The latent structure of national scientific development

Lili Miao¹, Dakota Murray¹, Woo-Sung Jung²,³, Vincent Larivièreme⁴,⁵, Cassidy R. Sugimoto⁶, Yong-Yeol Ahn*¹,⁷,⁸

¹Center for Complex Networks and Systems Research, Luddy School of Informatics, Computing and Engineering, Indiana University Bloomington, IN 47408, USA.
²Department of Industrial and Management Engineering, Pohang University of Science and Technology, Pohang 37673, Republic of Korea.
³Department of Physics, Pohang University of Science and Technology, Pohang 37673, Republic of Korea.
⁴École de bibliothéconomie et des sciences de l'information, Université de Montréal, Montréal, Québec H2L 2C4, Canada.
⁵Observatoire des sciences et des technologies, Université du Québec à Montréal, Montréal, Québec H2L 2C4, Canada.
⁶Luddy School of Informatics, Computing and Engineering, Indiana University Bloomington, IN 47408, USA.
⁷Network Science Institute, Indiana University, Bloomington (IUNI), IN 47408, USA
⁸Connection Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

*Yong-Yeol Ahn

Email: yyahn@iu.edu

Abstract

Science is considered essential to innovation and economic prosperity. Understanding how nations build scientific capacity is therefore crucial to promote economic growth and national development. Although studies have shown that national scientific development is affected by geographic, historic, and economic factors, it remains unclear whether there are universal structures and trajectories behind national scientific development that can inform forecasting and policy making. By examining countries’ scientific exportation—the publications that are internationally indexed—we reveal a three-cluster structure in the relatedness network of disciplines that underpin national scientific development and the organization of
global science. Tracing the evolution of national research portfolios reveals that while nations are proceeding to more diverse research profiles individually, scientific production is increasingly specialized in global science over the past decades. We further demonstrate that the revealed disciplinary clusters inform economic development, where the number of publications in applied research centered cluster significantly predicts economic growth. By uncovering the underlying structure of scientific development and connecting it with economic development, our results may offer a new perspective to study national scientific development and its relationships with economic development.

Introduction

Is there a universal trajectory of national scientific development? Several approaches have been taken to describe the evolution of science as well as the universal patterns of scientific development. Comte (1) argued—albeit not considering national context—that science develops along a natural trajectory from high-consensus physical sciences, towards more complex, low-consensus social sciences. Basalla (2) took a colonial perspective, arguing that scientific development in non-Western countries began with colonial exportation—providing natural resources to Western countries—and then developed scientific capacity within the Western tradition. In contrast to this “western recipe” of national scientific development, studies have examined how the interplay between geography (3), history (4), existing scientific strengths (5), and economic conditions (6) influence development. Chile exemplifies the influence of geopolitical opportunities and constraints on national knowledge production: despite relatively low scientific investments (7), Chile’s unique mountainous and remote terrain made it ideal for astronomical observatories, a comparative advantage that allowed the nation to become an
international hub for astronomy and astrophysics (8, 9). By contrast, South Korea, with its heavy investment in science (10, 11), has experienced diversified scientific expansion, developing into a science and innovation powerhouse (12).

With the increasing scientific capacity across countries, global science has been experiencing rapid transformation. During Cold War, the USSR and the United States competed in science; the collapse of the USSR in the late 1990s and the concurrent rise of China on the international stage significantly altered the power dynamics in science. Whereas China only accounted for 5% of the world’s scientific publications in 2000, it became the most productive country in the world by 2018, surpassing US’s production (13). The increase in scientific growth was also coupled with Asia’s economic take-off: for example, the rapid expansion and intense industrialization of the “Four Asian Tigers”— Hong Kong, Singapore, South Korea, and Taiwan—also occurred during this time. These contemporary shifts prompt the question of the relationship between economic and scientific development and the potential universality of the trajectory of scientific development.

The study of economic output with respect to the “product space”—i.e., the network of relatedness between exported products—revealed that the networked structure of industrial advantage is critical to understanding the economic development of nations (14). We apply this framework to examine national science production by considering scientific disciplines as types of “products” that are exported by countries. That is, we investigate a nation’s scientific development through scientific exports, in which research articles produced by a country and
indexed in international bibliographic databases represent the exported scientific “products” of the nation (15, 16). This is an important operationalization for our study; whereas there is a significant amount of scientific production that may happen within a country—e.g., in non-English languages, in grey literature, or governmental reports—we argue that it is those works that are made visible through indexation that are the best proxies for exportation. This is not to diminish localized scientific activity, but to create a measurement that approximates economic exportation.

**Results**

We employ Revealed Comparative Advantage (RCA) (2)—a common measure for quantifying the economic and production advantages of countries (14)—to assess each nation’s relative disciplinary strengths based on publications indexed by the Web of Science database (see Data & Methods). If country $c$ produces a greater share of its publications in field $i$ compared to the world average share in the discipline, then $\text{RCA}_{c,i} > 1$ and country $c$ is considered to have a revealed comparative advantage (or specialization) in discipline $i$.

We calculated the RCA for all combinations of 143 disciplines and all countries in our dataset. As expected, the patterns of relative advantage reflect a range of historical, geographical, and cultural factors (see Fig. 1). For instance, countries with relative strength in Botany are located in tropical areas rich in botanical resources; Anthropology and Archaeology features both wealthy and developing nations, reflecting the remnants of colonial science and alluding Basalla’s postulation that the science in colonial and post-colonial countries began with Western countries’
exploitation of natural resources (2). By contrast, far fewer countries—mostly in North America and Europe—specialize in *Biochemistry & Molecular Biology*, a discipline which requires sufficient funding and sophisticated technologies. Similarly, *Cancer* research is largely concentrated in countries with high cancer mortality (which is associated with longer lifespans) as well as advanced countries with the capacity to invest in clinical research (17). That research and innovation emerge as a response to local issues and threats can also be observed in other contexts. For example, *Agricultural & Food Science* and *Health Policy and Services* are prominent in nations across the global south, where infectious disease (18) and food security (19) are pressing issues. Large emerging economies like China and India are specialized in fields such as *Industrial Engineering* and *Applied Physics* that contribute to industrially-relevant research (20). Similarly, the relative strength of Russia, Ukraine, and Kazakhstan in Applied Physics may be explained as a remnant of the Soviet Union’s research priorities (4).
Figure 1. Disciplinary specialization reflects geographical, historical and economic factors. Eight examples illustrate the distribution of disciplinary specializations. Discipline specialization is measured by Revealed Comparative Advantage (RCA). Color represents the Logarithm of RCA; a nation is only colored if its $\log_{10} \text{RCA}_{c,i} > 0$. Grey corresponds to nations that were not represented in our dataset. Botany, Anthropology, and Archaeology reflect the presence and access to natural and anthropological resources in a country. Economic inequality underpin specialization in resource-intensive disciplines like Biochemistry & Molecular Biology and Cancer. Local issues also drive research, as can be seen from the distribution of Agricultural & Food Science and Health Policy & Services. The distribution of Industrial Engineering and Applied Physics likely reflects national economic priorities and policies.
The distribution of disciplinary specialization suggests scientific exportation is affected by geographic, historical, social, and economic factors. Do these idiosyncratic factors dominate the course of scientific development of a nation? Or is there an underlying structure that governs the scientific development of nations?

