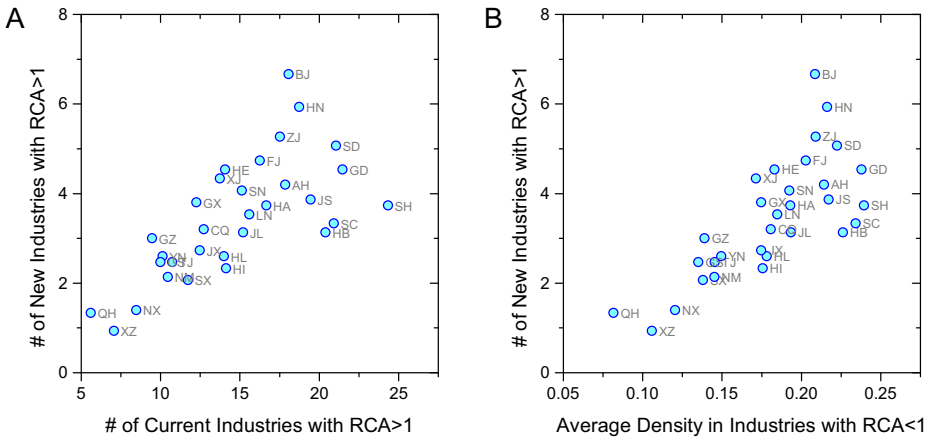

Figure D2: (A) Relationship between active industries at time t and new active industries at time $t+5$ for provinces. (B) Relationship between the average density of active industries at time t and new active industries at time $t+5$ for provinces. Results average for 2001–2015 using 5-year intervals. Abbreviations of province names are shown in Table D1. Source: Authors' calculations.

new industries emerge in 2006. We repeat the analysis for the period 2001–2015 by checking the pairs of years (2002, 2007), (2003, 2008), ..., (2010, 2015) and aggregate all these observations.

Figure D2A shows a positive relationship between the number of industries in which a province has $RCA > 1$ and the number of new industries in which the province develops in 5 years. The horizontal-axis shows the average number of current industries with $RCA > 1$ in year t, and the vertical-axis shows the average number of new industries with $RCA > 1$ in year $t+5$. Provinces did not have industries at the beginning (t) but developed them in 5 years ($t+5$). Figure D2B shows a positive relationship between the average density of industries with $RCA < 1$ and the number of new industries developed within 5 years, suggesting that the average density of industries with $RCA < 1$ predicts the number of new industries developed in the future. Table D1 shows the abbreviations of province names.

Table D1: Abbreviations of province names in China.

ID	Province Name	Abbreviation	ID	Province Name	Abbreviation	ID	Province Name	Abbreviation
1	Beijing	BJ	12	Anhui	AH	23	Sichuan	SC
2	Tianjin	TJ	13	Fujian	FJ	24	Guizhou	GZ
3	Hebei	HE	14	Jiangxi	JX	25	Yunnan	YN
4	Shanxi	SX	15	Shandong	SD	26	Tibet	XZ
5	Inner Mongolia	NM	16	Henan	HA	27	Shaanxi	SN
6	Liaoning	LN	17	Hubei	HB	28	Gansu	GS
7	Jilin	JL	18	Hunan	HN	29	Qinghai	QH
8	Heilongjiang	HL	19	Guangdong	GD	30	Ningxia	NX
9	Shanghai	SH	20	Guangxi	GX	31	Xinjiang	XJ
10	Jiangsu	JS	21	Hainan	HI			
11	Zhejiang	ZJ	22	Chongqing	CQ			

Source: Authors' calculations.

