Revealing and maximizing the collective learning effects in industrial diversification

Jian Gao[1,2,3,*], Bogang Jun[2], Tao Zhou[1,3], and César A. Hidalgo[2,*]

[1]CompleX Lab, Web Sciences Center, University of Electronic Science and Technology of China, Chengdu, 611731, China.
[2]Collective Learning Group, MIT Media Lab, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.
[3]Big Data Research Center, University of Electronic Science and Technology of China, Chengdu, 611731, China.
[*]E-mail: gaojian@mit.edu or hidalgo@mit.edu

ABSTRACT

Industrial development has been modeled as a collective learning process at the macro level. In this paper, we reveal the inter-industry and inter-regional collective learning effects in regional industrial diversification by exploring Brazilian labor data in the period of 2006-2013. We quantify the genuine core-periphery structure of industry space–a network representation of the relatedness between industries–and provide evidence of inter-industry learning from related industries within the same region and inter-regional learning from neighboring regions for the same industry. In particular, we show that the probability of developing new industries increases with the density of related industries and neighboring regions. Further, we explore the maximization of the collective learning effects by using a propagation model to simulate regional industrial diversification with different initial conditions. The results suggest the optimal strategies for initially developing industries according to the industry space structure and for building spatial connections between regions with consideration of their distance. Our work highlights the crucial role of collective learning and suggests its promising application in promoting regional industrial diversification.

Introduction

Understanding the underlying mechanisms of economic development and diversification is a long-standing challenge in the fields of development economics and economic geography.[1–3] To deal with the emerging complexity that originates in real-world economic systems,[4,5] in recent decades, the research paradigm has been extended from modeling with assumptions and explaining with multivariables[6] to data-driven approaches with new data resources and analytic tools.[7,8] On the one hand, the rapid development of information technology helps in collecting high quality and large-scale data that can be used to reveal and nowcast the status of economic development,[9,10] such as world trade data,[11] public firm data,[12] imagery data,[13] mobile phone data[14] and social media data.[15] On the other hand, the progress in the interdisciplinary fields of network science,[16] economic complexity,[11] computer science[13] and statistical physics[17] can help to better capture the structural information of economic development, like the product space,[18] and elucidate the underlying mechanisms of industrial diversification, like the spreading processes on complex networks.[19–21]

From the emerging concern to understand how regions manage to export new products and diversify into new industries, recent literature has focused on the effects of collective learning[22,23]–the learning that takes place at the scale of teams, organizations and nations[24,25]–by highlighting two learning channels.[22] One channel is inter-industry learning, focusing on the learning effects from related economic activities within the same region.[22] For example, countries are more likely to export new products that have a high relatedness with their existing products,[18] and regions have been found path-dependent to develop new industries that are related to the preexisting industries,[26,27] where the two kinds of relatedness are measured by the co-occurrence of products and co-locating of industries. The other channel is inter-regional learning, focusing on the learning effects from the neighboring regions with the same economic activity.[22] For example, countries with neighboring countries exporting a certain product are more likely to export that product,[28] and regions with an industry in neighboring regions have a high probability to develop and sustain that industry in the future.[29,30]

Collective learning has been implicated in exploring the underlying mechanisms of economic diversification at different scales from nations to regions,[18,26] across different continents in the North America, Europe and Asia,[27,28,31] and based on different types of data like product trade, public firms and manufacturing plants.[22,29,30] Due to a lack of data and methods, however, the problem requires further investigations in the following respects. First, learning from related industries and neighboring regions is more likely to originate in the input side of economies. Compared to the previously used product data from the output side,[11,18] the capabilities of economic development can be better extracted from the labor data.[32] Second, the regional learning effects could be further tested and generalized for countries at different stages, not only developed countries

like the US[29] and Sweden,[26] but also developing countries like China[22,27,30] in East Asia and Brazil in Latin America. Third, the optimal strategy to benefit from the learning effects for regions with different preexisting industries could be studied by using propagation models and simulations[19,20] after quantification of the core-periphery structure[33] of the industry space[22] and the spatial interactions among regions.[20] The recent availability of new data and methods motivates the addressing of these aforementioned points.

In this paper, we study regional industrial diversification by exploring Brazilian labor data covering the period of 2006-2013. We find evidence in support of both inter-industry learning and inter-regional learning. First, we map out the industry space by considering the co-hiring of occupations among industries and quantify its core-periphery structure. Then, we describe how Brazilian regions learned to diversify into new industries in terms of industry space and geographic illustration. Next, we quantify the collective learning effects by using graphical methods after measuring the density of economic activities for each channel. In particular, the probability of developing new industries or keeping previous industries increases strongly with both densities; however, their combination exhibits diminishing returns, meaning that the two learning channels are substitutes. Finally, we explore the maximization of two learning effects by using propagation models and we find optimal strategies for initially developing core/periphery industries and connecting short/long-range regions. Our results reveal collective learning effects with promising applications in advancing regional economic development practices.

Results

We use Brazilian labor data with the name RAIS (Annual Social Information Report), which covers about 97% of Brazilian formal labor market and was compiled by the Ministry of Labor and Employment (MET) of Brazil. The used dataset covers 76.62 million workers in 501 occupations during 2006 and 2013. The aggregation of occupations follows the Brazilian Occupations Classification (CBO 2002). Firms locate in 558 regions at the Microregion level (further aggregated into 137 Mesoregions, 27 States and 5 Regions) and operate in 669 industries at the Class level (further aggregated into 87 Divisions and 21 Sections) according to the National Classification of Economic Activities (CNAE). Relationship among attributes and summary statistic of the used RAIS dataset are present in Figure S1 and Table S1, respectively.

We first map the labor data to the industry space (see Figure 1A), in which each node represents an industry at the Class level and the link connecting nodes shows the proximity between two industries. The industrial proximity is measured by co-hiring of occupations, where two industries are considered to have a high proximity if they hire workers for the same occupations (see Methods). Briefly speaking, we first build an "Industry-Occupation" bipartite network, where the weight of link is the number of workers in the connected occupation and industry (see Figure S1B for illustration). Then, we calculate the proximity between industries using cosine similarity measures between two vectors, which summarize only occupations with the revealed comparative advantage (RCA) in each industry (see Methods). Figure 1B shows a log-normal like distribution of all proximity values with a long-tail. The results are robust if we use other alternative similarity calculation methods (see, for example, Figures S2A-C). Next, we build the industry space as shown in Figure 1A by overlapping two networks, which are both extracted from the proximity matrix (see Methods). In the space, the number of links is about 6.5 times of the number of nodes with the ratio of links to the complete proximity being 9.8×10^{-3}. The color of nodes differentiates industries at the aggregated Section level, the size of nodes is proportional to the number of workers in the industry, and the weight of links corresponds to the proximity value between the connected industries.

Three observations are notable from the illustrated industry space in Figure 1A. First, industries belonging to the same industry Section (with the same color) tend to connect with each other and have high proximity with each other (Figure 1C). This observation suggests the validation of proximity measure referring to the traditional CNAE classification (see Figures S2D-F for robustness check using alternative proximity measures). Second, the industry space has the core-periphery structure with a big tightly knit core locating at the center (together with some small cores) and various peripheries locating at the outsides. We further calculate a core-periphery value to verify this distinct structure of the industry space (see Methods). Figure 1D shows a inverse-U shape in the core-periphery value as a function of the ratio of added links. The illustrated industry space (Figure 1A) has the relatively high core-periphery value at 0.30, as marked by the red star in Figure 1D (see Figures S2H-I for results using alternative proximity measures). Third, industries with high level of complexity (and high relatedness within sectors) occupy the cores of the industry space, such as the big core of Processing Industries and some small cores of Information and Communication, Financial Activities and Extractive Industries (see Figure 1A). By comparison, industries with low level of complexity occupy the peripheries of the space, such as Trade, Public Administration, Agriculture and Animal Farming, etc.

We next explore how Brazil learned to diversify its regional economy by looking at the presences of competitive industries in the space, focusing on the effects of inter-industry learning. Here, one region is considered to have competitiveness (or active) in one industry if the revealed comparative advantage (RCA)[34] is over one (see Methods). Figure 2A illustrates the evolution of regional industrial structure with active industries (highlighted by color circles) and the market share of industries in two Brazilian microregions, Sao Paulo and Brasilia (see Figure S3A for Bello Horizorite). We find two main observations.

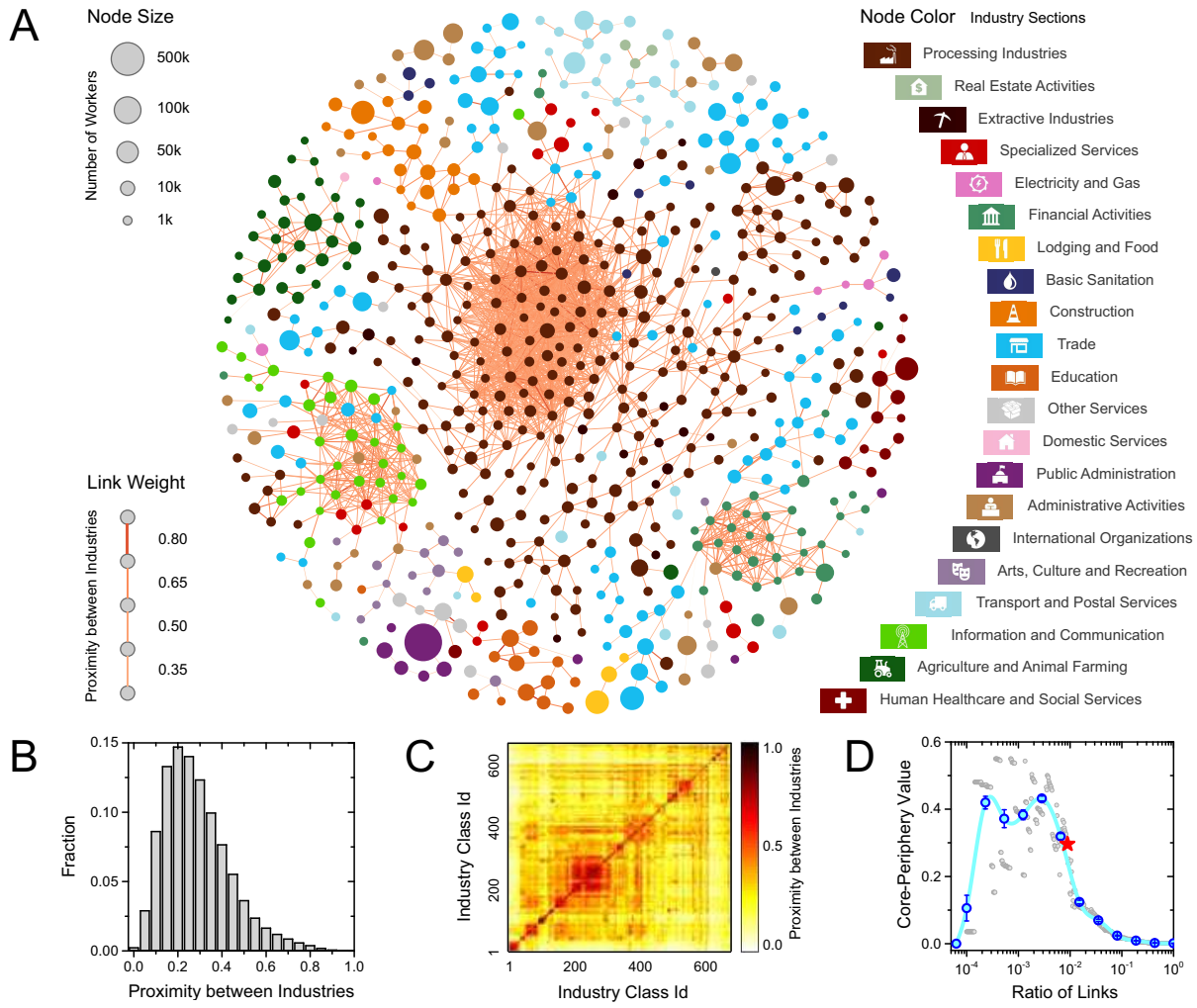

Figure 1. Brazilian industry space illustration. (A) Network representation of the industry space for 2013. Nodes represent 669 industries at the Class level, which are colored by the aggregated 21 Sections. The size of node is proportional to the number of workers. Links connect industries that are likely to hire workers for the same occupations. The weight and color of links correspond to the proximity value between industries. The average degree of the network is around 6.5. (B) Density distribution of proximity values across all industries. (C) Proximity matrix ordered by the industry Class Id referring to the CNAE, in which industries within the same aggregated categories are coded nearby. (D) Core-periphery value of the network as a function of the ratio of added links to the complete network. The network in panel (A) has the ratio 9.8×10^{-3} and the core-periphery value 0.30, as marked by the red star.