Inspired by the relatedness network of economic product exports that underpins national economic development (14), we construct a discipline relatedness network, in which the proximity between disciplines is defined by the minimum conditional probability that two disciplines are co-specialized in a country (see Data & Methods). The network builds on the idea that disciplines that are co-specialized are likely to require similar knowledge, skills, methods, or equipment. To reveal its most salient structure, we apply the multi-scale backbone extraction method (21). This “backbone” reveals three clusters—which we confirm in the full network with the Leiden Algorithm (22) (see Data & Methods)—which we call Natural, Physical, and Societal clusters (see Fig. 2a). These clusters—while resembling previous observations (23)—do not conform to the common high-level classifications of disciplines. None of the clusters exclusively coincide with major classifications such as natural sciences, engineering, or medical sciences. The high-level disciplinary classifications appearing in Natural cluster (left) are primarily Natural and Medical Sciences. Most disciplines are dependent upon natural resources (e.g., Geology, Entomology, and Agriculture & Food Science), or concern the prevalent medical concerns in low-income areas (e.g., Nutrition & Dietetic and Parasitology). The Physical cluster (right) contains primarily physical sciences and engineering, which are commonly considered as foundations for industry-based
economic growth (e.g., Chemistry and Applied Physics) and those that require technological investment (e.g., Civil Engineering, Astronomy & Astrophysics, and Aerospace Technology); this cluster suggests the intimate relationships between basic physical science and engineering. The *Societal* cluster (top) is formed by human-centric disciplines that are focused on improving societal welfare, including Medical Sciences (e.g., Psychiatry, Nursing, and Cancer) as well as Social Sciences and Arts & Humanities (e.g., Education, Sociology, and International Relations).
Figure 2. The Structure of the disciplinary proximity network and national development. (a) The backbone of the disciplinary relatedness network reveals three clusters, which we call Natural, Physical, and Societal. Each node corresponds to a discipline and the weight of an edge captures the minimum conditional probability of co-specialization (see Data & Methods). The area of a node is proportional to the number of total publications indexed in that discipline. Node color maps to five broad disciplinary categories. (b) Nations are classified into four groups by their income level: Low, Low-Middle, Upper-Middle and High (from left to right). Dots correspond to nations, and a nation’s position inside the simplex is calculated as the fraction of advantaged disciplines in each cluster normalized by its total number of advantaged disciplines. The density estimate of each income group is shown in red. (c) National research profile snapshots (2013-2017) and GDP. Points are colored according to the nation’s log-transformed GDP. (d) Four example countries, Ethiopia, Vietnam, China, and the United States (USA) at 2013-2017. Only the discipline with an advantage ($\log_{10}RCA > 0$) are colored. Node colors are the same as in (a).
These clusters offer a concise representation of each country’s research portfolio. Namely, each country’s scientific portfolio can be represented as a point in the simplex of the three clusters (see Fig. 2b; see Data & Methods). Aggregating countries based on their income-level classification (24) reveals that niches are largely related with national wealth (see Fig. 2b-d). Low-income countries (e.g., Afghanistan, Ethiopia, and South Sudan) tend to be confined to the *Natural* cluster; some of the low-middle countries extend towards the *Physical* disciplines whereas upper-middle income countries are located closer to the center. High-income countries (e.g., United States, France, and Japan) tend to occupy the center and the space between *Natural* and *Societal*, suggesting balanced exportation. This pattern suggests there might be a universal tendency that as a nation’s economic power increases, their scientific exports move towards a more balanced portfolio.

To understand the temporal evolution of national research portfolios, we first examine whether the development (or the loss) of revealed comparative advantage follows the “law of proximity” (5), which predicts that countries are more likely to develop a new advantage in a discipline that are close to their existing advantages (see Fig. 2). By examining the (de)activation of advantages across each subsequent time steps (see Data & Methods), we show the law of proximity indeed holds (see Fig. 3a-b). The probability of a new activation increases with the density of proximate specialized disciplines; the probability of deactivation follows the opposite pattern.
Figure 3. Disciplinary proximity dictates the development and loss of competencies (a) Probability of a new relative advantage in the next time period given the density of existing advantages surrounding the discipline. Dots represent the estimated probabilities from Bootstrapping and solid lines are the estimated regression line from bootstrapped samples. (b) Probability that an advantaged discipline will lose its advantage in the next period given the density of existing advantages. (c)-(d) Same plots where countries are grouped based on their income class. (e) We show the predicted and actual evolution in the simplex. Arrows point to the average simplex position of countries in the next period. Red arrows represent the empirical movement while gray arrows represent the movement predicted from the null model based only on the law of proximity.

Moreover, if we aggregate countries based on income groups, we further discover that low-income countries are more strongly constrained by the law of proximity than others (See Fig 3c-d). In other words, it is more difficult for low-income countries to develop a new relative advantage if it is not in the vicinity of already existing advantage, while wealthy countries are more likely to develop new advantage, thanks to the already existing diverse and complex research portfolio with broader and higher disciplinary production, allowing them to make less portfolio-dependent choices when building scientific capacity.
The law of proximity shapes scientific development, but to what extent? We compare the actual trajectories with a null model that is solely based on the law of proximity (see Data & Methods). As shown in Fig 3e, the predicted research profiles converge towards the center of the simplex; in other words, even with the constraining effect of the law of proximity, the connections across clusters are strong enough to attract countries towards a balanced research portfolio. By contrast, the aggregated actual trajectories display much weaker attraction towards the center, suggesting scientific development is not entirely dictated by the law of proximity but may also be conditioned by the three clusters (see Fig 3e and SI). The difference is particularly stark for countries specialized in Natural cluster, suggesting that low-income countries may face a heavy hurdle breaking into other disciplines.

Meanwhile, global science has been moving from a nested structure to a modular structure. It has been observed that the global economy exhibits a hierarchical (14, 25, 26) (or nested) structure, where rich countries can export a wide range of products—especially those that exported by only a few countries—whereas poor countries can only export a small number of products that can be exported by many (14, 27, 28). This pattern contrasts a more classical theory of specialization, where countries specialize and form a ‘modular’ structure. Inspired by the tension between these two ideas, we measure the nestedness and modularity of the scientific exports over time (see Data & Methods and Fig. 4). In contrast to the case of economic products, we do not observe strong evidence of nestedness; instead, we find that the modularity of the network has been increasing,
which is likely associated with the trapping of low-income countries in \textit{Natural} cluster (see Fig. 3) and the heavy investment and emphasis of applied sciences in rising economies such as China.

Motivated by the connection between the economic wealth and their scientific niches and the increasing diversity shown above, we investigate the relationships among scientific diversity, publication volume and economic performance. We measure the diversity of a scientific portfolio

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure4.png}
\caption{\textbf{Nestedness and modularity of global science} (a) The country-discipline RCA matrices of the earliest and the most recent periods where rows and columns are arranged in a descending order of number of advantages. (b) The $z$-score of nestedness over time which is calculated through Fixed-Fixed null model. (c) The country-discipline RCA matrices of the earliest and the most recent time periods where disciplines(columns) are arranged by its classification into the three clusters. (d) The $z$-score of modularity over time which shares the same null matrix as used for calculating nestedness.}
\end{figure}
with the Gini index of disciplinary RCA values (see Data & Methods). For convenience, we define the *scientific diversity* of a nation as one minus their Gini index. High scientific diversity corresponds to a more balanced and diversified portfolio, whereas low scientific diversity indicates more skewed and specialized exportation.