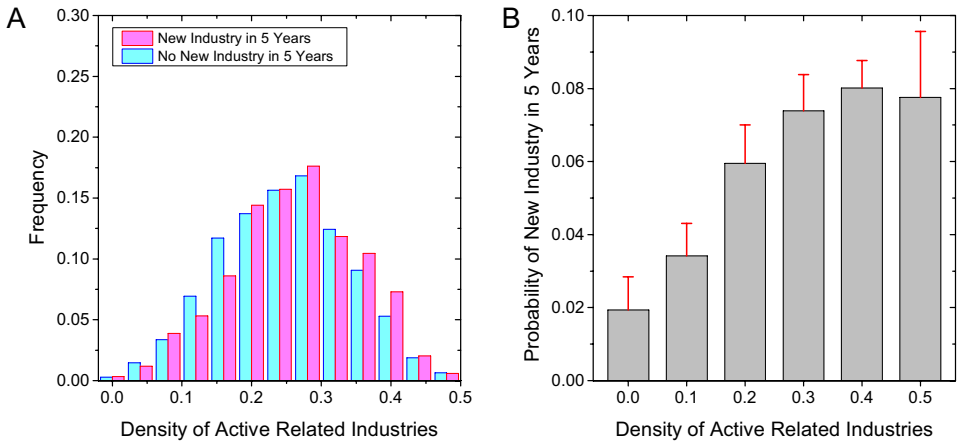

Figure D3: (A) Distributions of densities of active related industries for pairs of industries and provinces that developed RCA in an industry (in pink) and those that did not develop RCA (in blue) in a 5-year period. (B) Probability of a new industries emerging in a province as a function of the density of related industries (ω). Bars indicate average values and error bars indicate standard errors. Results average for 2001–2015 using 5-year intervals. The density is calculated using industrial proximity ($\phi_{\alpha,\beta,t}$). Source: Authors' calculations.

In addition to a fixed proximity matrix used in our main analysis, here, we use time-varying proximity ($\phi_{\alpha,\beta,t}$) to calculate the density of related industries. Figure D3A shows the distribution of related industry densities for pairs of industries and provinces that developed RCA (in pink) and that did not (in blue) within 5 years. We find that the average related industry density for the pairs of industries and provinces in which developed RCA is significantly larger (ANOVA p-value < 0.01). Figure D3B shows that the probability of an industry developing RCA in a province increases strongly with the density of related industries that are already present in that province. Overall, these observations largely support the robustness of our findings on inter-industry spillovers.

Appendix E: Robustness check of regional spillover

We first explore whether provinces in close physical proximity tend to have a more similar industrial structure, that is, a negative correlation between geographic proximity and industrial similarity. Figure E1A shows the distribution of the industrial similarities ($\varphi_{i,j}$) in 2015 for pairs of neighboring (in pink) and non-neighboring (in blue) provinces. The industrial similarity of neighboring provinces is significantly larger than the similarity of non-neighboring provinces (ANOVA p-value=8.1×10^{-4}). Figure E1B shows the industrial similarity ($\varphi_{i,j}$) as a function of geographic distance ($D_{i,j}$). We find that pairs of provinces in close physical proximity tend to be more similar. Figure E2 shows equivalent charts using other distance and travel time measures. Again, the industrial similarity is highly correlated with transit time (A), normal-train time (B), driving time (C), and driving distance (D). These results show that shorter travel time or closer distance between two regions corresponds to more similar industrial structure.

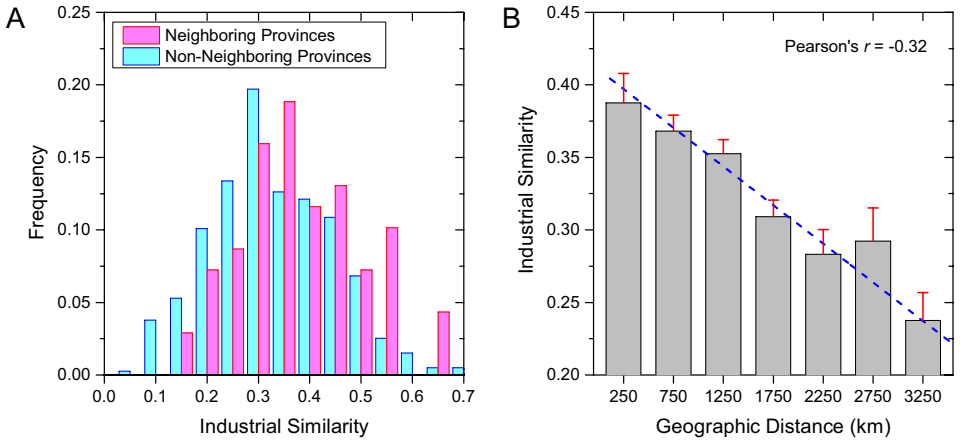

Figure E1: Industrial similarity and geographic distance. Panel (A) Distribution of industrial similarity between pairs of neighboring provinces (in pink) and non-neighboring provinces (in blue). The red and blue curves are normal fits for the distributions for neighboring and non-neighboring province pairs, respectively. Panel (B) Industrial similarity between all pairs of provinces as a function of their geographic distance. Bars correspond to the average industrial similarity of pairs of provinces at that distance and error bars correspond to standard errors. The blue dashed line represents a linear fit of the unbinned data. Pearson's correlation between industrial similarity and geographic distance is $r = -0.32$. Source: Authors' calculations.