On the one hand, developed regions have competitiveness in more and core-located industries. For example, Sao Paulo occupies more Processing Industries. By comparison, developing regions occupy less and periphery-located industries. For example, Brasilia are only active in some Agriculture and Animal Farming Industries and Public Administration Industries (see also the bottom tree map). On the other hand, active industries tend to have large proximity with each other, as indicated by their close locations in the space. Also, industries that are surrounded by active industries in the space are more likely to become active in the future. For example, Sao Paulo diversified into more Information and Communication Industries in 2013 when it already had more active related industries in 2006. These results provide evidence of inter-industry learning.

Analogously, we explore how regions diversified into new industries that their neighboring regions already had competitiveness, focusing on the effects of inter-regional learning. Figure 2B illustrates the evolution of active microregions (highlighted by color circles) for one industry and the market share of all regions in two industries, Manuf. N.F. Ceramic Products and Cooperative Banks industries (see Figure S3B for another example). Two observations are notable. On the one hand, regions that are geographically close located tend to have the same industry. For example, Cooperative Banks Industries especially dominate the South Region (see also the bottom tree map). Neighboring regions are more similar in the industrial structure

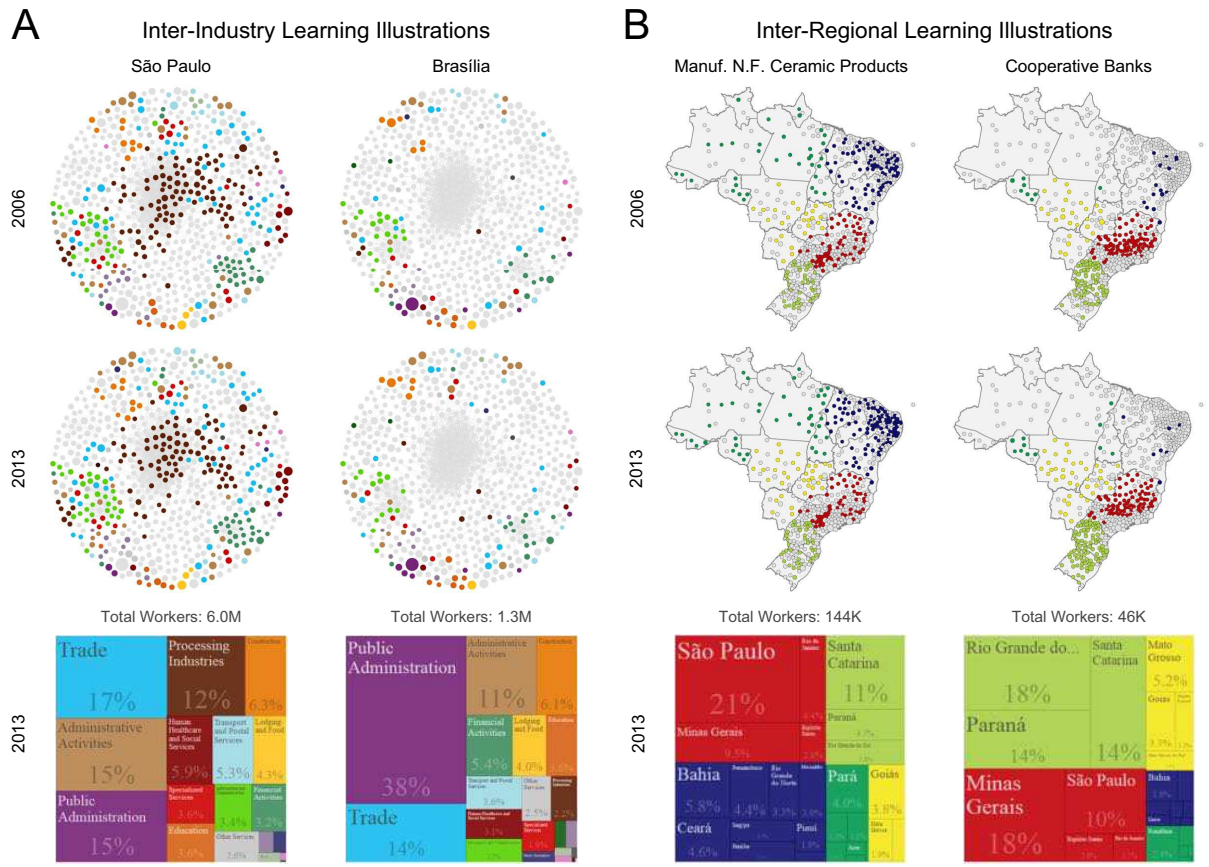

Figure 2. Illustrating Brazilian industrial diversification in 2006 and 2013. (A) Inter-industry learning illustrations by using Sao Paulo (left) and Brasilia (right). Colored circles highlight comparative industries in 2006 and 2013. Tree maps show the percentage of market share (number of workers) of industry Sections in 2013. (B) Inter-regional learning illustrations by using Manufacture of Non-Refractory Ceramic Products (left) and Cooperative Banks (right). Colored circles highlight geographic locations of microregions with competitiveness. Tree maps show the percentage of market share (number of workers) of Regions in 2013.

of active industries (see Figure S4A and SI for details). The industrial similarity among regions decays linearly with their geographic distances, either for all industries (see Figure S4B) or for industries at Section level (Figure S5). On the other hand, regions that are surrounded by many active neighboring regions for one industry are more likely to become active in the future. For example, more microregions in the Northeast Region became active in Manuf. N.F. Ceramic Products in 2013 when they already had many active neighboring microregions in 2006. These results show evidence of inter-regional learning.

We further formalize these collective learning observations in developing new industries by using a graphical method. Here, we apply more strict roles by restricting "developing a new industry" at $t+2$ (compared to t) to two conditions: 1) Presence. Region should have "no workers" at t and "at least five workers" at $t+2$ in that industry. 2) Stability. Region should keep "no workers" at $t-1$ (backward condition) and "at least five workers" until $t+3$ (forward condition) in that industry. These additional restrictions help to filter out some temporal presences of new industries.

Figure 3A presents the inter-industry learning results. We find the probability for a region to develop a new industry in next two years increases strongly with the density of active related industries. In other words, industries that will be developed in the future are with higher density of active related industries (see Figure S6A). Here, this density metric is used to measure how many related industries in the industry space are already active in the region (see Methods). Figure 3B presents the inter-regional learning results. We find an increasing tend in the probability for a region to develop a new industry in next two years as a function of the density of active neighboring regions. Also, regions that will develop new industries have higher density of active neighboring regions (see Figure S6B). Here, another density metric is used to measure how many neighboring regions of the region are already active in one industry (see Methods). Figure 3C shows the results of both learning channels. We find the probability of developing new industries increases with both densities of active related industries and active neighboring provinces. These results are robust if using other definitions of "developing a new industy" (see, for example, using RCA≥ 1

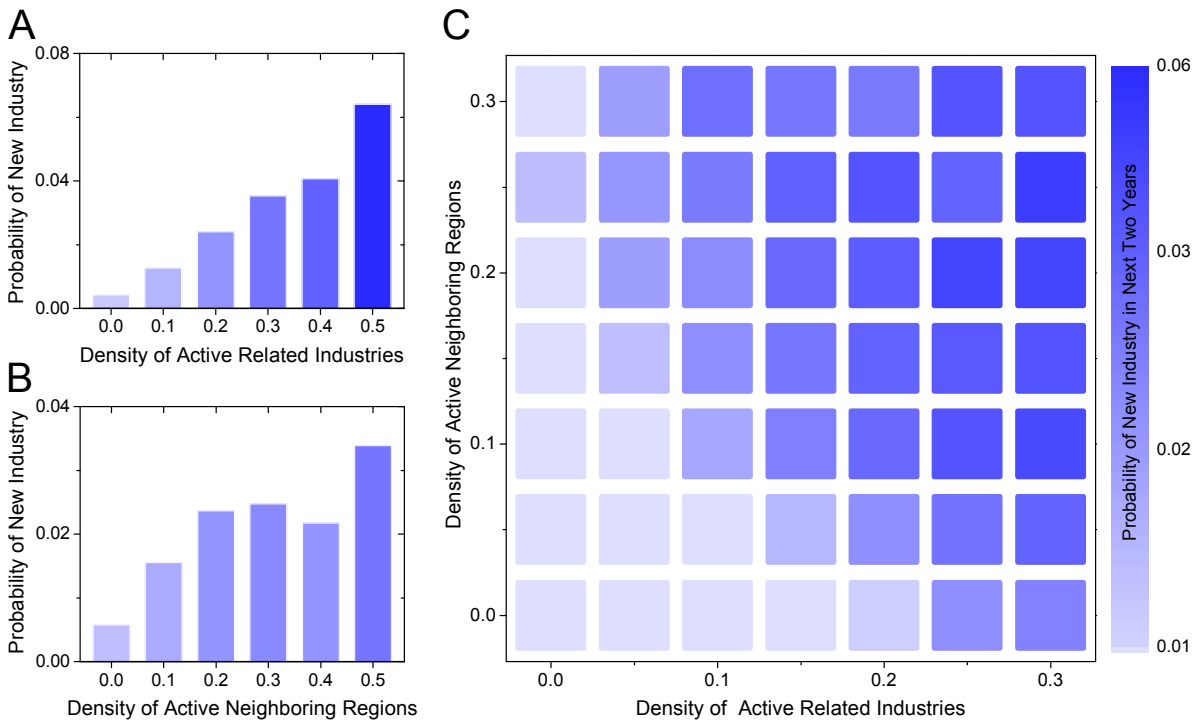

Figure 3. Formalization of collective learning effects. (A) Inter-industry learning: Probability of developing a new industry in a region in next two years as a function of the density of active related industries. (B) Inter-regional learning: Probability of developing a new industry in a region in next two years as a function of the density of active neighboring regions. (C) Combining two learning channels: Joint probability of developing a new industry in a region in next two years given the density of active related industries in horizontal-axis and the density of active neighboring regions in vertical-axis. Results are averaged for 2006-2013.

in Figure S7).

To cross-valid the results observed in Figures 2 and 3, we further introduce a multivariate statistical method to quantify the two collective learning effects. The regression results are summarized in Table 1 after using a probit model (see Methods for models, Table S2 for summary statistics and Table S3 for variable correlations). The regression table shows how the probability of developing a new industry changes with the density of active related industries in columns (1-3) and the density of active neighboring regions in columns (4-6). In all specifications, we find the two densities are both strong, positive and significant predictors of developing new industries, with controlling for the effects of the number of active industries of the region and the number of regions in which the industry is active. It suggests that our findings are not just a reflection of the ubiquity of an industry or the diversity of a region, and the effects of inter-industry learning are stronger with larger regression coefficient. In addition, we find both densities have positive and significant predictability for regions to keep previous industries (see SI for models, Table S4 for summary statistics, Table S5 for variable correlations and Table S6 for regression results).

We further check the interactions between the two learning channels by using the probit model (see Methods for models, Table S7 for summary statistics and Table S8 for variable correlations). Table 2 presents the regression results on developing new industries in columns (1-3) and keeping previous industries in columns (4-6). We find both densities are jointly significant in developing (see column (1)) and keeping industries (see column (4)), and we confirm the inter-industry learning effects are stronger (see columns (1) and (4)). Once the interaction term is added, we find its regression coefficient is negative and significant and regression coefficients of two densities are increased (see in column (2)). These observations are robust when using the ratio of active industries/regions (see columns (3)) and other alternative definitions of density measures (see Table S9 and SI). The negative interaction term suggests the presence of diminishing returns, meaning that the two learning channels are substitutes in developing new industries. However, it is still hard to tell if they are substitutes in keeping previous industries, since the coefficient of the interaction term is not significant for density in column (5) but negative and significant for ratio in column (6) and other density measures (see Table S9).

Finally, we explore how to maximize the collective learning effects by using simulations based on a simple threshold propagation model,[19,20] where inactive nodes become active in the next step if over half of its neighboring nodes in the

Table 1. Probit regressions for inter-industry learning and inter-regional learning in developing new industries.

Developing New Industry	Probit Model					
	Density of Active Related Industries			Density of Active Neighboring Regions		
	(1)	(2)	(3)	(4)	(5)	(6)
Density	3.0073***	3.2439***	2.0810***	1.8167***	1.9864***	1.3446***
	(0.0366)	(0.0364)	(0.1115)	(0.0272)	(0.0283)	(0.0786)
Number of Active Provinces in Industry		0.0030***	0.0033***			0.0014***
		(0.0001)	(0.0001)			(0.0001)
Number of Active Industries in Province			0.0021***		0.0047***	0.0049***
			(0.0002)		(0.0001)	(0.0001)
Constant	−2.5784***	−2.7866***	−2.8354***	−2.4134***	−2.8174***	-2.8512***
	(0.0080)	(0.0086)	(0.0098)	(0.0072)	(0.0092)	(0.0096)
Observations	1,083,154	1,083,154	1,083,154	1,083,154	1,083,154	1,083,154
Pseudo R^2	0.0337	0.0477	0.0484	0.0184	0.0479	0.0484

Notes: Probit regressions on the probability of developing a new industry as a function of the density of active related industries or the density of active neighboring regions with controlling for the effects of the number of active industries that one region has and the number of regions in which an industry is active. Data are for the 2006-2013 period. Probit regressions include year-fixed effects. Robust standard errors are reported in parentheses. Significant level: $*p < 0.1$, $**p < 0.05$, and $***p < 0.01$.

network are currently active (see Methods). The purpose is to study the optimal strategies to advance the final activation of all industries if the capabilities or investments are limited at an early economic development stage, in our case, the ratio of initial activated industries in the industry space and the ratio of initial activated regions for one industry. In propagation simulations, three metrics are of interest, namely, the ratio of active nodes to all nodes (S_a), the ratio of the giant component of active nodes (S_{gc}), and the number of iterations (time steps) before reaching the steady state (NOI).