![Figure 5. Scientific production is correlated with national development indicators. (a) Number of publications is strongly correlated with scientific diversity (defined as one minus the GINI index of the RCA values of a country). (b)-(c) The relationship between scientific publication volume and nation’s GDP (c) and their economic complexity index. (d) The temporal development of the number of productions by income group. Error bars represent the confidence interval drawn from bootstrapping.](image)

We find that the number of publications, GDP, ECI (Economic Complexity Indicator), and scientific diversity are all strongly correlated with each other (see Fig. 5). Over the past 40 years
the number of publication as well as scientific diversity of nations have been steadily increasing across all income groups (see Fig.5d-e). However, this steady growth is not enough to close the gap between income groups, the gap between high-income countries and low-income countries remains wide. Although scientific diversity is correlated with the number of publications, the diversification of research portfolio cannot be explained by the increase in number of publications alone (see Fig. S9).

Our results with two-way fixed effects panel regression models corroborate a previously reported, mutual influence relationship between GDP and publication volume (29–31) (see Table 1, Table S3). We further reveal that, if we divide the publications into the three clusters we identified, only the number of publications in Physical cluster significantly predicts GDP growth (see Table 1). In contrast to the results that economic complexity captures the capabilities available in a country and further helps to predict the long-term GDP growth (32), we could not find evidence that either ECI or the scientific diversity contributes to economic development or publication growth (see Table 1, Table S3), while GDP predicts the growth rate of scientific diversity. This suggests a possibility that the scientific production of a country, rather than the exported product portfolio, may better capture its capabilities for innovation and growth. Our results also suggest that a balanced research portfolio may be a result, rather than the cause of economic development (see Table S4). However, scientific diversity is negatively associated with the similarity between new activated disciplines and the current existing advantage, and a balanced research profile is associated with more flexibility to develop research areas (see Table 2). Countries with high diversity tend to develop more easily beyond their current research advantages while with low
diversity prefer to develop similar disciplines.

Table 1 Regression results of predicting GDP growth

|                  | (1)        | (2)        | (3)        | (4)        |
|------------------|------------|------------|------------|------------|
| Dependent variable: GDP Growth (log-ratio) |            |            |            |            |
| GDP              | -0.418***  | -0.441***  | -0.432***  | -0.439***  |
|                  | (-0.478, -0.358) | (-0.503, -0.380) | (-0.495, -0.369) | (-0.500, -0.377) |
| ECI              | 0.007      | 0.002      | 0.005      | 0.002      |
|                  | (-0.024, 0.037) | (-0.029, 0.032) | (-0.026, 0.036) | (-0.028, 0.033) |
| Num.Pub          | 0.051***   | 0.059***   |            |            |
|                  | (0.021, 0.081) | (0.023, 0.094) |            |            |
| Num.Natural      | -0.0002    |            |            |            |
|                  | (-0.061, 0.060) |            |            |            |
| Num.Physical     | 0.052**    |            |            |            |
|                  | (0.005, 0.098) |            |            |            |
| Num.Societal     | -0.005     |            |            |            |
|                  | (-0.056, 0.047) |            |            |            |
| Diversity        | -0.077     |            |            |            |
|                  | (-0.267, 0.112) |            |            |            |
| Observations     | 838        | 838        | 825        | 838        |
| $R^2$            | 0.208      | 0.220      | 0.214      | 0.221      |
| Adjusted $R^2$   | 0.064      | 0.077      | 0.065      | 0.076      |

*Note: *p<0.1 **p<0.05 ***p<0.01
Table 2 Regression results of predicting average similarity of new activated disciplines.

|                | (1)   | (2)   | (3)   | (4)   | (5)   |
|----------------|-------|-------|-------|-------|-------|
| **Dependent variable:** Similarity |       |       |       |       |       |
| GDP            | -0.103|       |       | -0.080|       |
|                | (-0.328, 0.121) |       |       | (-0.306, 0.145) |       |
| ECI            | -0.021|       |       |       | -0.017|
|                | (-0.133, 0.090) |       |       | (-0.129, 0.094) |       |
| Num.Pub        | -0.024| 0.022 |       | 0.038 |       |
|                | (-0.133, 0.086) | (-0.031, 0.075) |       | (-0.092, 0.168) |       |
| Num.Natural    |       |       | -0.107|       |       |
|                |       |       | (-0.237, 0.023) |       |       |
| Num.Physical   |       |       | 0.091*|       |       |
|                |       |       | (-0.012, 0.195) |       |       |
| Num.Societal   |       |       | -0.010|       |       |
|                |       |       | (-0.127, 0.108) |       |       |
| Diversity      | -0.304|       | -0.608*|       |       |
|                | (-0.689, 0.081) |       | (-1.298, 0.081) |       |       |
| Observations   | 837   | 1,496 | 1,284 | 1,496 | 837   |
| R²             | 0.002 | 0.001 | 0.004 | 0.002 | 0.006 |
| Adjusted R²    | -0.182| -0.165| -0.187| -0.163| -0.178|

*Note:* *p<0.1 **p<0.05 ***p<0.01
Conclusion

It is widely believed that scientific development holds the key for a nation’s future prosperity (33, 34). Yet, whether there are universal structural patterns of scientific development at the national level has been an open question. By analyzing more than 30 million scientific publications across 217 countries spanning the period 1973-2017, we provide a large-scale temporal analysis of national science development. We find that the disciplinary proximity network constructed from these publications exhibits three clusters of disciplines which roughly capture the relative advantages of countries across the spectrum of economic wealth. Although each country’s position in the network is shaped by various historical, geographical, social, and economic factors, the three-cluster structure still conditions their scientific development. We further reveal the relationship between economic growth, publication volume, and scientific diversity. The size of scientific enterprise significantly predicts the GDP growth; we show that the publication volume, especially those in Physical cluster, is a much more significant predictor of future GDP growth than ECI, suggesting that it is critical to consider scientific capacity in assessing the economic potential of a country. Finally, we find evidence that the economic growth leads to higher scientific diversity not vice versa.

Our results are reminiscent of classical theories. The clusters and the niches that are occupied by nations show some semblance of Comte’s “Hierarchy of the Sciences” (1855) hypothesis—that
science progresses from natural sciences that require readily-available simple subjects, towards social sciences that deal with more complex subjects. At the same time, the prominence of *Natural* disciplines in low-income countries resonates with Basalla’s “Spread of Western Science” (Basalla, n.d.), pointing to the colonial exploitation of natural resources.