We then explore the effects of regional spillovers on developing new industries. In addition to RCA used in the main text, here, we present the spatial evolution of the presence of industries in provinces by showing the number of firms. Figure E3 illustrates the same four industries as shown in the main text: Chemical products manufacturing industry, Pharmaceutical industry, Electric machinery manufacturing industry, and Wholesale industry. The keys of labels correspond to Figure A2, and the saturation of the color indicates the number of firms. We find that provinces

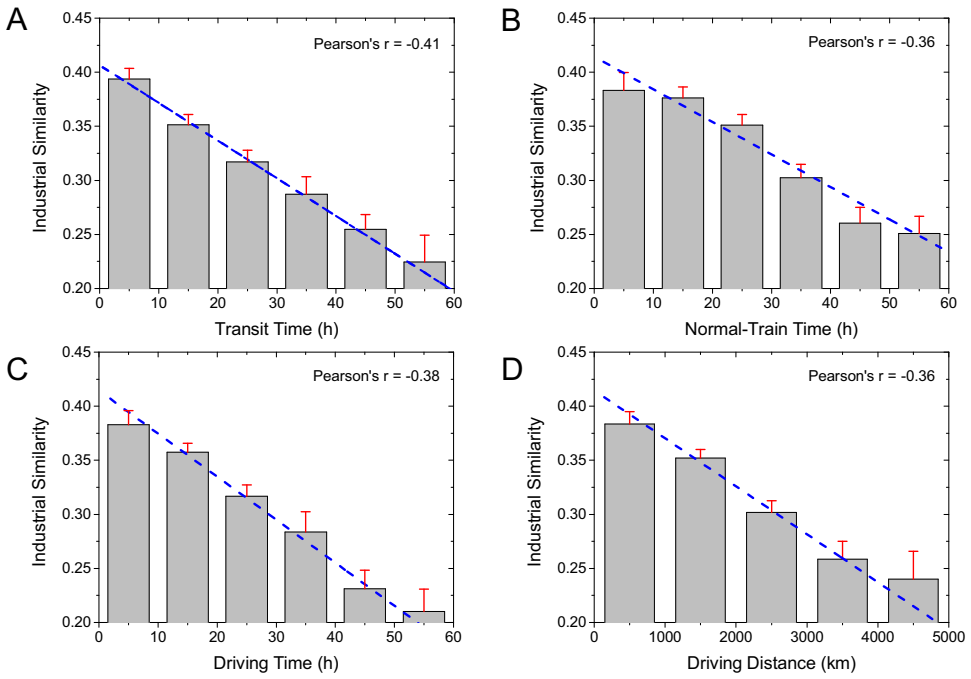

Figure E2: Relationship between industry similarity and (A) transit time, (B) normal-train time, (C) driving time, and (D) driving distance. Bar charts with error bars correspond to average values with stand errors in bins. Blue dash lines are linear fits of the corresponding bar charts. Source: Authors' calculations.

that have large numbers of firms in an industry tend to be neighbors of provinces that already had a large number of firms in that industry, supporting the robustness of our findings.

Next, we test the robustness of our results on regional spillovers. In addition to the geographic distance $D_{i,j}$ used in the main text, we additionally use the neighboring distance $B_{i,j}$ to measure the density of active neighboring provinces. The alternative density measure is given by

$$\Omega_{i,\alpha,t} = \sum_j \frac{U_{j,\alpha,t}}{B_{i,j}} \Big/ \sum_j \frac{1}{B_{i,j}} , \tag{E1}$$

where $U_{j,\alpha,t}$ takes 1 if province j has already developed industry α in year t, and 0 otherwise. Figure E4A shows the distribution of densities (Ω) for industry-province pairs that developed RCA in an industry in a 5-year period (in pink) and those who did not (in blue). The average density of active neighboring province for the pairs that developed RCA is significantly larger (ANOVA p-value=7.3×10^{-36}). Figure E4B shows the positive correlations between the probability for a province to develop RCA in an industry and the density of active neighboring provinces in that industry 5 years before. These observations support the robustness of our findings on inter-regional spillovers.