The maximization of the collective learning effects is studied by a distinct strategy for each channel. For the inter-industry learning, the strategy decides which set of industries are suggested to be initially developed.[35] Considering that the industry space[22] has core-periphery structure and the degree distribution of nodes is heterogeneous,[36] different industries face different opportunities to be developed.[37,38] The structural information of a region's current industries will condition its future growth paths,[30] leading to an optimal strategy to diversify into new industries. For the inter-regional learning, the strategy decides whether to choose nearby or distant regions to establish new spatial connections. For example, building high-speed rails will connect short-distance regions,[39] while opening flights can significantly reduce the commuting time between long-distance regions (in some sense, the connected distant region becomes a neighboring region) but with relatively large costs.[40] The different operating costs will lead to an nontrivial strategy in determining the length of spatial connections between regions for advancing industrial diversification.

Figure 4A presents the phase diagram that describes the ratio of active industries to all industries (S_a) as a function of the ratio of initially activated industries (in horizontal-axis) and the balance of core and periphery industries (in vertical-axis) when initially activating industries. Here, the balance index controls the strategy in determining which set of industries will be initially activated in the industry space (see Methods). Briefly, the balance value -1 means always selecting periphery-located industries, 0 means selecting industries by random, and 1 means always selecting core-located industries. From Figure 4A, we notice that the diagram is trivial when the ratio of initially activated industries is below 0.3 or above 0.8 as different strategies give almost the same results. However, an non-trivial area emerges in the middle of the diagram, where random selection (with the balance index being 0) leads to the full activation of all industries.

In particular, when the initial ratio is between 0.3 and 0.5, to initially active core-located or periphery-located industries is not competitive, because only part of industries can be finally activated (see Figure 4A) and the size of the giant component is relatively small (see Figure 4B). When the initial ratio is between 0.5 and 0.8, to initially active periphery-located industries is the worst, while to initially active core-located industries is the best because it has fewer time steps (NOI) (see Figure 4C) but full activation of all industries (see Figures 4A and 4B). The findings can be verified by using the bootstrap percolation model (see Figures S8A-C). In short, results suggest an optimal strategy, the strategy that makes a balance between core and periphery industries, to initially active industries for maximizing the range and minimizing the time of industrial diversification, benefiting most from the inter-industry learning effects.

Figure 4D presents the phase diagram of the inter-regional learning simulations. It shows the ratio of active regions (S_a) as a function of the ratio of initially activated regions (in horizontal-axis) and the balance of long and short distance (in vertical-axis) regions that are spatially connected. Here, the balance index is used to control the strategy in determining the length of newly added spatial connection between regions (see Methods). Briefly, the balance value -1 means always connecting to

Table 2. Interaction between inter-industry learning and inter-regional learning in developing new industries and keeping previous industries.

Independent Variables	Probit Model					
	Developing New Industry			Keeping Previous Industry		
	(1)	(2)	(3)	(4)	(5)	(6)
Density of Active Related Industries	2.9384***	3.6988***		3.3308***	3.2942***	
	(0.0378)	(0.0501)		(0.0368)	(0.0649)	
Density of Active Neighboring Regions	1.7240***	2.5772***		2.3785***	2.3449***	
	(0.0289)	(0.0481)		(0.0246)	(0.0537)	
Interaction Term 1		−6.9244***			0.2108	
		(0.3372)			(0.3019)	
Ratio of Active Related Industries			0.4874***			0.4230***
			(0.0118)			(0.0146)
Ratio of Active Neighboring Regions			0.8348***			0.9109***
			(0.0201)			(0.0162)
Interaction Term 2			−0.5815***			−0.2874***
			(0.0525)			(0.0421)
Constant	−2.7538***	−2.8484***	−2.3582***	0.2593***	0.2653***	1.0440***
	(0.0084)	(0.0095)	(0.0070)	(0.0100)	(0.0133)	(0.0069)
Observations	1,083,154	1,083,154	1,083,154	410,054	410,054	410,054
Robust R^2	0.0494	0.0511	0.0203	0.0877	0.0877	0.0329

Notes: Probit regressions consider both effects of inter-regional learning and inter-industry learning using different measures of density and ratio. Data are for the 2006-2013 period. Probit regressions include year-fixed effects. Robust standard errors are reported in parentheses. Significant level: $*p < 0.1$, $**p < 0.05$, and $***p < 0.01$.

nearby regions, 0 means connecting to regions randomly, and 1 means always connecting to distant regions. From Figure 4D, we find that the diagram is trivial when the ratio of initially activated regions is below 0.18 or above 0.24. However, non-trivial diagram emerges at the middle part, which shows that the random connecting strategy (with the balance index being around 0) and the distant-preferred strategy (with the balance index being above 0) give the full activation of all regions (see Figure 4D) and a full giant component (see Figure 4E).

Particularly interesting observations can be found when the initial ratio is between 0.18 and 0.21. Even though all strategies give full activation of regions (see Figures 4D and 4E), the distant-preferred strategy is the most efficient because it takes the shortest time (NOI) (see Figure 4F). The nearby-preferred strategy is the worst because it gives partial activation but takes long time. The findings can be verified as well by using the bootstrap percolation model (see Figures S8D-F). In short, results suggest an optimal strategy, the strategy that makes a balance between nearby and distant regions, in adding spatial connections between regions. In particular, the randomly connecting strategy (e.g., partially open long-range flights and build short-distance rails) costs relatively less but performs better than the distant-preferred strategy (only opening flights), benefiting most from the inter-regional learning effects.

Conclusions and Discussion

In this paper, we revealed the collective learning effects in regional industrial diversification by exploring Brazilian labor data in the period of 2006-2013. First, we mapped out the industry space based on the co-hiring of occupations and quantified its core-periphery structure. Then, we described how Brazilian regions learned to diversify into new industries in two ways: the evolution of competitive industries in the industry space for each region (the inter-industry learning channel) and the evolution of active regions for each industry (the inter-regional learning channel). Next, we used a graphical method to quantify the two learning effects by introducing two density measures of active related industries and active neighboring regions. We found that the probability of developing new industries or keeping previous industries increases strongly with both densities; however, their combination exhibits diminishing returns, meaning that the two learning channels are substitutes especially in developing new industries. The result suggests that the presence of either learning channel is sufficient for regional industrial diversification, and having both learning channels doesn't linearly increase the probability of developing new industries.

Moreover, we showed how to maximize the collective learning effects by using simulations under a simple threshold propagation process, where we considered the structural information of preexisting industries and the spatial locations of regions. For the inter-industry learning, we found an optimal strategy of balancing core and periphery industries in initially activating industries. Regions can benefit most from the inter-industry learning effects by adopting the optimal strategy in

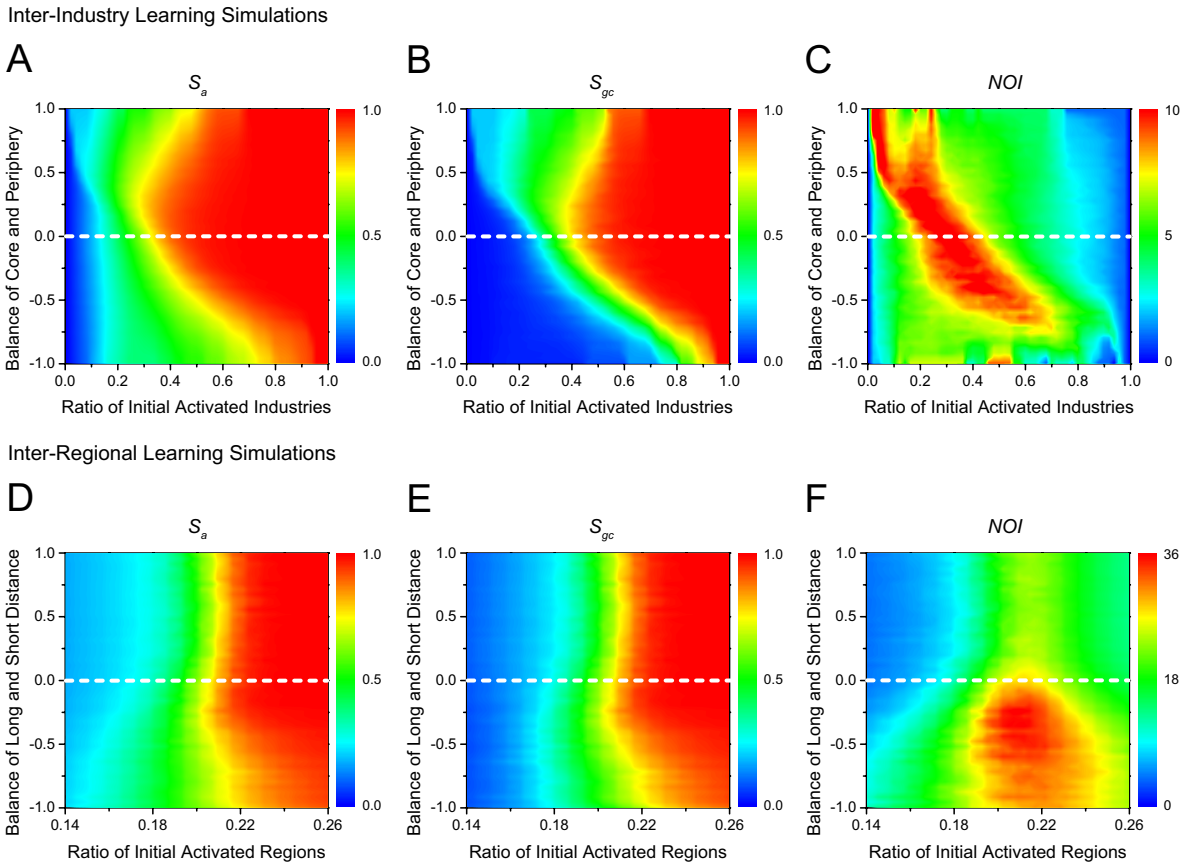

Figure 4. Simulation results on maximizing the two collective learning effects. Inter-industry learning: (A), (B) and (C) show respectively the ratio of active nodes (S_a), the ratio of giant component of active nodes (S_{gc}) and the number of iterations (NOI) before reaching the steady state. The horizontal-axis is the ratio of initially activated industries, and the vertical-axis is the balance index, which controls the strategy of selecting periphery-located (-1), random-located (0) and core-located (1) industries in the industry space for initial activation. Inter-regional learning: (D), (E) and (F) show respectively S_a, S_{gc} and NOI. The horizontal-axis is the ratio of initially activated regions, and the vertical-axis is the balance index, which controls the strategy of connecting nearby (-1), random (0) and distant (1) regions.

choosing initial industries at an early stage of development. Analogously, for the inter-regional learning, we found an optimal strategy of balancing nearby and distant regions in establishing new spatial connections. In particular, the randomly connecting strategy instead of distant-preferred strategy performs best by give the full activation of all regions but relative low operating cost. Our work suggests promising applications for regions with arbitrary conditions to make a better use of the collective learning effects in advancing industrial diversification.

The explorations of collective learning in modeling economic development remain open and our analysis should be interpreted in light of its limitations. For example, the used Brazilian labor dataset covers only a relatively short period, which makes it hard to capture the whole picture of rapid industrial transformations like the case of China.[22,30] There is still no widely accepted method that can best identifies if a region successfully developed a new industry,[41] even though we verified our results by using two methods: the number of workers and the revealed comparative advantages. The proximity between industries might be not only measured by the co-hiring of occupations, but also cross-validated by the spatial co-location method[22] and other vertex similarity measures.[42,43] Moreover, the threshold propagation model is too simple to reproduce the complex and realistic industrial diversification, even though it is a representative of various adoption, activation and spreading processes in complex networks.[19,20]

These aforementioned respects urge future works towards addressing these limitations and digging deeper into this topic. For example, after revealing the collective learning effects in regional industrial development at the macro-level, the possible next step is to explore the specific learning channels at the micro-level, such as learning-by-hiring[44] and labour mobility.[45] This exploration will contribute to a more comprehensive understanding and explanation of underlying mechanisms of collec-

tive learning in economic development. Besides, simulation results suggest the existence of optimal strategies in maximizing collective learning effects. Therefore, it will be worth it to develop new models that incorporate both learning effects and to explore the optimal strategic diffusion[35] in economic development for real-word applications.