This study is subject to limitations. First, our study relies on a bibliographic database created and maintained by a Western scientific enterprise. Therefore, it overestimates the research from western countries and the publications in English while underestimating the production in other nations and languages (see SI Data). Still, we argue that our operationalization is reasonable under the analogy to product exportation (14, 27) and the status of English as the de-facto lingua franca (35) of science. Second, many analyses considered the RCA matrix as a bipartite network. This approximation is not strictly valid because the edges are not independent from each other. Finally, we also note that reliable causal inference with country-level data is often infeasible and our results do not necessarily confirm a direct causal relationship between national scientific development and economic growth. There would exist unobserved hidden confounders and complex feedback mechanisms between scientific and economic development.

Even with these limitations, our empirical framework may provide a useful perspective to study the structure and evolution of national scientific portfolios and the relationship to economic development. Our results call for attention to the barriers faced by low-income countries in building their scientific capacity, and the potential consequences on future scientific capacity and
economic growth. Our results also highlight the importance of considering scientific capacity in the study of economic development. We hope our analysis opens a new avenue towards the understanding of the mechanisms of scientific development as well as its relationship to economic prosperity.

**Data & Methods**

**Data.** The dataset was drawn from the Clarivate Analytics’ Web of Science database hosted and managed by the *Observatoire des Sciences et des Technologies* at the University of Montreal. The Web of Science database contains three main citation indices: The Science Citation Index Expanded, the Social Science Citation Index, and the Arts and Humanities Citation Index. We used all indexed publication records listed as being published between 1973 to 2017, which included 37,479,532 papers published across 20,252 scholarly journals. To examine temporal patterns, we split the data into nine five-year snapshots. We limited this set to only journal articles, review articles, and notes (discontinued in 1991 but included in articles). We also excluded any publication that did not list any institutional address, and publications that could not be assigned a disciplinary category according to the steps below. After these filters, the dataset contained 35,793,320 papers published across 20,123 scholarly journals (See Fig. S2).

Discipline classification of publications is based on the National Science Foundation typology of journals, which categorizes papers into a hierarchy of disciplines. The high-level and granular classification was further complemented with an in-house classification of the Arts and Humanities (36). The resulting classification scheme contains 144 granular categories. After removing
“Unknown” from the 144 granular categories, we manually classified each of the 143 categories into one of five broad categories: “Natural Science”, “Medical Science”, “Engineering”, “Social Science”, and “Arts and Humanities”; this scheme is used to color nodes in Figure 2.

Publications are associated with nations using the institutional addresses listed by the authors. We assign a full unit credit of a publication to every country of affiliation represented on the paper’s author byline (“full counting”). For example, a paper listing five authors—two with affiliations in the United States, two in Canada, and one in the Netherlands—would count as one paper to all three countries. Full counting method assumes each author contributes equally to the publication. Different counting methods are highly correlated at the macro level (37). See SI data section for more details.

We use data on national GDP from the World Bank (24, 38) to approximate the economic wealth of each country. The dataset covers 264 countries from 1960 to 2019. Income classification comes from World bank database (24) which contains 224 countries between 1987 and 2018. We convert the annual classification to a time snapshot classification by assigning each country to its most frequent income group during each period. See SI data section for more details.

**Revealed Comparative Advantage.** The revealed comparative advantage (RCA) of country $c$ in discipline $i$ is defined as:
\[
\text{RCA}_{c,i} = \frac{P(c, i) / \sum_i P(c, i)}{\sum_c P(c, i) / \sum_{c,i} P(c, i)}
\]

where \(P(c, i)\) is the number of publications produced and “exported”—the number of publications which is indexed in the Web of Science—by country \(c\) in discipline \(i\), \(\sum_i P(c, i)\) is the total number of publications produced by country \(c\), \(\sum_c P(c, i)\) is the total number of publications produced in a discipline globally, and \(\sum_{c,i} P(c, i)\) is the total number of publications across all countries and disciplines.

**Disciplinary Proximity.** The proximity between disciplines \(i\) and \(j\) is defined as the minimum of the pairwise conditional probabilities of a country having an advantage (RCA > 1) in one discipline given an advantage in another:

\[
\phi_{ij} = \min\{P(\text{RCA}_i > 1|RCA_j > 1), P(\text{RCA}_j > 1|RCA_i > 1)\}
\]

\(\phi\) is a 143 \times 143 matrix that captures the proximity between pairs of disciplines (see SI Fig. S4).

**Identifying the disciplinary clusters.** The relatedness network is constructed from the disciplinary proximity matrix derived from aggregating data across all years (from 1973 to 2017). The network is fixed over the analysis. Although the network structure changes over time, networks derived from a snapshot of data closely resemble with the aggregated network (see SI Fig. S5). The multi-scale backbone extraction method (21) exposes three visual clusters when laid out with a force-directed layout algorithm (Gephi’s ForceAtlas2). We formally confirm the structure by applying a community detection algorithm (the Leiden algorithm (22)) to the full
network. We use modularity as the quality function and ran the model 50 times to obtain consensus. Other methods produce similar results although some partitions the network into smaller communities (see SI Disciplinary Relatedness Network section).

**Position within the simplex.** Position within the simplex measure countries’ cluster-level specialization concentration. We first calculate $C_i = n_i/N_i$, where $n_i$ is the number of disciplines in cluster $i$ with $RCA > 1$, $N_i$ is the total number of disciplines in cluster $i$. Then we normalize $C_i$ so that $\sum_i C_i = 1$.

**The Density of existing advantages and the null development model.** The density of existing advantages around a given discipline is defined as follows:

$$\omega^k_j = \frac{\sum_i x_i \phi_{ij}}{\sum_i \phi_{ij}}$$

where $\phi_{ij}$ is the proximity between discipline $i$ and $j$, and $x_i = 1$ if $RCA_{kl} > 1$ else $x_i = 0$, and the density of existing advantages, $\omega^k_j$, is the proximity-weighted sum of all disciplines that are connected to $j$ with $RCA_{kl} > 1$. We bin the density values and aggregate across countries and time periods to calculate the probability of activation and deactivation, given the density. We also perform a bootstrap sampling with 20 samples to estimate the uncertainty of the slope and report the mean and standard deviation of the slopes across bootstrapped samples. A linear regression model (OLS) is fit by pooling all bootstrap samples to obtain the parameters (intercept and slope) for the null model. The null model works as following: for every inactive ($RCA < 1$) discipline,
we assign a probability the discipline will be activated (RCA > 1) in the subsequent time period based on its current density using the intercept and slope obtained from the pooled regression model that include all countries. We use the same procedure for the deactivation. For each time period and each country, the newly activated and deactivated disciplines are sampled using the null model while preserving the number of new activation and deactivation in the next time period. We repeat this procedure 100 times. When visualizing the actual profile and the predicted profile on the simplex, to reduce the influence of extreme cases, we remove data points located on the boundary of the simplex. To smooth out the noise, we aggregate data points within each rhombus with the side length of 0.1 that tessellates across the simplex. We observe that the difference between actual trajectory and the predicted trajectory is robust against the direction of rhombus.