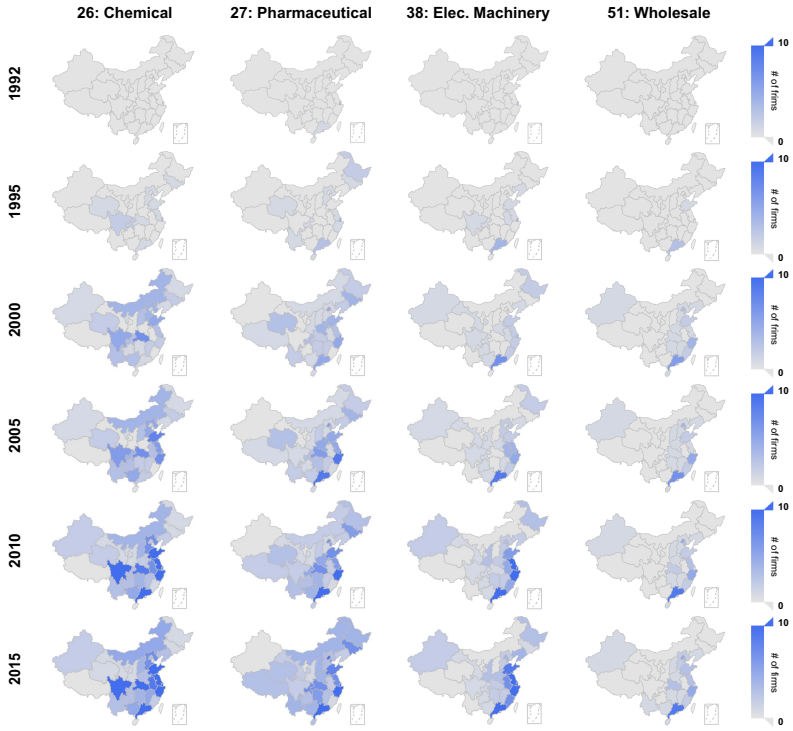

Figure E3: The evolution of the presence of industries in provinces of China between 1992 and 2015. Four illustrated industries are Chemical products manufacturing industry, Pharmaceutical industry, Electric machinery manufacturing industry, and Wholesale industry. The keys of labels correspond to Figure A2, and the saturation of the color indicates the number of firms. Source: Authors' calculations.

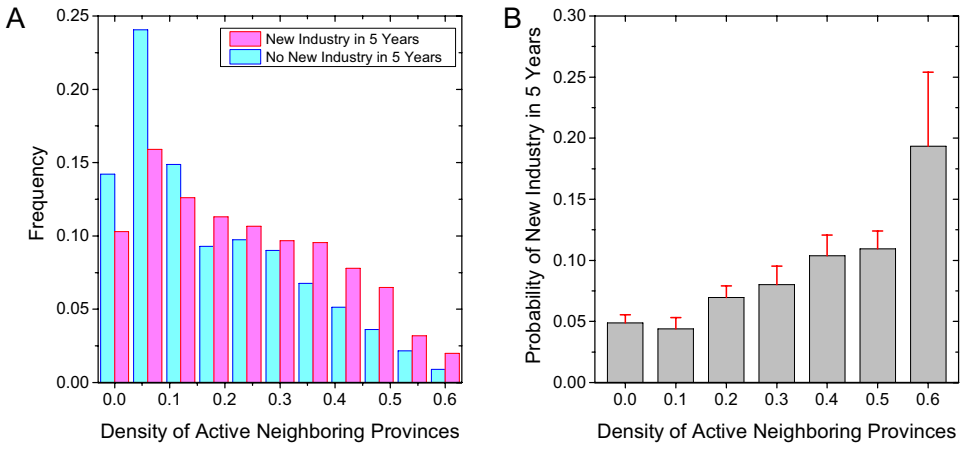

Figure E4: (A) Distribution of densities of active neighboring provinces for province-industry pairs that developed RCA in a 5-year period (in pink) and those who did not (in blue). (B) Probability of developing RCA in an industry as a function of the density of active neighboring provinces 5 years before. Bars show average values, and error bars show standard errors. Results are averaged for 2001–2015 using a 5-year interval. Source: Authors' calculations.

Appendix F: Robustness check of joint effects

In the main text, we consider the joint effects of inter-industry and regional spillovers. Table F1 presents the summary statistics of variables used in the econometric considering both spillovers. Three different groups of metrics are included in the multivariable regressions: density, ratio, and number. Here, the density of active related industries and that of the number of related industries are based on the illustrated industry space in 2015.