Methods

Revealed comparative advantage. We define the revealed comparative advantage (RCA) measure following Balassa.[34] For industry and occupation, we define the $\mathrm{RCA}_{i,\alpha,t}$ of industry α in occupation i at year t using the ratio between the observed number of workers and the expected number. Formally, $\mathrm{RCA}_{i,\alpha,t}$ is given by:

$$\mathrm{RCA}_{i,\alpha,t} = \frac{x_{i,\alpha,t}}{\sum_{\alpha} x_{i,\alpha,t}} \Big/ \frac{\sum_{i} x_{i,\alpha,t}}{\sum_{\alpha}\sum_{i} x_{i,\alpha,t}}, \tag{1}$$

where $x_{i,\alpha,t}$ is the number of workers in industry α working in occupation i at year t. We say occupation i has competitiveness in industry α at year t if $\mathrm{RCA}_{i,\alpha,t} \geq 1$. Analogously, for region and industry, we define the $\mathrm{RCA}_{i,\alpha,t}$ of region i in industry α at year t using the ratio between the observed number of workers and the expected number,[22] which is also given by Eq. (1). We say industry α has competitiveness in region i at year t if $\mathrm{RCA}_{i,\alpha,t} \geq 1$.

Proximity and industry space. First, we map the labor data to an "Industry-Occupation" bipartite network (see Figure S1B), where the link weight $x_{i,\alpha}$ is the number of workers in industry α and in occupation i. Then, with the RCA calculated by Eq. (1), we measure the proximity between two industries using the cosine similarity index.[22] Formally, let $x_{i,\alpha,t} = ln(\mathrm{RCA}_{i,\alpha,t} + 1)$ and $x_{i,\beta,t} = ln(\mathrm{RCA}_{i,\beta,t} + 1)$. The proximity between industries α and β at year t is given by:

$$\phi_{\alpha,\beta,t} = \frac{\sum_{i} x_{i,\alpha,t} x_{i,\beta,t}}{\sqrt{\sum_{i} (x_{i,\alpha,t})^2}\sqrt{\sum_{i} (x_{i,\beta,t})^2}}. \tag{2}$$

Finally, the industry space is built based on the proximity matrix, following the three steps: (i) Maximum spanning network. All nodes are connected by the minimum number of links with the maximum weights. (ii) Maximum weighted network. A number of links with the maximum weights are included. Here, the number of included links is six times of nodes. (iii) Superposed network. By overlapping the maximum spanning network and the maximum weighted network, a superposed network is built and visualized as the industry space by using the ForceAtlas and Fruchterman-Reingold layout methods.

Core-periphery value of networks. We use a simplified variant of core-periphery (CP) measure to quantify the CP structure of networks.[33] The used method is based on the k-core decomposition method,[46] which progressively prunes a network by recursively removing nodes that are with at least number of degree and assigns them a "coreness" value k_s if they are put in the shell where nodes have at least k_s degree. Considering that a network with a genuine CP structure will have many nodes with small k_s value (periphery) and little nodes with large k_s value (core),[33] we define the CP value as

$$\lambda = (\tau_{max} - \tau_{min})\frac{S_{\tau_{min}}}{S_{\tau_{max}}}, \tag{3}$$

where $S_{\tau_{min}}$ and $S_{\tau_{max}}$ are respectively the number of nodes that have the maximal k_s value τ_{max} and the minimal k_s value τ_{min}.

Density of active related industries. For each industry in a region, the density of active related industries measures how many related industries that are already active in that region.[22,41] Here, the active industries are identified by $\mathrm{RCA} \geq 1$. Formally, the density $\omega_{i,\alpha,t}$ of active related industries for industry α in region i at year t is given by

$$\omega_{i,\alpha,t} = \frac{\sum_{\beta} \phi_{\alpha,\beta,t} U_{i,\beta,t}}{\sum_{\beta} \phi_{\alpha,\beta,t}}, \tag{4}$$

where $\phi_{\alpha,\beta,t}$ is the proximity between industries α and β, and $U_{i,\beta,t}$ takes the value of 1 if $\mathrm{RCA}_{i,\beta,t} \geq 1$ and 0 otherwise.

Density of active neighboring regions. For each region in an industry, the density of active neighboring regions measures how many neighboring regions that are already active in that industry.[22,29] The active regions are also identified by $\mathrm{RCA} \geq 1$. Formally, the density $\Omega_{i,\alpha,t}$ of active neighboring regions for region i in industry α at year t is given by

$$\Omega_{i,\alpha,t} = \sum_{j} \frac{U_{j,\alpha,t}}{D_{i,j}} \Big/ \sum_{j} \frac{1}{D_{i,j}}, \tag{5}$$

where $D_{i,j}$ is the geographic distance between regions i and j, and $U_{j,\alpha,t}$ takes the value of 1 if $\mathrm{RCA}_{j,\alpha,t} \geq 1$ and 0 otherwise.

Regression model for each learning channel. Probit model is used to explain the probability for region i to develop new industry α in the future with controlling for the effects of the number of regions with RCA ≥ 1 in industry α ($M_{\alpha,t} = \sum_i U_{i,\alpha,t}$) and the number of industries with RCA ≥ 1 in region i ($N_{i,t} = \sum_\alpha U_{i,\alpha,t}$). The empirical specification is given by

$$U_{i,\alpha,t+2} = \beta_0 + \beta_1 \omega_{i,\alpha,t} + \beta_2 \Omega_{i,\alpha,t} + \beta_3 M_{\alpha,t} + \beta_4 N_{i,t} + \mu_t + \varepsilon_{i,\alpha,t}, \tag{6}$$

where $U_{i,\alpha,t+2}$ ($U_{i,\alpha,t}$) takes the value of 1 if $RCA_{i,\beta,t+2} \geq 1$ ($RCA_{i,\beta,t} \geq 1$) and 0 otherwise. The $\omega_{i,\alpha,t}$ and $\Omega_{i,\alpha,t}$ are densities of active related industries and active neighboring regions for industry α and province i at year t, respectively. μ_t are the year-fixed effects, and $\varepsilon_{i,\alpha,t}$ is the error term.

Regression model for both learning channels. Probit model is used to explain the probability for a region to develop a new industry (i.e., U jumps to 1 from 0) or keep a previous industry (i.e., U sustains 1) in the future with the consideration of both learning effects: the inter-industry density ω and the inter-regional density Ω. The empirical specification is given by

$$U_{i,\alpha,t+2} = \beta_0 + \beta_1 \omega_{i,\alpha,t} + \beta_2 \Omega_{i,\alpha,t} + \beta_3 \omega_{i,\alpha,t} \Omega_{i,\alpha,t} + \mu_t + \varepsilon_{i,\alpha,t}, \tag{7}$$

where $\omega_{i,\alpha,t}\Omega_{i,\alpha,t}$ is the interaction term between the two densities, μ_t are the year-fixed effects, and $\varepsilon_{i,\alpha,t}$ is the error term.

Threshold propagation model. A simple threshold propagation process,[19] similar to the bootstrap percolation,[47] is used to simulate the industrial diversification. The process works as follows: (i) Nodes are in either active or inactive status; (ii) A node remains active once it is activated; (iii) A given ratio of nodes (p) are initially activated; (iv) Inactive nodes become active if their q fraction of neighbors are already active; (v) Nodes are activated in an iterative manner until no new nodes can be activated.[20] Here, the threshold is set as $q = 0.5$, which asks at least half of active neighbors to trigger the activation.

Balance index in simulations. For the inter-industry learning, the initially activated industries are selected with ratio p according to the balance index (q) of core and periphery for the fixed industry space as illustrated in Figure 1A. First, a randomized industry list is generated. Then, q ratio of industries are randomly selected from the list and rearranged by their coreness in descending order (to generate the case that q varies from 0 to 1) or in ascending order (to generate the case that q varies from 0 to -1). Finally, p ratio top-listed industries in the rearranged list are selected to be initially activated. The balance index $q = 1$ and $q = -1$ correspond to initially activating core-located and periphery-located industries, respectively.

For the inter-regional learning, the adjacent networks of regions are built by adding a new spatial connection to each region according to the balance index (Q) of long and short distance, and the initially activated regions are randomly selected. For each region, a random distance r between 2 and $d/2$ is generated with probability $P(r) \sim r^{5Q}$,[20] where d is the maximal neighboring distance between the considered region and all the other regions. The decay parameter $5Q$ is set to satisfy the boundary conditions, where all added links should be the longest or the shortest. The neighboring distance between two regions is defined as the minimum number of regions that one region has to cross, in order to reach the target region based on the original adjacent network of regions.[22] Then, one region with the neighboring distance r to the considered region is randomly selected to establish the new spatial connection. Finally, such procedure is repeated for the rest regions to finalize the new adjacent network of regions, where each region has an undirected spatial connection to the other region.

References

1. Schweitzer, F. *et al.* Economic networks: The new challenges. *Science* **325**, 422–425 (2009).
2. Youn, H. *et al.* Scaling and universality in urban economic diversification. *Journal of The Royal Society Interface* **13**, 20150937 (2016).
3. Lucas, R. E. On the mechanics of economic development. *Journal of Monetary Economics* **22**, 3–42 (1988).
4. Eagle, N., Macy, M. & Claxton, R. Network diversity and economic development. *Science* **328**, 1029–1031 (2010).
5. Hidalgo, C. A. Disconnected, fragmented, or united? A trans-disciplinary review of network science. *Applied Network Science* **1**, 6 (2016).
6. Henderson, J. V., Storeygard, A. & Weil, D. N. Measuring economic growth from outer space. *American Economic Review* **102**, 994–1028 (2012).
7. Varian, H. R. Big data: New tricks for econometrics. *Journal of Economic Perspectives* **28**, 3–27 (2014).
8. Tomasello, M. V., Perra, N., Tessone, C. J., Karsai, M. & Schweitzer, F. The role of endogenous and exogenous mechanisms in the formation of R&D networks. *Scientific Reports* **4**, 5679 (2014).
9. Einav, L. & Levin, J. Economics in the age of big data. *Science* **346**, 1243089 (2014).
10. Gao, J. & Zhou, T. Big data reveal the status of economic development. *Journal of the University of Electronic Science and Technology of China* **45**, 626–633 (2016).

11. Hidalgo, C. A. & Hausmann, R. The building blocks of economic complexity. *Proceedings of the National Academy of Sciences, USA* **106**, 10570–10575 (2009).

12. Gao, J. & Zhou, T. Quantifying China's regional economic complexity. *arXiv:1703.01292* (2017).

13. Jean, N. *et al.* Combining satellite imagery and machine learning to predict poverty. *Science* **353**, 790–794 (2016).

14. Steele, J. E. *et al.* Mapping poverty using mobile phone and satellite data. *Journal of The Royal Society Interface* **14**, 20160690 (2017).

15. Liu, J.-H., Wang, J., Shao, J. & Zhou, T. Online social activity reflects economic status. *Physica A* **457**, 581–589 (2016).

16. Barabási, A.-L. *Network Science* (Cambridge University Press, 2016).

17. Schulz, M. *Statistical Physics and Economics: Concepts, Tools, and Applications* (Springer Science & Business Media, 2003).

18. Hidalgo, C. A., Klinger, B., Barabási, A.-L. & Hausmann, R. The product space conditions the development of nations. *Science* **317**, 482–487 (2007).

19. Watts, D. J. A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences, USA* **99**, 5766–5771 (2002).

20. Gao, J., Zhou, T. & Hu, Y. Bootstrap percolation on spatial networks. *Scientific Reports* **5**, 14662 (2015).

21. Caraglio, M., Baldovin, F. & Stella, A. L. Export dynamics as an optimal growth problem in the network of global economy. *Scientific Reports* **6**, 31461 (2016).

22. Gao, J., Jun, B., Pentland, A., Zhou, T. & Hidalgo, C. A. Collective learning in China's regional economic development. *arXiv:1703.01369* (2017).

23. Lawson, C. & Lorenz, E. Collective learning, tacit knowledge and regional innovative capacity. *Regional Studies* **33**, 305–317 (1999).

24. Capello, R. Spatial transfer of knowledge in high technology milieux: Learning versus collective learning processes. *Regional Studies* **33**, 353–365 (1999).

25. Kao, A. B., Miller, N., Torney, C., Hartnett, A. & Couzin, I. D. Collective learning and optimal consensus decisions in social animal groups. *PLoS Computational Biology* **10**, e1003762 (2014).

26. Neffke, F., Henning, M. & Boschma, R. How do regions diversify over time? Industry relatedness and the development of new growth paths in regions. *Economic Geography* **87**, 237–265 (2011).

27. Guo, Q. & He, C. Production space and regional industrial evolution in China. *GeoJournal* **82**, 379–396 (2017).

28. Bahar, D., Hausmann, R. & Hidalgo, C. A. Neighbors and the evolution of the comparative advantage of nations: Evidence of international knowledge diffusion? *Journal of International Economics* **92**, 111–123 (2014).