**Modularity and Nestedness.** We use the country-discipline bipartite network to represent knowledge exportation. Country $c$ is connected to discipline $i$ if $RCA_{c,i} > 1$. Modularity (39) of the country-discipline bipartite network is defined as:

$$Q = \frac{1}{m} \sum_{i=1}^{p} \sum_{j=1}^{q} (A_{ij} - P_{ij}) \delta(g_i, h_j)$$

Where $m$ is the number of links, $A_{ij}$ equals to 1 if there is a link from node $i$ to node $j$, $P_{ij}$ is the probability the edge between $i$ and $j$ exists under the null model, $g_i$ and $h_j$ are communities that the country and discipline belong to. The community of a country is decided by its largest cluster level revealed comparative advantage; for example, China is classified to Physical cluster since it has highest cluster-level RCA value in Physical cluster. The community of disciplines is defined
by the Leiden algorithm. Although the elements of the RCA matrix are not strictly independent from each other, we use $P_{ij} = \frac{k_i d_j}{m}$ (where $k_i$ and $d_j$ are the degree of node $i$ and $j$ respectively) as an approximation. Larger modularity means countries tend to be specialized in one of the three clusters rather than having advantages spread across multiple clusters.

Nestedness is measured by the overlap and decreasing fill (NODF) (40) method. NODF measures the degree of overlapping between row pairs and column pairs in the adjacency matrix. The metric is defined as

$$NODF = \frac{\sum N_{\text{paired}}}{\left[ \frac{n(n-1)}{2} \right] + \left[ \frac{m(m-1)}{2} \right]}$$

Where $\sum N_{\text{paired}}$ is the averaged degree of nestedness for each pair of row and column based on the principles of decreasing fill and paired overlap (40), $n$ and $m$ are the number of rows and columns.

We use a null model to test whether modularity and nestedness are significant. We construct the null model of the bipartite network by swapping edges between node pairs while constraining the degree of each node which we refer to as the Fixed-Fixed null model.

**Scientific Diversity.** The Gini index of a nation’s RCA values across disciplines is used to capture the scientific diversity of a nation. For convenience, we use 1 minus the Gini index as a measure of scientific diversity. If all disciplines have the same RCA value in the country, the
diversity value would be 1. If a country only produces scientific publications in one discipline, then the diversity value would be 0. To investigate the dynamic relationship between scientific diversity and economic power, we project countries’ evolution into the diversity-GDP plane. To smooth out noise, we averaged the trajectory in each grid with width equals to 0.1 and height equals to 0.5. The starting point of arrow represents the average of all displacements whose starting points were in the grid. The direction and length of arrows are computed by averaging the subsequent displacements of all countries within a grid.

**Regression Analysis**: We use fixed-effect panel regression model to investigate the relationship between economic growth and scientific development. The model is written as following:

\[
\log \left( \frac{N_{c,t+1}}{N_{c,t}} \right) = \beta_0 + \beta_1 \log (GDP_{c,t}) + \beta_2 ECl_{c,t} + \beta_3 \log (Num.Pub_{c,t}) + \beta_4 Diversity_{c,t}
\]

\[+ \beta_5 T_{N,t+1} + \beta_6 T_{P,t+1} + \beta_7 T_{S,t+1} + \alpha_c + \alpha_t\]

Where \(c\) denotes countries, \(t\) denotes time periods, \(N_{c,t}\) is the investigated dependent variable, \(GDP_{c,t}\) is the averaged GDP value of country \(c\) during time period \(t\), \(Diversity_{c,t}\) is the scientific diversity value of country \(c\) during time period \(t\), \(T_N, T_P\) and \(T_S\) are the number of publications in \(Natural, Physical\) and \(Societal\) clusters during time \(t\), \(\alpha_c\) and \(\alpha_t\) are the country-specific and time specific intercepts that capture the heterogeneity across countries and across time periods. The dependent variables involved in our analysis are GDP growth rate, publication growth rate, scientific diversity growth rate and averaged disciplinary similarity.

Averaged disciplinary similarity measures how similar the new activated disciplines are comparing with the current existing advantages. The averaged disciplinary similarity is defined
as:

\[ \rho_c = \sum_{i} \frac{n \omega_i^c}{n} \]

Where \( n \) is the number of new activated disciplines and \( \omega_i^c \) is the normalized density of newly activated disciplines. Normalized density is measured as the z-score of raw disciplinary density of all inactivated disciplines. High averaged disciplinary similarity indicates the new activated disciplines have higher similarity with existing advantaged disciplines compared with average similarity between advantaged disciplines and inactivated disciplines.

**Acknowledgements**

This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-19-1-0391. We thank Staša Milojević, Jisung Yoon, Sadamori Kojaku, Byungkyu Lee, Bruce Weinberg, Paula Stephan, Santiago Schnell, Tao Zhuo, Junming Huang, Jian Gao, Yi Bu, Qianming Zhang and Nima Dehmamy for helpful discussion and comments.

**Author Contributions:**

L.M. and D.M. conceived the study; All authors contributed to the design of the study; V.L. prepared the primary datasets; L.M., D.M., V.L., Y.Y.A performed analysis; All authors contributed to the interpretation of the results and writing of the manuscript.

**Data availability.** Data will be available at
https://figshare.com/account/projects/96980/articles/13623035.

**Code availability.** The code used for data processing and analysis will be available https://github.com/yy/national-science-exports.

**References**

1. A. Comte, *The Positive Philosophy of Auguste Comte* (C. Blanchard, 1855).
2. G. Basalla, The Spread of Western Science. *Science* **156**, 611–622 (1967).
3. D. N. Livingstone, *Putting Science in Its Place: Geographies of Scientific Knowledge* (University of Chicago Press, 2010).
4. J. Kozłowski, S. Radosевич, D. Ircha, History matters: The inherited disciplinary structure of the post-communist science in countries of central and eastern Europe and its restructuring. *Scientometrics* **45**, 137–166 (1999).
5. C. A. Hidalgo, *et al.*, “The Principle of Relatedness” in *Unifying Themes in Complex Systems IX*, Springer Proceedings in Complexity., A. J. Morales, C. Gershenson, D. Braha, A. A. Minai, Y. Bar-Yam, Eds. (Springer International Publishing, 2018), pp. 451–457.
6. S. Paula, *How Economics Shapes Science* (Harvard University Press, 2012).
7. T. Hornyak, Chilean research grows despite poor investment (2016).
8. L. Bronfman, A panorama of Chilean astronomy. *The Messenger* **107**, 14–18 (2002).
9. A. Bajak, Chile’s chance to embrace science for the twenty-first century. *Nature* **552**, S53–S55 (2017).
10. H. W. Yeom, South Korean science needs restructuring. *Nature* **558**, 511–513 (2018).
11. National Science Board, “Science and Engineering Indicators 2018” (2018).
12. R. V. Noorden, Science in East Asia — by the numbers. *Nature* **558**, 500–501 (2018).
13. National Science Board, National Science Foundation. 2019, “Publications Output: U.S. Trends and International Comparisons” (2019).
14. C. A. Hidalgo, B. Klinger, A.-L. Barabasi, R. Hausmann, The Product Space Conditions the Development of Nations. *Science* **317**, 482–487 (2007).
15. D. A. King, The scientific impact of nations. *Nature* **430**, 311–316 (2004).
16. R. M. May, The Scientific Wealth of Nations. *Science* **275**, 793–796 (1997).
17. F. Bray, *et al.*, Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for
36 cancers in 185 countries. CA. Cancer J. Clin. 68, 394–424 (2018).