Table F1: Summary statistics of regression variables in the analysis of the emergence of new industries.

Variable	Observations	Min	Max	Mean	Std. Dev.
Density of Active Neighboring Provinces	25,713	0	0.7127	0.1894	0.1551
Density of Active Related Industries	25,713	0.0109	0.5939	0.2283	0.0866
Interaction Term 1	25,713	0	0.3022	0.0453	0.0441
Ratio of Active Neighboring Provinces	25,713	0	1	0.1816	0.2358
Ratio of Active Related Industries	25,713	0	1	0.1949	0.2884
Interaction Term 2	25,713	0	1	0.0457	0.1063
Number of Neighboring Provinces	25,713	1	8	4.4463	1.8048
Number of Related Industries	25,713	1	15	3.8851	3.3691
Interaction Term 3	25,713	1	120	17.271	17.712

Source: Authors' calculations.

To check the robustness of our results on the joint effects, we use an alternative index (ratio) to measure the density of active neighboring provinces (Ω) and the density of related industries (ω). For provinces, the ratio is the proportion of active neighboring provinces (R). For industries, the ratio is the proportion of active related industries (r) according to the illustrated industry space in 2015.

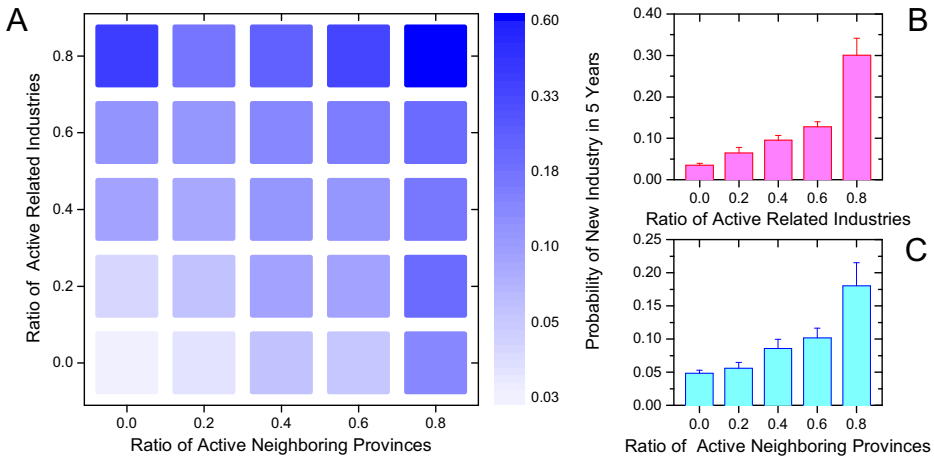

Figure F1: (A) Joint probability of a new industry developing revealed comparative advantage in a province given the ratio of active neighboring provinces on the horizontal axis, and the ratio of active related industries on the vertical axis. The color marks the joint probability of new industries present after dividing the two densities into bins. (B) and (C) are the corresponding marginal probability distributions of developing new industries. Source: Authors' calculations.

Figure F1A shows that both ratios have significant effects on developing new industries, supporting the robustness of our findings on the joint effects. Moreover, the ratio of active related industries (Figure F1B) and the ratio of active neighboring provinces (Figure F1C) both exhibit increasing relationship with the probability of developing new industries.

Appendix G: Robustness check of difference-in-differences estimation

We explore whether pair of provinces connected by HSR are more similar in their industrial structure. Figure G1A shows that the average value of industrial similarity between province pairs connected by HSR (in pink) is significant larger (ANOVA p-value=1.2×10^{-18}) than that pairs that are not connected (in blue). Figure G1B shows the timing of HSR entry and its effects on the industrial similarity of province pairs. The average industrial similarity increases remarkably after train speed-up in 2005, 2008, and 2012 (the years after "speed-up" campaigns are used for illustration), showing the positive and significant effects of HSR on regional spillovers.