29. Boschma, R., Martín, V. & Minondo, A. Neighbour regions as the source of new industries. *Papers in Regional Science* **96**, 227–245 (2017).

30. Zhu, S., He, C. & Zhou, Y. How to jump further and catch up? Path-breaking in an uneven industry space. *Journal of Economic Geography* **17**, 521–545 (2017).

31. Boschma, R. & Iammarino, S. Related variety, trade linkages, and regional growth in Italy. *Economic Geography* **85**, 289–311 (2009).

32. Nelson, R. R. & Phelps, E. S. Investment in humans, technological diffusion, and economic growth. *American Economic Review* **56**, 69–75 (1966).

33. Verma, T., Russmann, F., Araújo, N., Nagler, J. & Herrmann, H. Emergence of core-peripheries in networks. *Nature Communications* **7**, 10441 (2016).

34. Balassa, B. Trade liberalisation and "revealed" comparative advantage. *Manchester School* **33**, 99–123 (1965).

35. Alshamsi, A., Pinheiro, F. L., & Hidalgo, C. A. When to target hubs? Strategic diffusion in complex networks. *arXiv:1705.00232* (2017).

36. Csermely, P., London, A., Wu, L.-Y. & Uzzi, B. Structure and dynamics of core/periphery networks. *Journal of Complex Networks* **1**, 93–123 (2013).

37. Vespignani, A. Modelling dynamical processes in complex socio-technical systems. *Nature Physics* **8**, 32–39 (2012).

38. Castellano, C. & Pastor-Satorras, R. Thresholds for epidemic spreading in networks. *Physical Review Letters* **105**, 218701 (2010).

39. Zheng, S. & Kahn, M. E. China's bullet trains facilitate market integration and mitigate the cost of megacity growth. *Proceedings of the National Academy of Sciences, USA* **110**, E1248–E1253 (2013).

40. Brueckner, J. K. Airline traffic and urban economic development. *Urban Studies* **40**, 1455–1469 (2003).

41. Boschma, R., Minondo, A. & Navarro, M. The emergence of new industries at the regional level in Spain: A proximity approach based on product relatedness. *Economic Geography* **89**, 29–51 (2013).

42. Lü, L. & Zhou, T. Link prediction in complex networks: A survey. *Physica A* **390**, 1150–1170 (2011).

43. Chen, L.-J., Zhang, Z.-K., Liu, J.-H., Gao, J. & Zhou, T. A vertex similarity index for better personalized recommendation. *Physica A* **466**, 607–615 (2017).

44. Song, J., Almeida, P. & Wu, G. Learning–by–hiring: When is mobility more likely to facilitate interfirm knowledge transfer? *Management Science* **49**, 351–365 (2003).

45. Petersen, A. M. & Puliga, M. High-skilled labour mobility in Europe before and after the 2004 enlargement. *Journal of The Royal Society Interface* **14**, 20170030 (2017).

46. Carmi, S., Havlin, S., Kirkpatrick, S., Shavitt, Y. & Shir, E. A model of internet topology using k-shell decomposition. *Proceedings of the National Academy of Sciences, USA* **104**, 11150–11154 (2007).

47. Baxter, G. J., Dorogovtsev, S. N., Goltsev, A. V. & Mendes, J. F. Bootstrap percolation on complex networks. *Physical Review E* **82**, 011103 (2010).

Acknowledgments

We thank Cristian Jara-Figueroa for sharing the cleaned dataset, Dominik Hartmann and Flávio Pinheiro for helpful discussions. Jian Gao acknowledges the China Scholarship Council for partial financial support.

Author contributions

J.G. and C.A.H. designed the research. J.G. executed the experiments and prepared the figures. All authors analyzed the results. J.G. and B.J. wrote the manuscript. All authors reviewed the manuscript.

Additional information

Supplementary information:

Supplementary information accompanies this paper.

Competing financial interests:

The authors declare no competing financial interests.

Data Availability:

All relevant data are within the paper and its Supplementary Information files.

Revealing and maximizing the collective learning effects in industrial diversification

(Supplementary Information)

Jian Gao, Bogang Jun, Tao Zhou, César A. Hidalgo

1 Brazilian Labor Data

The Brazilian labor data, named RAIS (Annual Social Information Report), was compiled by the Ministry of Labor and Employment (MET) of Brazil (http://www.rais.gov.br). The RAIS depicts the Brazilian formal market by annual data provided for all businesses based on their standing at the end of the previous year. Specifically, the RAIS data include the demographic, occupational, industrial and income characteristics of employees. For example, firm and occupation in which one worker works, region where a firm locates, and industry category in which a firm operates. Currently, RAIS covers about 97% of the Brazilian formal market according to the Brazilian Institute of Geography and Statistics (IBGE). The visualization of the RAIS data together with the Brazilian education and international trade data from 2002 to 2013 is provided by the DataViva platform (http://dataviva.info), which is the Brazil's largest platform for social and economic data search.

The used RAIS dataset covers the period of 2006-2013 and 76.62 Million workers with 501 Occupations in firms, locating in 558 Microregions (aggregated into 137 Mesoregions, 27 States and 5 Regions) and operating in 669 industry Classes (aggregated into 87 Divisions and 21 Sections). The classification of industries of the RAIS data set is according to the National Classification of Economic Activities (CNAE) of Brazil (see IBGE website http://www.ibge.gov.br). Due to the insufficient data that caused by the adoption of the CNAE new edition, we excluded four industries (at the Class-level) in our analysis. These industries are Manufacture of Military Combat Vehicles, Exchange Banks, Company's Head Offices, Local Administrative Units, and Exchange Banks. Basic statistics of the used data are shown in Table S1.

Table S1. Basic statistics of the used Brazilian labor data.

Year	# of Microregions	# of Industry Classes	# of Industry Sections	# of Occupations	# of Workers
2006	558	669	21	501	33.70M
2007	558	669	21	501	36.00M
2008	558	669	21	501	37.76M
2009	558	669	21	501	39.47M
2010	558	669	21	501	42.21M
2011	558	669	21	501	44.34M
2012	558	669	21	501	45.53M
2013	558	669	21	501	47.00M

2 Mapping Data to Networks

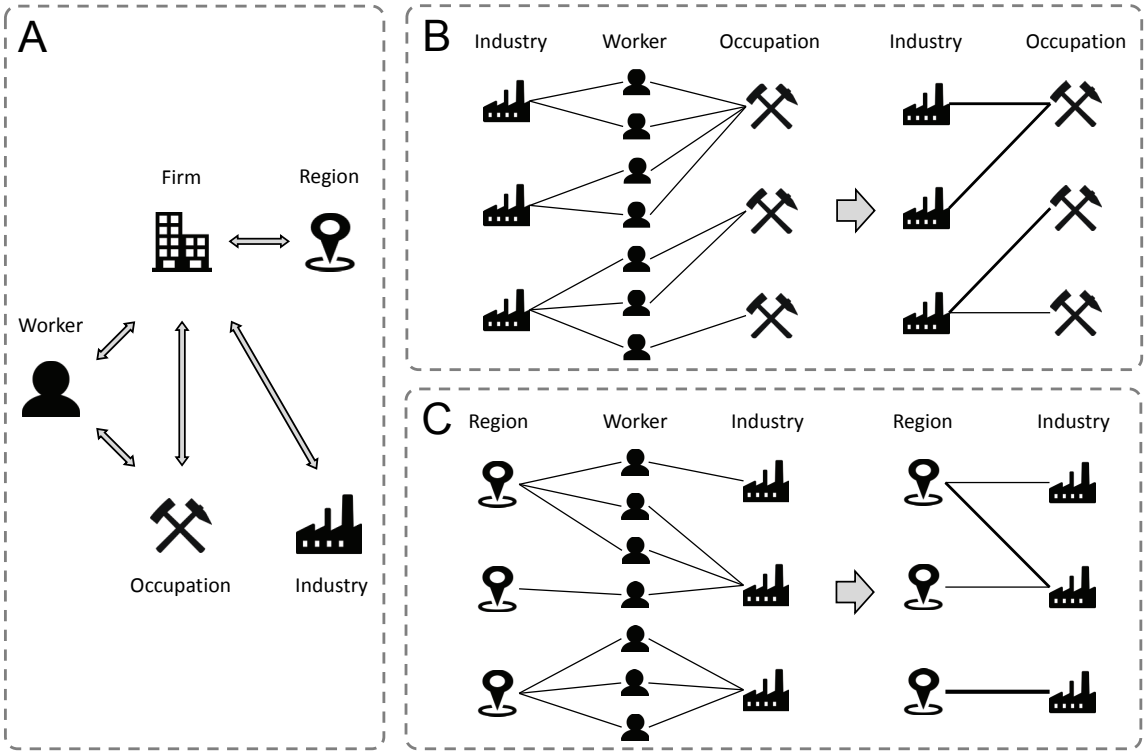

Figure S1. (A) Relationship between attributes of the used Brazilian labor data. (B) "Industry-Occupation" bipartite network. The weight of links is the number of workers working in the corresponding industry and occupation. (C) "Region-Industry" bipartite network. The weight of links is the number of workers in the corresponding region and industry.

Figure S1A illustrates the relationship between the used attributes of the Brazilian labor data. One worker takes one occupation and works in one firm, while one firm can hire workers in various occupations. One firm operates in one industry and locates in one region. Accordingly, by linking industries to occupations (and vice versa) through workers, we can build an "Industry-Occupation" network (see Figure S1B), where the weight of links is the number of workers in that occupation and industry. Similarly, by linking regions to industries (and vice versa) through workers, we can also build a "Region-Industry" network (see Figure S1C), where the weight of links is the number of workers in that region and industry.

3 Industry Space Based on Alternative Proximity Measures

The "co-hiring of occupations using the proximity index" between industries is based on the revealed comparative advantage (RCA), which is defined as

$$\mathrm{RCA}_{i,\alpha} = \frac{x_{i,\alpha}}{\sum_{\alpha} x_{i,\alpha}} \Big/ \frac{\sum_{i} x_{i,\alpha}}{\sum_{\alpha}\sum_{i} x_{i,\alpha}}, \tag{S1}$$

where $x_{i,\alpha}$ is the number of workers in occupation i and industry α. If $\mathrm{RCA}_{i,\alpha} \geq 1$, we say occupation i is effectively hired by industry α. Accordingly, the proximity $\phi_{\alpha,\beta}$ between industries α and β is defined as

$$\phi_{\alpha,\beta} = min\Big\{P(RCA_{\alpha}|RCA_{\beta}), P(RCA_{\beta}|RCA_{\alpha})\Big\}, \tag{S2}$$

where $P(RCA_{\alpha}|RCA_{\beta})$ is the conditional probability of effectively hired by industry α given occupations are effectively hired by industry j. The minimum value of both conditional probabilities is used.

The "cosine similarity by RCA ≥ 1" between industries is defined by two vectors summarizing the occupations that are effectively hired by two industries. Formally, the cosine similarity by RCA ≥ 1 between industries α and β is given by

$$\phi_{\alpha,\beta} = \frac{\sum_{i} x_{i,\alpha} x_{i,\beta}}{\sqrt{\sum_{i} (x_{i,\alpha})^2}\sqrt{\sum_{i} (x_{i,\beta})^2}}, \tag{S3}$$