18. J. A. Evans, J.-M. Shim, J. P. A. Ioannidis, Attention to Local Health Burden and the Global Disparity of Health Research. PLoS ONE 9, e90147 (2014).

19. World Health Organization, “World health statistics 2018: monitoring health for the SDGs” (2018) (August 29, 2020).

20. P. Zhou, L. Leydesdorff, The emergence of China as a leading nation in science. Res. Policy 35, 83–104 (2006).

21. M. A. Serrano, M. Boguna, A. Vespignani, Extracting the multiscale backbone of complex weighted networks. Proc. Natl. Acad. Sci. 106, 6483–6488 (2009).

22. V. A. Traag, L. Waltman, N. J. van Eck, From Louvain to Leiden: guaranteeing well-connected communities. Sci. Rep. 9, 5233 (2019).

23. F. Moya-Anegón, V. Herrero-Solana, Worldwide Topology of the Scientific Subject Profile: A Macro Approach in the Country Level. PLoS ONE 8, e83222 (2013).

24. N. Fantom, U. Serajuddin, The World Bank’s Classification of Countries by Income (The World Bank, 2016) (August 30, 2020).

25. S. Bustos, C. Gomez, R. Hausmann, C. A. Hidalgo, The Dynamics of Nestedness Predicts the Evolution of Industrial Ecosystems. PLoS ONE 7, e49393 (2012).

26. J. Gao, Y.-C. Zhang, T. Zhou, Computational Socioeconomics. Phys. Rep. 817, 1–104 (2019).

27. C. A. Hidalgo, R. Hausmann, The building blocks of economic complexity. Proc. Natl. Acad. Sci. 106, 10570–10575 (2009).

28. G. Morrison, et al., On Economic Complexity and the Fitness of Nations. Sci. Rep. 7, 15332 (2017).

29. R. Inglesi-Lotz, M. Balcilar, R. Gupta, Time-varying causality between research output and economic growth in US. Scientometrics 100, 203–216 (2014).

30. L.-C. Lee, P.-H. Lin, Y.-W. Chuang, Y.-Y. Lee, Research output and economic productivity: a Granger causality test. Scientometrics 89, 465–478 (2011).

31. P. Vinkler, Correlation between the structure of scientific research, scientometric indicators and GDP in EU and non-EU countries. Scientometrics 74, 237–254 (2008).

32. C. A. Hidalgo, Economic complexity theory and applications. Nat. Rev. Phys. 3, 92–113 (2021).

33. S. M. S. Kramer, Science, technology, and innovation for economic competitiveness: The role of smart specialization in less-developed countries. Technol. Forecast. Soc. Change 123, 95–107 (2017).

34. W. W. Powell, K. Snellman, The Knowledge Economy. Annu. Rev. Sociol. 30, 199–220 (2004).

35. M. D. Gordin, Scientific Babel: How science was done before and after global English (University of Chicago Press, 2015).
36. A. Éric, V.-G. Étienne, C. Grégoire, L. Vincent, Y. Gingrasb, Benchmarking scientific output in the social sciences and humanities: The limits of existing databases. *Scientometrics* **68**, 329–342 (2006).

37. C. R. Sugimoto, L. Vincent, *Measuring research: What everyone needs to know* (Oxford University Press, 2018).

38. World Bank, World Development Indicators. (2019). GDP (current US$). https://data.worldbank.org/indicator/NY.GDP.MKTP.CD?locations=1W.

39. M. J. Barber, Modularity and community detection in bipartite networks. *Phys. Rev. E* **76**, 066102 (2007).

40. M. Almeida-Neto, P. Guimarães, P. R. Guimarães, R. D. Loyola, W. Ulrich, A consistent metric for nestedness analysis in ecological systems: reconciling concept and measurement. *Oikos* **117**, 1227–1239 (2008).
Supplementary Information

Data

*Methods for crediting publications to nations.*

In bibliometrics there are three classic methods of assigning credit of a publication to individual countries: full counting, fractional counting, and corresponding author counting. Each of these methods is associated with distinct advantages, drawbacks, and implications. Full counting attributes one “article unit” to each country appearing on the article. Full counting is the simplest method. The main drawback of this counting method is that it leads to an overestimation of nation’s production. This means that the sum of all nation’s publications will be *more* than the global total of papers. Full counting can also inflate the publications of a country if the authors in the country tend to play a more marginal role in collaborations. Alternative to full counting, fractional counting attributes a fraction of the article based on the share of authorship. Instead of measuring how many publications are produced, fractional counting measures the proportion of contribution, implicitly assuming that the number of authors in the paper from a country is a good approximation of the contribution of the country to the paper. However, our dataset does not document the national affiliation of each author, but the national affiliation of each institution. Therefore, the fraction counting would measure the fractional contribution estimated by the institutional affiliation recorded in the bibliographic data. Finally, corresponding author counting is supposed to capture the country of the corresponding author—usually the principal investigator of the project—in a paper. Unfortunately, Web of Science has an inaccurate coverage on corresponding author information before 2008, where the first author is marked as the corresponding author. Considering these data limitations and for the sake of simplicity in interpretation(1), we focus on full counting in this analysis.

The structure of the disciplinary relatedness network remains robust regardless of the counting method. We observe a similar three-cluster structure in every type of network (see Fig S1). In general, the network derived from fractional counting has higher similarity with the full counting network. In terms of the discipline classification, the fractional counting network is differentiated from the full counting network in 9 disciplines while corresponding network has 13 disciplines that have inconsistent classification compared with the full counting network (see Table S1). The major discrepancy between full counting network and corresponding network happens in *Natural* cluster and *Societal* cluster. Among the 13 disciplines, 8 disciplines (e.g., Education, history, Law, and Literature) are classified to *Natural* cluster in the corresponding network while they originally belong to *Societal* cluster in the full counting network.
Figure S1. Backbone networks derived from different counting methods (alpha=0.2). The area of a node is proportional to the number of total publications indexed in that discipline. Node color maps to five broad disciplinary categories.

Table S1. Discipline classification difference across networks derived from different counting methods. N, P and S stands for Natural cluster, Physical cluster, and Societal cluster respectively.

| Discipline                        | Full | Fractional | Corresponding |
|-----------------------------------|------|------------|---------------|
| Genetics & Heredity               | N    | S          | S             |
| Miscellaneous Engineering & Technology | N    | P          | P             |
| Social Studies of Medicine        | N    | P          | P             |
| Civil Engineering                 | P    | P          | N             |
| Education                         | S    | N          | N             |
| History                           | S    | S          | N             |
| Information Science & Library Science | S    | N          | N             |
| Language & Linguistics            | S    | N          | N             |
| Law                               | S    | S          | N             |
| Literature                        | S    | S          | N             |
| Political Science and Public Administration | S    | N          | N             |
| Religion                          | S    | N          | N             |
| Urology                           | S    | S          | P             |

Evolution of Web of Science indexing and coverage

The coverage of the Web of Science database has changed over time, as shown in figure S2-3. While the number of papers has increased in a relatively stable manner, the number of journals indexed has followed a much less stable pattern of growth. These differences are influenced by
Clarivate’s indexing practices, which underwent major changes in 1975 (with the addition of the Arts and Humanities Citation Index), the early 1990s, and between 2006 and 2010. On the whole, number of papers increased from 303,393 in 1973 to 1,601,947 in 2017, and number of journals from 3,240 in 1973 to 12,788 in 2017.