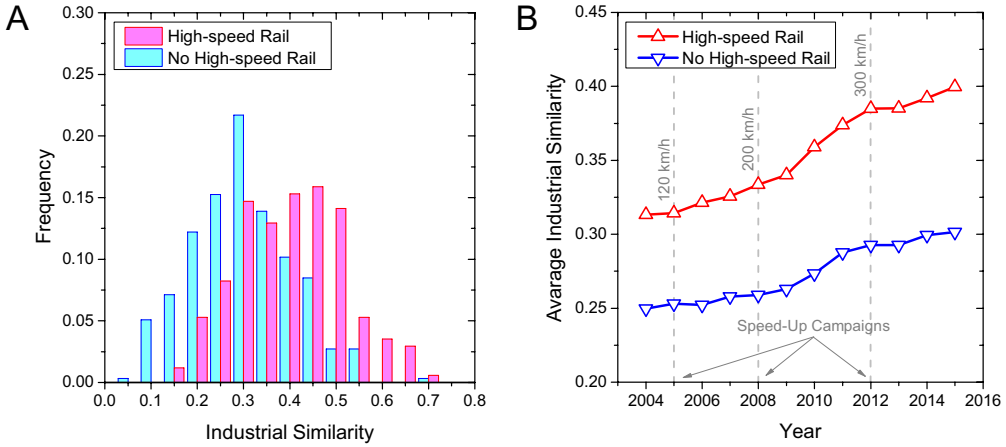

Figure G1: (A) Density distributions of industrial similarity for province pairs with (in red) or without (in blue) high-speed rail. Red and blue curves are normal fits of the bar charts. (B) Average industrial similarity between province pairs with (in red) or without (in blue) high-speed rail. Source: Authors' calculations.

In our DID design, province pairs belong to the treatment group if they are connected by HSR in 2014, and otherwise belong to the control group. In the DID regressions, control variables include gravity considerations: the difference between population, GDP per capita, urbanization (defined as the share of urban area over the entire area of a province), and trade (defined as total exports and imports of each province). Table G1 shows the summary statistics of variables used in the DID analysis in the main text.

We test the robustness of our DID estimations in the main text. Specifically, to reduce sampling bias, we additionally control for the geographic distance between provinces in the DID analysis. As shown in Table G2, the estimates of the effects of HSR entry are also significant and robust in the presence of controls for the geographic distance, supporting the positive and significant effects of HSR entry on regional spillovers.

Table G1: Summary statistics of variables in difference-in-differences analysis.

Independent Variables	Before		After		DID
	Control	Treatment	Control	Treatment	
Industrial Similarity	0.2496	0.3133	0.2994	0.3921	0.0290
Δ Population (log)	1.1394	0.7314	1.0953	0.6514	-0.0358
Δ GDP pc (log)	0.5717	0.6255	0.4603	0.4327	-0.0814
Δ Urbanization	0.1066	0.2115	0.1098	0.2151	0.0005
Δ Trade (log)	2.0719	1.5899	2.2047	1.3863	-0.3365
Observations	295	170	295	170	930

Notes: The table shows mean values of industrial similarity and differences in population (log), GDP per capita (log), urbanization and trade (log) between province pairs before and after the HSR entry. Source: Authors' calculations.

Table G2: Difference-in-differences regressions with controlling for the geographic distance between provinces.

Independent Variables	DID Regressions Using OLS Model: Industrial Similarity		
	(1)	(2)	(3)
HSR Entry	0.0290*	0.0270*	0.0274*
	(0.0151)	(0.0150)	(0.0151)
Treatment Group	0.0504***	0.0455***	0.0473***
	(0.0112)	(0.0114)	(0.0113)
After Entry	0.0498***	0.0470***	0.0504***
	(0.0089)	(0.0089)	(0.0089)
Distance (log)	-0.0297***	-0.0261***	-0.0278***
	(0.0069)	(0.0069)	(0.0070)
Δ Population (log)		-0.0182***	
		(0.0048)	
Δ GDP per Capita (log)		-0.0176**	
		(0.0081)	
Δ Urbanization			0.0146
			(0.0130)
Δ Trade (log)			-0.0049**
			(0.0024)
Observations	930	930	930
Robust R^2	0.1826	0.1983	0.1859
RMSE	0.1096	0.1087	0.1095

Notes: DID regressions using the OLS model considering the effects of HSR entry on the industrial similarity after controlling macroeconomic indicators, including geographic distance, population, GDP per capita, urbanization, and trade. Data are for 2004 (before HSR entry) and 2014 (after HSR entry). Robust standard errors are reported in parentheses. Significance level: $*p < 0.1$, $**p < 0.05$, and $***p < 0.01$. Source: Authors' calculations.