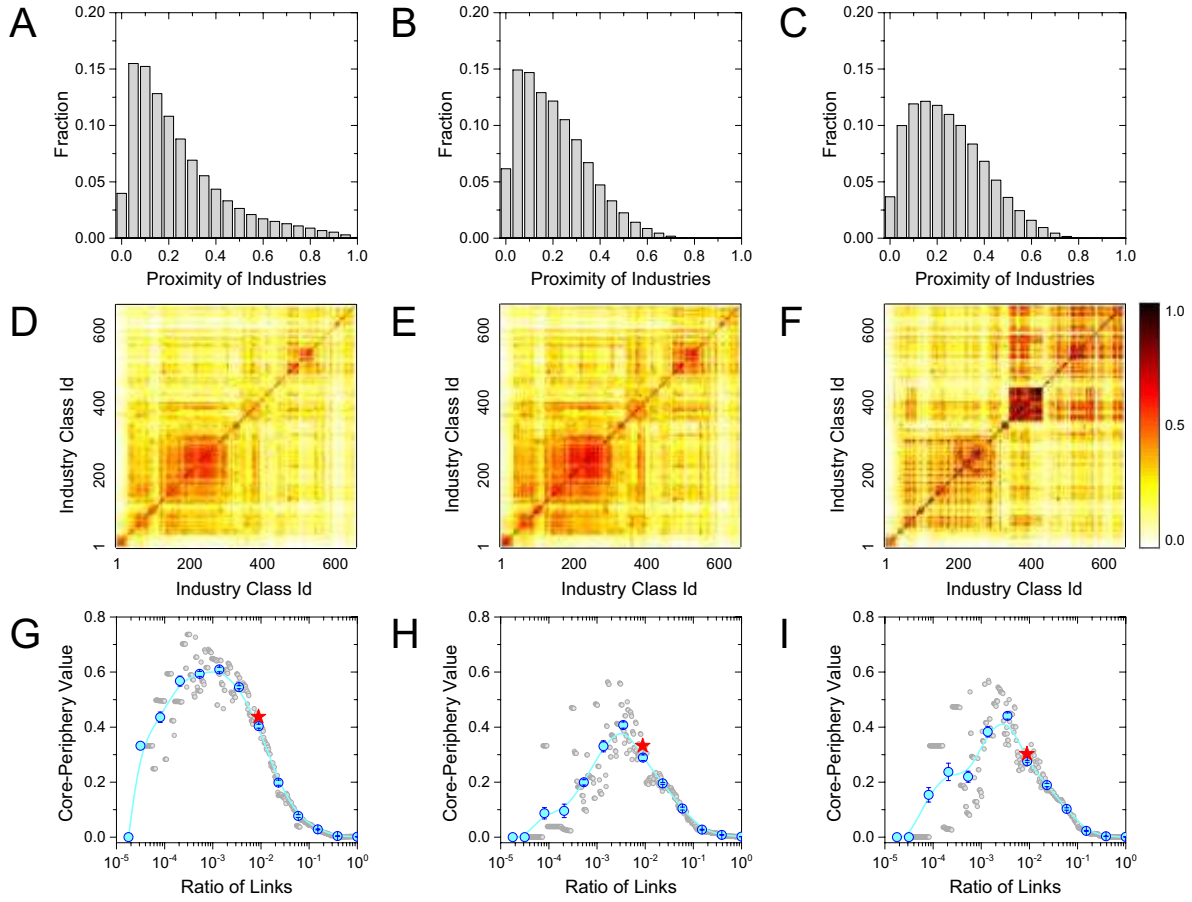

Figure S2. (A), (B) and (C) are density distributions of industrial proximity based on the "co-hiring of occupations using the proximity index", the "cosine similarity by RCA ≥ 1", and the "cosine similarity by number of workers", respectively. (D), (E) and (F) are the corresponding proximity matrix ordered by the industry Class Id referring to the CNAE, in which industries within the same aggregated category are nearby coded. The color marks the proximity value. (G), (H) and (I) are the core-periphery value of networks as a function of the ratio of added links. As marked by the red star, when the ratio of added links is 9.8×10^{-3}, the core-periphery value of the industry space is 0.332 for the "co-hiring of occupations using the proximity index", 0.303 for the "cosine similarity by RCA ≥ 1" and 0.438 for the "cosine similarity by number of workers".

where $x_{i,\alpha}$ takes the value of 1 if RCA ≥ 1 and 0 otherwise.

The "cosine similarity by number of workers" between industries is defined by two vectors summarizing the number of workers that work in two industries. Formally, the similarity number of workers between industries α and β is given by Eq. (S3), where $x_{i,\alpha}$ is the number of workers in industry α and occupation i.

Figures S2A, S2B and S2C present the density distributions of the proximity values using the "co-hiring of occupations using the proximity index", the "cosine similarity by RCA ≥ 1" and the "cosine similarity by number of workers", respectively. All distributions are log-normal like, with most values being small and some values being big in the long-tail.

Figures S2D, S2E and S2F present the corresponding proximity matrix ordered by the industry Class Id referring to the CNAE. The three figures share the similar pattern. Industries of the same industry Section (with the same color) tend to have high proximity with each other, suggesting the validation of the used proximity measures and the robustness of the findings.

Figures S2G, S2H and S2I present the corresponding function of core-periphery value as the increasing of the ratio of added links (to the complete proximity matrix). As marked by the red star, when the ratio of added links is 9.8×10^{-3} (the number of added links is about 6.5 times of the number of nodes), the core-periphery value of the industry space is 0.332 for the "co-hiring of occupations using the proximity index", 0.303 for the "cosine similarity by RCA ≥ 1" and 0.438 for the "cosine similarity by number of workers". The core-periphery value is relatively high for each proximity measure.

4 Illustrations of the Inter-Industry and Inter-Regional Learning

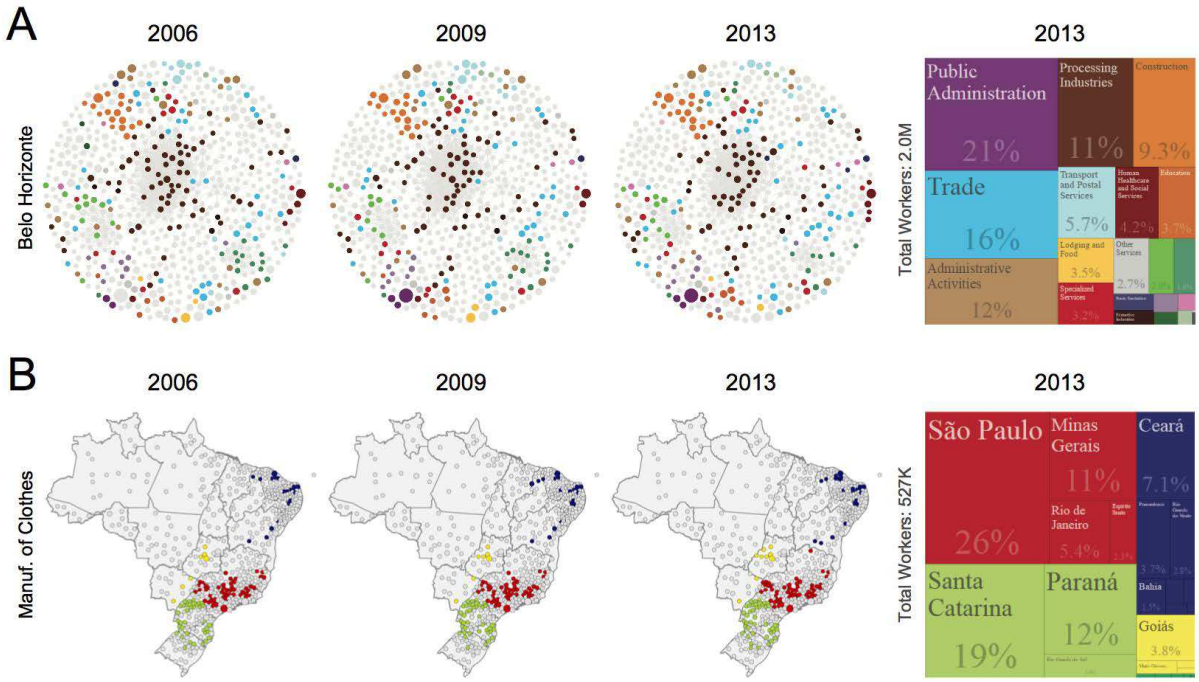

Figure S3. Illustrating the evolution of Brazilian industrial diversification in 2006, 2009 and 2013. (A) Inter-industry learning illustrations using microregions Bello Horizonte. Color circles in the space highlight comparative industries that one region has. Colored circles highlight comparative industries in 2006 and 2013. Tree map (right column) shows the percentage of market share (number of workers) of industry Sections in 2013. (B) Inter-regional learning illustrations using Manufacture of Clothes. Colored circles highlight geographic locations of microregions with competitiveness. Tree map (right column) shows the percentage of market share (number of workers) of Regions in 2013.

Figure S3A illustrates the evolution of industrial structure with active industries (highlighted by color circles in 2006, 2009 and 2013) and the market share (number of workers) of industries (colored by industry Sections in 2013) in Bello Horizorite. The industries with competitiveness are highly localized in the industry space. Industries that are surrounded by more active industries in the industry space are more likely to become active in the future, showing the evidence of inter-industry learning.

Figure S3B illustrates the evolution of active microregions (highlighted using color circles at geographic locations) for one industry and the market share (number of workers) of all regions in Manufacture of Clothes. Regions that geographically close located tend to be activated in the same industry. Inactive regions that are surrounded by many active neighboring regions in the same industry are more likely to become active in the future, showing the evidence of inter-regional learning.

5 Decay of Industrial Similarity with Distance

We measure the industrial similarity between regions using the cosine similarity of the vectors summarizing the RCA values of industries in each region. Formally, the industrial similarity $\varphi_{i,j}$ between regions i and j is given by

$$\varphi_{i,j} = \frac{\sum_\alpha y_{i,\alpha} y_{j,\alpha}}{\sqrt{\sum_\alpha (y_{i,\alpha})^2}\sqrt{\sum_\alpha (y_{j,\alpha})^2}}. \tag{S4}$$

where $y_{i,\alpha,t} = ln(RCA_{i,\alpha} + 1)$. The $\mathrm{RCA}_{i,\alpha}$ of region i in industry α is defined as $\mathrm{RCA}_{i,\alpha} = \frac{x_{i,\alpha}}{\sum_\alpha x_{i,\alpha}} \Big/ \frac{\sum_i x_{i,\alpha}}{\sum_\alpha \sum_i x_{i,\alpha}}$, where $x_{i,\alpha}$ is the number of workers in industry α in region i.

Figure S4A presents the industrial similarity distributions for neighboring regions (in pink) and non-neighboring regions (in blue) for the period of 2006-2013. The industrial similarity of neighboring regions is significantly larger than that of non-neighboring regions. Figure S4B presents the industrial similarity as a function of geographic distance. Closer regions have higher industrial similarly, and the industrial similarly highly decays with distance.

Figure S5A presents the industrial similarity between regions as a function of their geographic distance for each industry Section, and Figure S5B shows the number of workers. Two main observations are notable. On the one hand, for most industry

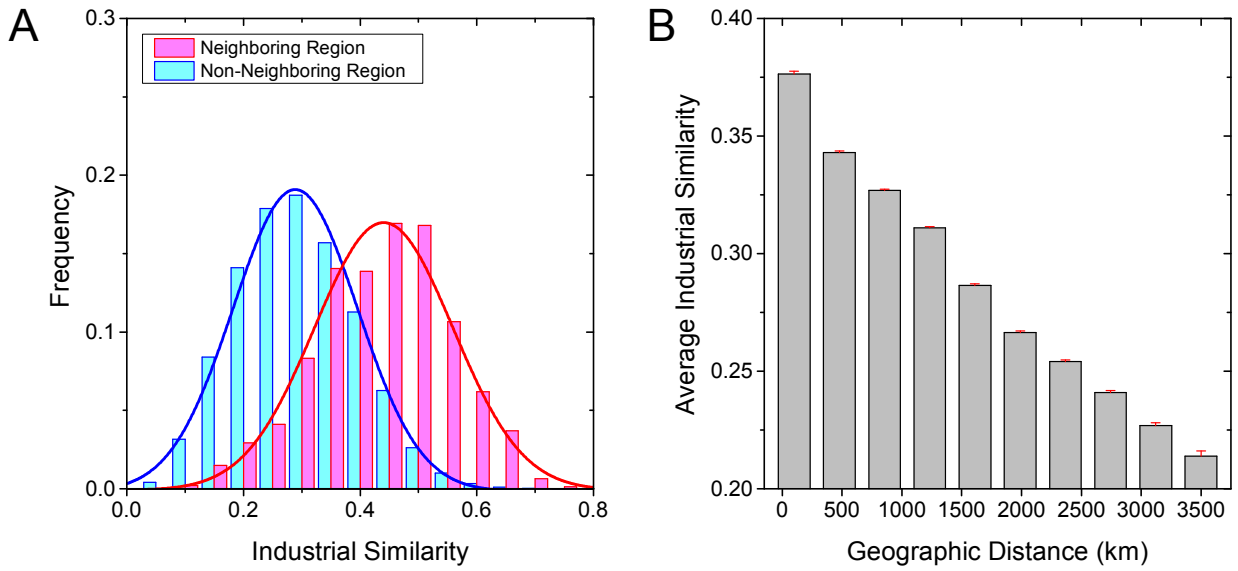

Figure S4. (A) Distribution of industrial similarity between pairs of neighboring regions (in pink) and non-neighboring regions (in blue). Curves show the normal fits. (B) Industrial similarity between regions as a function of their geographic distance across all industries. Bars correspond to the average industrial similarity. Error bars correspond to the standard errors.