Figure 2. Coverage of the Web of Science over time: Number of papers (articles, notes and reviews) and number of journals indexed within the Web of Science database, for each year between 1973 and 2017.
Like all bibliographic databases, the Web of Science has been shown to have geographic and linguistic biases. A recent survey of Web of Science journal coverage showed that the database over-indexes journals that publish papers in English-language, as well as journals from countries whose main language is English(2). This overestimation of English language literature is stronger in the social sciences and humanities, whose research topics are more likely to happen in the context of a particular country, and are therefore more likely to be published in languages other than English(3, 4) Therefore, the indexing of national social science and humanities literature is overestimated for English-speaking countries, and underestimated for non-English speaking countries. Although this is an important limitation, we chose our operationalization due to its simplicity and strong parallel to the operationalization of exports and the product space.

**Economic data**

Both GDP data and income group classification data are taken from World Bank database(5, 6). The GDP dataset contains annual GDP value in current dollar amounts from 1960 to 2019 which is available at [https://data.worldbank.org/indicator/NY.GDP.MKTP.CD](https://data.worldbank.org/indicator/NY.GDP.MKTP.CD). Income group data contains income classification per country from 1987 to 2018 which can be found at [https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-](https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-).
lending-groups. Among the 217 countries covered by WoS, 198 countries are covered by the World Bank GDP data and 200 countries are covered by the World Bank income group data. Economic Complex Indicator data is available at https://legacy.oec.world/en/. The data covers 131 unique countries from 1964 to 2017. All of the 131 countries have publication records in the WoS database. To make the annually updated economic data fit into our 5-year time interval, each country’s GDP and ECI value in each 5-year interval is calculated by averaging GDP and ECI value across the 5 years and the income group classification of each country is decided by its most frequent income group during the time period. However, since all three economic data sources have missing data in every year, the exact number of countries that are included in our analysis not only varies over time period but also based on the specific economic dataset we use in the analysis.

Table S2. Number of countries that are included over time periods

| Period     | WoS | GDP | GDP & ECI | Income Group |
|------------|-----|-----|-----------|--------------|
| 1973-1977  | 178 | 127 | 83        |              |
| 1978-1982  | 176 | 136 | 87        |              |
| 1983-1987  | 183 | 146 | 92        |              |
| 1988-1992  | 204 | 174 | 107       | 185          |
| 1993-1997  | 206 | 184 | 117       | 192          |
| 1998-2002  | 203 | 189 | 119       | 191          |
| 2003-2007  | 205 | 189 | 119       | 193          |
| 2008-2012  | 203 | 190 | 118       | 195          |
| 2013-2017  | 208 | 190 | 126       | 197          |

Disciplinary Relatedness Network

The aggregated disciplinary proximity matrix is showed in Figure S4. Due to the exponential growth of publications, the structure of network might be dictated by more recent data. To estimate the influence of recent data on network structure, we investigate whether the network structure is stable over time. We calculate Pearson’s Correlation Coefficient (PCC) of each disciplinary similarity between the aggregated network (network derived from whole time period
data) and the networks derived from each time snapshot. Although networks change over time, temporal snapshots share high resemblance with the aggregated network (see Fig S5).

Figure S4. Disciplinary proximity matrix: A 143 x 143 matrix records the pairwise similarity between disciplines. Discipline similarity is calculated by conditional co-occurrence. The hierarchical clustering shows the clustered structure with 3 clusters.
Even though the snapshot matrixes are close to the aggregated proximity matrix, there are still some differences in discipline classification over time, especially during the first two time periods when the publication data is sparse. There are around 60 disciplines have inconsistent classifications with the aggregated data during the first two time periods (1973-1977 and 1978-1982). The discipline relatedness network is divided into four clusters during 1978-1982. The number of inconsistent disciplines quickly decreases to 30 in the third time period (1983-1987). There are 20 disciplines have different classifications with the aggregated network during 1988-1992. From then on, the number of inconsistencies decreases to around 10. Although we observe that the aggregated network shares higher similarity with the most recent data, the network structure appears to be stable as early as 1983-1987. Therefore, we believe the aggregated network captures the general structure behind discipline relatedness. The change of discipline relatedness over time and the precise reasons behind the temporal change are topics for future research.

To confirm the robustness of the cluster structure, in addition to Leiden algorithm, we also apply Infomap(7) to detect the community structure. Infomap gives similar cluster classifications while it further breaks clusters to subclusters. Here we use the community structure obtained by Leiden algorithm as a benchmark to illustrate the results of Infomap. Infomap partitions network to 5 clusters. Cluster 1 contains 43 disciplines and all of them belong to Natural cluster under Leiden algorithm. Cluster 2 contains 36 disciplines and all of them are identified to Physical cluster under Leiden algorithm. Infomap breaks Societal cluster into 2 subclusters. The first subcluster consists of social science disciplines (e.g., Law, Education, History, and Sociology). The second subcluster contains medical disciplines (e.g., Acoustics, Cancer, Hematology, and Pathology). The only difference between the Societal cluster gained from Leiden algorithm and the two
subclusters gained from Infomap is Social Studies of Medicine is classified to the Natural cluster in Leiden algorithm while it is classified to the social science subcluster in Infomap. The fifth cluster in Infomap contains 4 disciplines: Anatomy & Morphology, Dentistry, Fertility and Pharmacology. As shown here, the overall structure is robust across different community detection algorithms. We use the result of the Leiden algorithm because of its higher modularity, interpretability, and simplicity.

**Evolution of Countries**

It is widely believed that developing applied science (*Physical* cluster) will contribute to economic growth. To investigate whether the developmental trap of low-income countries is related with the lack of development in the *Physical* cluster, we compare the economic growth of countries with different developmental trajectories by aggregating countries with their initial cluster specialization and the most recent cluster specialization. Countries are assigned to a single cluster based on its cluster level specialization in each time period (see Method). As we see from Fig S6, the majority of countries have been developing within *Natural* cluster. Countries started with *Natural* cluster and end up with *Physical* cluster have in average the highest GDP growth which is consistent with our regression results that the amount of publication in Physical cluster significantly predicts GDP growth rate. Countries have been developing within *Societal* have the lowest growth rate potentially due to the fact that many of them are rich countries with slowing growth.

*Figure S6. National scientific evolution: (a) cluster transition distribution by aggregating with their initial cluster specialization and the most recent cluster level specialization. Countries are color coded based on their latest GDP value. N, P, S stand for Natural cluster, Physical cluster and Societal cluster respectively for instance: P-S represents countries started from Physical cluster and ended up with Societal cluster. Numbers indicate the number of countries in each transition group classification. Due to the availability of GDP data, only 121 countries are included. (b) Nine countries are selected to illustrate cluster transition scenario as showed in panel (a).*
One noticeable outlier in panel b is the cluster level specialization of United Arab Emirates mismatch with its simplex position. Based on our dataset, United Arab Emirates have only six publications in the initial time period where the six publications are evenly distributed in the three clusters. Therefore, United Arab Emirates is assigned to Societal cluster based on the cluster level relative advantage meanwhile it locates closer to Physical cluster within the simplex. The discrepancy stems from different aspects the cluster level specialization and simplex visualization are measuring. Cluster level specialization takes the sheer number of publications within each cluster into calculation while the position within simplex is decided by the number of advantaged disciplines within each cluster.