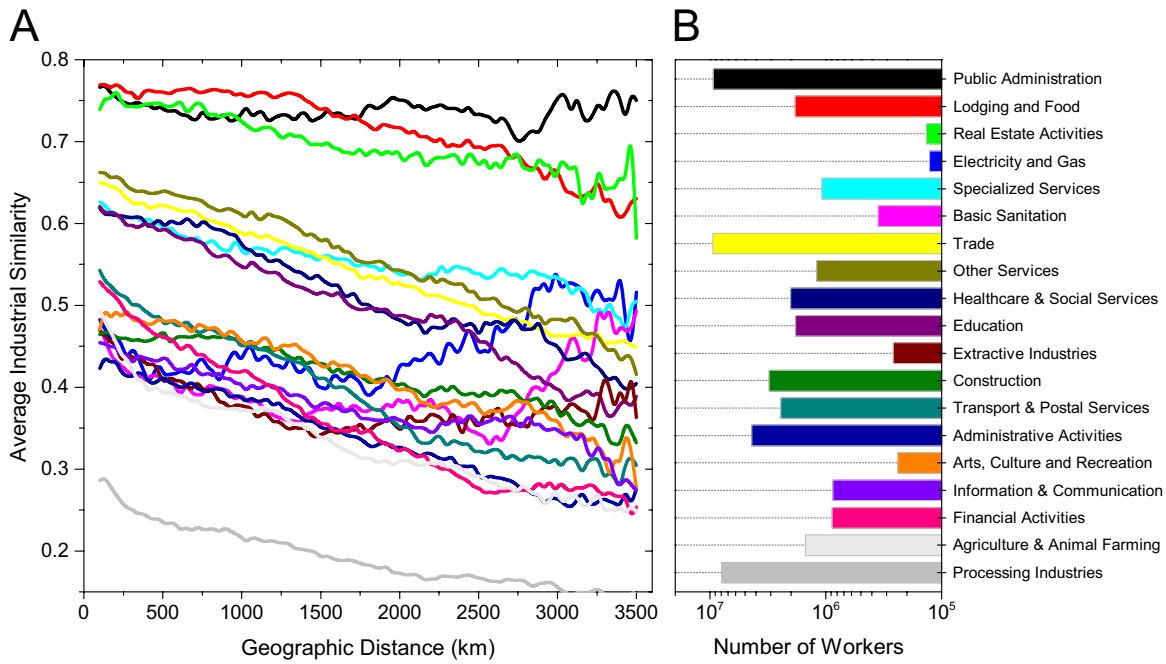

Figure S5. (A) Industrial similarity between regions as a function of their geographic distance. In calculating the industrial similarity, industries are grouped at the industry Section level. Data are divided into bins by geographic distance, and the average industrial similarity in each bin is shown. (B) Number of workers in each industry. Results are for the year 2013.

Sections, the average industrial similarity of regions decays linearly with distance, and the curves share almost the same slope. On the other hand, different industries have different similarity at the same distance, suggesting that neighboring regions have different levels of collaboration and competition for different industries.

6 Effects of Related Industries and Neighboring Regions

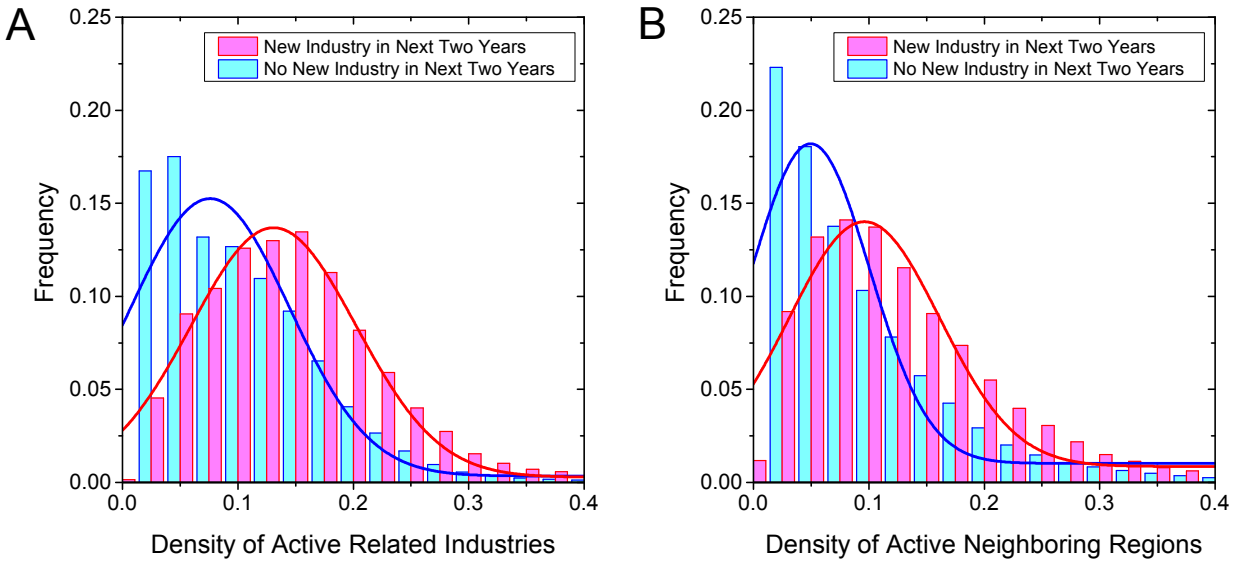

Figure S6. Effects of related industries and neighboring regions. (A) and (B) are respectively the distributions of the densities of active related industries and active neighboring regions for each pair of regions and industries. The pink and blue distributions respectively focus on pairs of regions and industries that developed new industries and pairs of regions and industries that didn't develop new industries in next two years. The mean value of the pink distribution is significantly larger than that of blue. Results show averages for 2006-2013 using two-year intervals with one-year backward and forward conditions.

Figure S6A presents the distribution of the density of active related industries for pairs of industries and provinces that developed new industries in next two years (in pink) and pairs of industries and provinces that didn't develop new industries (in blue) in next two years. The distributions show that, on average, the density of an industry in the provinces who developed that industry two years later is significantly larger than in those who didn't developed that industry two years later. The result suggests the effects of related industries in developing new industries, showing the evidence of the inter-industry learning.

Figure S6B presents the distribution of the density of active neighboring regions for pairs of industries and provinces that developed new industries in next two years (in pink) and pairs of industries and provinces that didn't develop new industries (in blue) in next two years. The distributions show that, on average, the density of neighboring regions in one industry for the region who developed that new industry two years later is significantly larger than for those that did not. The result suggests the effects of neighboring regions in developing new industries, showing the evidence of the inter-regional learning.

7 Robustness Check of the Two Learning Channels

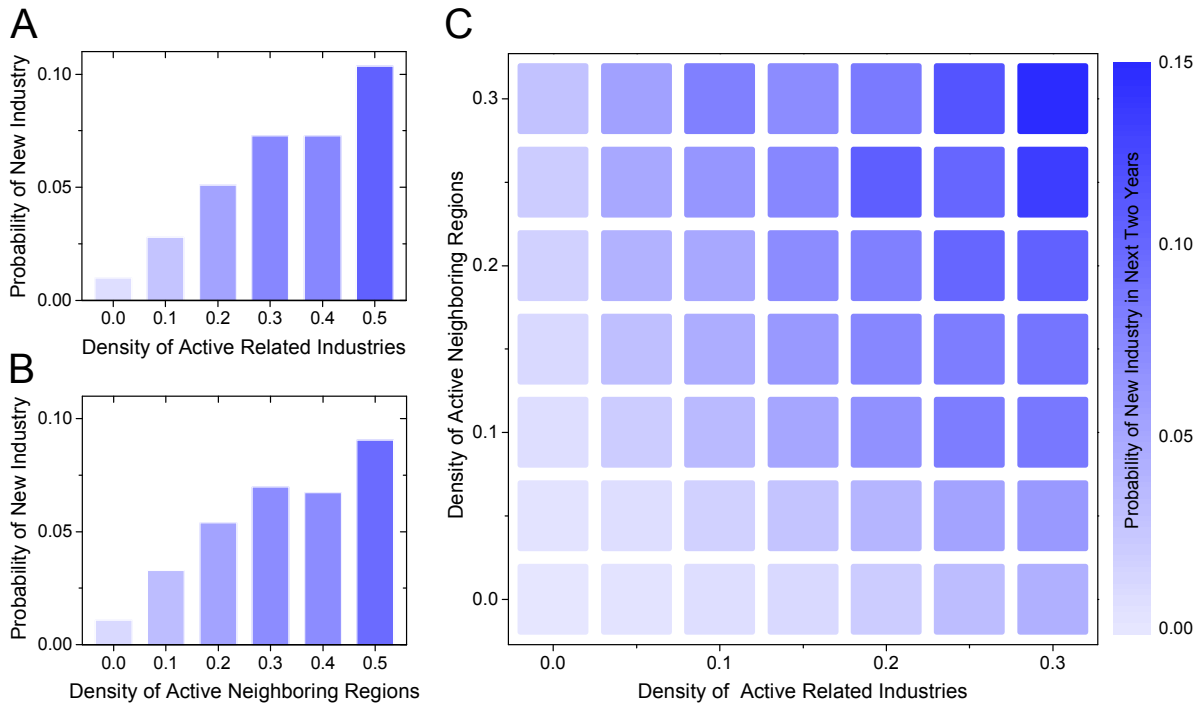

Figure S7. Collective learning in industrial diversification using a graphical method. The identification of developing new industries is based on the RCA ≥ 1. (A) Inter-industry learning: Probability of developing a new industry in a region in next two years as a function of the density of active related industries. (B) Inter-regional learning: Probability of developing a new industry in a region in next two years as a function of the density of active neighboring regions. (C) Combining two learning channels: Joint probability of developing a new industry in a region in next two years given the density of active related industries in horizontal-axis and the density of active neighboring regions in vertical-axis. Results are averaged for 2006-2013.

To check the robustness, we use a soft condition to identify the developing of new industries for one region in one industry. In terms of RCA ≥ 1, we restrict the development to two conditions: 1) Presence. The RCA value of a region in one industry should be below one before the period and at least one after the period. 2) Stability. A backward condition requires the region to have RCA < 1 in the industry one year prior to the beginning of the period, and a forward condition requires the region to sustain RCA ≥ 1 for one year after the end of the period. The additional restrictions are to filter out some temporal activation.

Figure S7A presents the inter-industry learning results. We find the probability for a region to develop a new industry in next two years increases strongly with the density of active related industries. In other words, industries that will become active in next two years have higher density of active related industries. Figure S7B presents the inter-regional learning results. We find the increasing tend in the probability for a region to develop a new industry in next two years as a function of the density of active neighboring regions. In other words, regions that will become active in next two years have higher density of active neighboring regions. Figure S7C combines both learning channels, to test whether the two collective learning channels work together. Results show that the probability of new industry development increases with both, the density of active related industries (in horizontal-axis) and the density of active neighboring provinces (in vertical-axis).

8 Regression Results and Variable Summary Statistics

Table S2. Summary statistics for developing new industries.

Variable	Obs	Mean	Std. Dev.	Min	Max
Jump in Next Two Years	1,083,154	0.0132	0.1140	0	1
Density of Active Related Industries	1,083,154	0.1010	0.0691	0.0007	0.5760
Density of Active Neighboring Regions	1,083,154	0.0908	0.0812	0.0006	0.8260
Interaction Term 1	1,083,154	0.0098	0.0125	0.0000	0.1800
Ratio of Active Related Industries	1,083,154	0.0994	0.2330	0	1
Ratio of Active Neighboring Regions	1,083,154	0.0749	0.1440	0	1
Interaction Term 2	1,083,154	0.0124	0.0554	0	1
Number of Active Related Industries	1,083,154	0.4610	1.0920	0	34
Number of Active Neighboring Regions	1,083,154	0.4020	0.7700	0	8
Interaction Term 3	1,083,154	0.3180	1.6590	0	136
Number of Related Industries	1,083,154	6.5770	9.1430	1	62
Number of Neighboring Regions	1,083,154	5.4730	1.5200	0	11
Interaction Term 4	1,083,154	35.960	52.940	0	682
Number of Active Provinces in Industry	1,083,154	53.240	45.400	1	421
Number of Active Industries in Region	1,083,154	72.680	44.080	4	310

Table S3. Correlations for developing new industries: Single channel

Variable	Jump	Density_I	Density_R	Num_I	Num_R
Jump (Jump in Next Two Years)	1				
Density_I (Density of Active Related Industries)	0.0741	1			
Density_R (Density of Active Neighboring Regions)	0.0556	0.104	1		
Num_I (Number of Active Provinces in Industry)	0.0396	−0.0665	0.8833	1	
Num_R (Number of Active Industries in Region)	0.0651	0.9377	0.0011	−0.1827	1

Table S4. Summary statistics for keeping previous industries.