To better understand the development trap of low-income countries, we investigate how countries with different economic power have been moving among different clusters. As we can see from Fig. S7, most groups have been moving away from the Physical cluster, likely due to the strong emphasis of Physical sciences and Engineering by emerging countries like China. In particular, the center of the low-income countries has been moving even more into the Natural cluster, capturing the increasing pattern of specialization.

![Figure S7. Average evolution trajectory of countries across income groups. Colors are color coded by time which recent time is represented by dark color.](image)

**The law of proximity and null model**

The constraining force of revealed clusters is further corroborated by the null model that is constructed using the law of proximity. The null model significantly underestimates the number
of newly activated disciplines in each dominant cluster, for instance, the null model predicts fewer newly activated disciplines in the Physical cluster when the country currently possesses relative advantage in the Physical cluster. The underestimation is particularly significant for countries that show advantage in the Natural cluster. In other words, the constraining force of the Natural cluster may be stronger than that of the other clusters.

![Figure S8](image)

*Figure S8 Difference between the number of actual activated disciplines and predicted activated disciplines in each cluster. Countries are aggregated by the cluster-level classification.*

**Scientific Diversity**

The increase in scientific diversity across income groups is coupled with the increase in number of publications over time. To investigate whether the scientific diversity growth is caused by the increased number of publications, we created a simulated research portfolio for every country during each time period by resampling the actual number of publications from the initial publication portfolio—the research portfolio of countries during 1973-1977. We further measure the GINI value of the simulated research portfolio. As shown in Fig S9, the simulated portfolio is far more skewed than the actual portfolio. The difference between the resampled data and actual data indicates scientific diversity growth comes from an increasing balanced research profile. The diversity difference across income groups is not related with the difference in the size of scientific enterprise. The actual GINI value here is slightly different with the GINI value in the main text. Due to the infeasibility to normalize publication count by world average, the sampled GINI value is derived directly from the actual number in each discipline in countries. To make a fair comparison, the actual GINI value here is also derived directly from publication count instead of RCA value.
Figure S9. Difference of scientific diversity between the simulated research profile and the actual research profile. Countries are aggregated by the income-level classification.
## Regression Models

Table S3 Regression results of predicting publication growth

|                | (1)          | (2)          | (3)          | (4)          | (5)          |
|----------------|--------------|--------------|--------------|--------------|--------------|
| Dependent variable:   |              |              |              |              |              |
| Publication Growth (log-ratio) |              |              |              |              |              |
| GDP             | 0.105***     | 0.279***     | 0.228***     | 0.286***     |
|                 | (0.004, 0.205)| (0.193, 0.365)| (0.146, 0.309)| (0.200, 0.373)|
| ECI             | -0.058**     | -0.021       | -0.011       | -0.020       |
|                 | (-0.109, -0.007)| (-0.064, 0.021)| (-0.051, 0.030)| (-0.062, 0.023)|
| Num.Pub         | -0.410***    | -0.381***    | -0.362***    |
|                 | (-0.445, -0.375)| (-0.423, -0.339)| (-0.411, -0.312)|
| Num.Natural     |              | -0.283***    |
|                 |              | (-0.362, -0.205)|                   |
| Num.Physical    |              | -0.016       |
|                 |              | (-0.076, 0.044)|                   |
| Num.Societal    |              | -0.054       |
|                 |              | (-0.121, 0.013)|                   |
| Diversity       |              |              | -0.192       |
|                 |              |              | (-0.456, 0.072)|                   |
| Observations    | 839          | 1,503        | 839          | 826          | 839          |
| R²              | 0.012        | 0.290        | 0.317        | 0.283        | 0.319        |
| Adjusted R²     | -0.168       | 0.174        | 0.192        | 0.147        | 0.193        |

Note: *p<0.1 **p<0.05 ***p<0.01
Table S4 Regression results of predicting diversity growth

|                | (1)    | (2)     | (3)      | (4)      | (5)     |
|----------------|--------|---------|----------|----------|---------|
| Dependent variable: Diversity Growth |         |         |          |          |         |
| GDP            | -0.143 | 0.331** | 0.270*** | 0.418*** |         |
|                | (-0.481, 0.194) | (0.022, 0.641) | (0.105, 0.435) | (0.112, 0.725) |         |
| ECI            | -0.078 | 0.023   | 0.086**  | 0.039    |         |
|                | (-0.249, 0.094) | (-0.131, 0.177) | (0.005, 0.167) | (-0.112, 0.191) |         |
| Num.Pub        | -1.858*** | -1.037*** | -0.802*** |         |
|                | (-2.136, -1.581) | (-1.189, -0.885) | (-0.978, -0.625) |         |
| Num.Natural    |         | -0.088  |         | (-0.246, 0.071) |         |
| Num.Physical   |         | -0.288*** |         | (-0.410, -0.166) |         |
| Num.Societal   |         | -0.229*** |         | (-0.365, -0.093) |         |
| Diversity      |         |         |         | -2.329*** | (-3.269, -1.389) |
| Observations   | 839    | 1,503   | 839      | 826      | 839     |
| R²             | 0.002  | 0.118   | 0.204    | 0.219    | 0.230   |
| Adjusted R²    | -0.179 | -0.027  | 0.058    | 0.071    | 0.087   |

*Note:* *p<0.1 **p<0.05 ***p<0.01
References

1. C. R. Sugimoto, L. Vincent, *Measuring research: What everyone needs to know* (Oxford University Press, 2018).

2. P. Mongeon, A. Paul-Hus, The Journal Coverage of Web of Science and Scopus: a Comparative Analysis. *Scientometrics* 106, 213–228 (2016).

3. A. Éric, V.-G. Étienne, C. Grégoire, L. Vincent, Y. Gingrasb, Benchmarking scientific output in the social sciences and humanities: The limits of existing databases. *Scientometrics* 68, 329–342 (2006).

4. G. Sivertsen, B. Larsen, Comprehensive bibliographic coverage of the social sciences and humanities in a citation index: an empirical analysis of the potential. *Scientometrics* 91, 567–575 (2012).

5. World Bank, World Development Indicators. (2019). GDP (current US$). https://data.worldbank.org/indicator/NY.GDP.MKTP.CD?locations=1W.

6. World Bank. World Development Indicators. (2019). World Bank Country and Lending Groups. Retrieved from https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups.

7. M. Rosvall, C. T. Bergstrom, Maps of random walks on complex networks reveal community structure. *Proc. Natl. Acad. Sci.* 105, 1118–1123 (2008).

8. S. Albeaik, M. Kaltenberg, M. Alsaleh, C. A. Hidalgo, Improving the Economic Complexity Index. *Preprint at https://arxiv.org/abs/1707.05826* (2017).

9. C. A. Hidalgo, R. Hausmann, The building blocks of economic complexity. *Proc. Natl. Acad. Sci.* 106, 10570–10575 (2009).