Variable	Obs	Mean	Std. Dev.	Min	Max
RCA in Next Two Years	410,054	0.9180	0.2750	0	1
Density of Active Related Industries	410,054	0.1960	0.0960	0.0009	0.6950
Density of Active Neighboring Regions	410,054	0.2430	0.1640	0.0000	0.9030
Interaction Term 1	410,054	0.0469	0.0408	0.0000	0.4790
Ratio of Active Related Industries	410,054	0.2740	0.3090	0	1
Ratio of Active Neighboring Regions	410,054	0.2780	0.2830	0	1
Interaction Term 2	410,054	0.0879	0.1600	0	1
Number of Active Related Industries	410,054	1.6430	2.8290	0	49
Number of Active Neighboring Regions	410,054	1.5550	1.6380	0	11
Interaction Term 3	410,054	3.1080	8.9630	0	264
Number of Related Industries	410,054	6.5460	7.9580	1	62
Number of Neighboring Regions	410,054	5.6030	1.5730	0	11
Interaction Term 4	410,054	36.780	47.530	0	660
Number of Active Provinces in Industry	410,054	120.50	84.350	1	423
Number of Active Industries in Region	410,054	121.10	55.830	4	310

Table S5. Correlations for keeping previous industries: Single channel

Variable	RCA	Density_I	Density_R	Num_I	Num_R
RCA (RCA in Next Two Years)	1				
Density_I (Density of Active Related Industries)	0.1350	1			
Density_R (Density of Active Neighboring Regions)	0.1456	−0.0522	1		
Num_I (Number of Active Provinces in Industry)	0.1228	−0.1571	0.8998	1	
Num_R (Number of Active Industries in Region)	0.1134	0.9176	−0.1761	−0.2947	1

Table S6. Probit regression results for inter-industry learning and inter-regional learning in keeping previous industries.

Keeping Previous Industry	Probit Model					
	Density of Active Related Industries			Density of Active Neighboring Regions		
	(1)	(2)	(3)	(4)	(5)	(6)
Density	3.0256***	3.8141***	2.0071***	2.1995***	2.6172***	1.5496***
	(0.0355)	(0.0388)	(0.0976)	(0.0235)	(0.0251)	(0.0458)
Number of Active Provinces in Industry		0.0046***	0.0050***			0.0025***
		(0.0000)	(0.0001)			(0.0001)
Number of Active Industries in Region			0.0034***		0.0056***	0.0062***
			(0.0002)		(0.0001)	(0.0001)
Constant	0.8040***	0.1900***	0.0897***	0.8863***	0.1801***	0.0801***
	(0.0082)	(0.0106)	(0.0119)	(0.0073)	(0.0106)	(0.0117)
Observations	410,054	410,054	410,054	410,054	410,054	410,054
Pseudo R^2	0.0367	0.0846	0.0865	0.0468	0.0861	0.0887

Notes: Probit regressions on the probability of keeping previous industries as a function of the density of active related industries or the density of active neighboring regions with controlling for the effects of both the number of active industries that one region has and the number of regions in which an industry is active. Data is for the 2006-2013 period. Probit regressions include year-fixed effects. Robust standard errors are reported in parentheses. Significant level: $*p < 0.1$, $**p < 0.05$, and $***p < 0.01$.

Table S7. Correlations for developing new industries: Combining two channels

Variable	Jump	Industry	Region	Interaction
Jump in Next Two Years	1			
Density of Active Related Industries	0.0741	1		
Density of Active Neighboring Regions	0.0556	0.1040	1	
Interaction Term 1	0.0750	0.5743	0.7370	1
Ratio of Active Related Industries	0.0433	1		
Ratio of Active Neighboring Regions	0.0455	0.1458	1	
Interaction Term 2	0.0374	0.5294	0.4953	1
Number of Active Related Industries	0.0568	1		
Number of Active Neighboring Regions	0.0476	0.158	1	
Interaction Term 3	0.0404	0.6045	0.4339	1
Number of Related Industries	0.0030	1		
Number of Neighboring Regions	0.0089	−0.0027	1	
Interaction Term 4	0.0044	0.9449	0.1884	1

Table S8. Correlations for keeping previous industries: Combining two channels

Variable	Jump	Industry	Region	Interaction
RCA in Next Two Years	1			
Density of Active Related Industries	0.1350	1		
Density of Active Neighboring Regions	0.1456	−0.0522	1	
Interaction Term 1	0.1661	0.5135	0.7368	1
Ratio of Active Related Industries	0.0703	1		
Ratio of Active Neighboring Regions	0.1140	0.1319	1	
Interaction Term 2	0.0852	0.648	0.6019	1
Number of Active Related Industries	0.0648	1		
Number of Active Neighboring Regions	0.1141	0.1192	1	
Interaction Term 3	0.0646	0.7612	0.4085	1
Number of Related Industries	−0.0024	1		
Number of Neighboring Regions	0.0064	0.0082	1	
Interaction Term 4	−0.0009	0.9441	0.2177	1

Table S9. Interaction between inter-industry learning and inter-regional learning using alternative definitions of density.

Independent Variables	Probit Model					
	Developing New Industry			Keeping Previous Industry		
	(1)	(2)	(3)	(4)	(5)	(6)
Number of Active	0.0850***	0.1168***		0.0631***	0.0758***	
Related Industries	(0.0019)	(0.0023)		(0.0019)	(0.0025)	
Number of Active	0.1338***	0.1646***		0.1520***	0.1639***	
Neighboring Regions	(0.0033)	(0.0035)		(0.0021)	(0.0026)	
Interaction Term 3		−0.0300***			−0.0089***	
		(0.0017)			(0.0010)	
Number of			0.0022*			−0.0007
Related Industries			(0.0013)			(0.0014)
Number of			0.0208***			0.0072***
Neighboring Regions			(0.0027)			(0.0024)
Interaction Term 4			−0.0002			0.0000
			(0.0002)			(0.0002)
Constant	−2.3462***	−2.3683***	−2.3451***	1.0609***	1.0465***	1.3046***
	(0.0068)	(0.0069)	(0.0165)	(0.0066)	(0.0068)	(0.0145)
Observations	1,083,154	1,083,154	1,083,154	410,054	410,054	410,054
Robust R^2	0.0234	0.0255	0.0006	0.0355	0.0358	0.0007

Notes: The regressions consider both effects of the inter-regional learning and the inter-industry learning. Data are for the 2006-2013 period. The development of new industries in a two-year period asks for backward and forward conditions. The probit regressions include the year-fixed effects. Significant level: $*p < 0.1$, $**p < 0.05$, and $***p < 0.01$.

Besides the density, we used two alternative measures to count how many active related industries that one industry has or how many active neighboring regions that one region has: the ratio and the active number. The two measures are based on the illustrated industry space for the year 2013 and the original adjacent network of regions. The ratio of active related industries is calculated by the number of active related industries (the active number) to the number of all related industries (the number) in the industry space. The ratio of active neighboring regions is calculated by the number of active neighboring regions (the active number) to the number of all neighboring regions (the number) in the region adjacent network.

Table S9 presents the regression results for developing new industries in columns (1-2) and keeping previous industries in columns (4-5) using the two active numbers. Results suggest that both effects of the two learning channels are jointly significant, but their combination exhibits diminishing returns, meaning that the two learning channels are substitutes. Columns (3) and (6) of Table S9 present a negative control, where we use the number (no matter whether they are active or not). We noticed that the model remarkably loses its explanatory power, and the effects are small. These results suggest that, instead of having many related industries or neighboring regions, what real matters is whether they are active or not.

9 Simulations for Inter-industry Learning and Inter-regional Learning

We use the bootstrap percolation model to check the robustness of our simulations on the inter-industry learning and the inter-regional learning. The bootstrap percolation process works as follows: (i) Nodes are in either active or inactive status; (ii) A node remains active forever once it is activated; (iii) A given ratio of nodes (p) are initially activated; (iv) Inactive nodes become active if their at least k neighbors are already active; (v) Nodes are activated in an iterative manner according to the condition in (iv), until no more nodes can be activated. Here, the threshold is set as $k = 3$ for inter-industry learning based on the illustrated industry space of 2013 and $k = 4$ for the new region adjacent network. The used threshold is slightly larger than the half of the average degree of the corresponding network. Figure S8 presents the phase diagrams of inter-industry learning and inter-regional learning after simulations using the bootstrap percolation model.

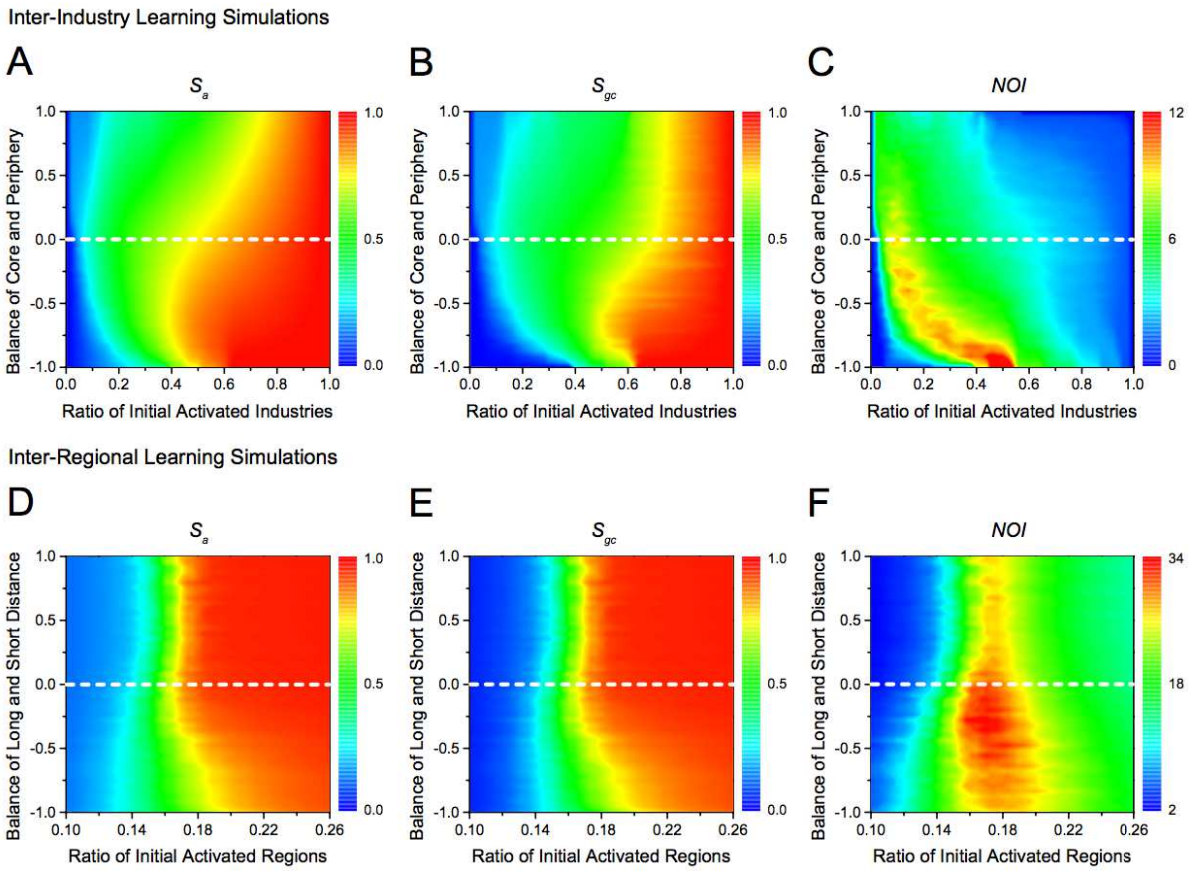

Figure S8. Simulation results on maximizing the two collective learning effects. Inter-industry learning: (A), (B) and (C) show respectively the ratio of active nodes (S_a), the ratio of giant component of active nodes (S_{gc}) and the number of iterations (NOI) before reaching the steady state. The horizontal-axis is the ratio of initially activated industries, and the vertical-axis is the balance index, which controls the strategy of selecting periphery-located (-1), random-located (0) and core-located (1) industries in the industry space for initial activation. Inter-regional learning: (D), (E) and (F) show respectively S_a, S_{gc} and NOI. The horizontal-axis is the ratio of initially activated regions, and the vertical-axis is the balance index, which controls the strategy of connecting nearby (-1), random (0) and distant (1) regions. All figures use the bootstrap percolation model with $k = 3$ for the inter-industry learning (average degree of the network is around 6.5) and $k = 4$ for the inter-regional learning (average degree of the network is around 7.5).

For inter-industry learning in Figures S8A-C, the diagram is trivial when the ratio of initially activated industries is below 0.4 or above 0.8 because different strategies of choosing initial active industries give almost the same results. However, an non-trivial area emerges in middle of the diagram, where the periphery-preferred strategy (with balance index being around -0.5) leads to the full activation of all industries. The results suggest an optimal strategy of the balance between core and periphery industries in maximizing the range and minimizing the time of industrial diversification through the inter-industry learning channel.

For inter-regional learning in Figures S8D-F, the diagram is trivial when the ratio of initially activated regions is below

0.16 or above 0.20, because different strategies in selecting the distance of spatial connections between regions give almost the same results. However, the middle part of the diagram becomes non-trivial, showing that the random connecting strategy (with balance index being 0) and the distant-preferred strategy (with balance index being larger than 0) give both the full activation of all regions and the connected active component. The results suggest that, given a ratio of initially activated regions, there is an optimal strategy of the balance between long and short distance regions in adding the new connection. Distant-preferred strategy (for example, always building long-range flights) will not do better than a random strategy (for example, building some long-range flights and some short-range rails) in maximizing the range and minimizing the time of industrial diversification through the inter-regional learning channel.