nature portfolio

https://doi.org/10.1038/s41562-022-01367-x

Supplementary information

The latent structure of global scientific development

In the format provided by the authors and unedited

Supplementary Information for

The latent structure of global scientific development

Lili Miao, Dakota Murray, Woo-Sung Jung, Vincent Larivière, Cassidy R. Sugimoto, Yong-Yeol Ahn
Corresponding author: Yong-Yeol Ahn
Email: yyahn@iu.edu

This PDF file includes:

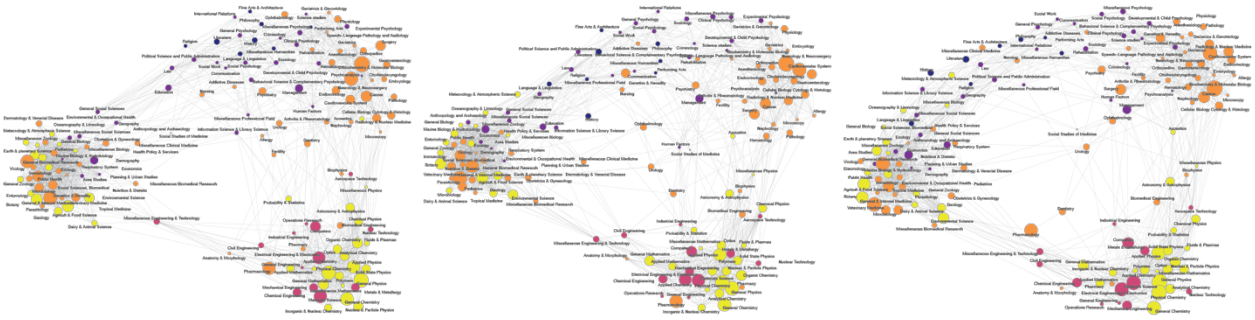
Supplementary Information

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 the 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 Supplementary Figure 1). 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 Supplementary Table 1). The major discrepancy between full counting network and corresponding network happens in *Natural* cluster and *Societa*l 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.



Full Counting Fractional Counting Corresponding Counting

Figure 1. ***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 1. 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	N	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 supplementary figure 2-3. 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. This growth in papers and journals is not associated with more journals from national (or local) communities, but, rather, from the "international" research community. As shown in Figure 1 of a related paper (2), the bulk of new journals actually come from major commercial/international publishers (Elsevier, Springer, Wiley, mostly), but also Sage, Taylor and Francis, and ACS. Conversely, national journals and publishers are accounting for a smaller share of our data over time.

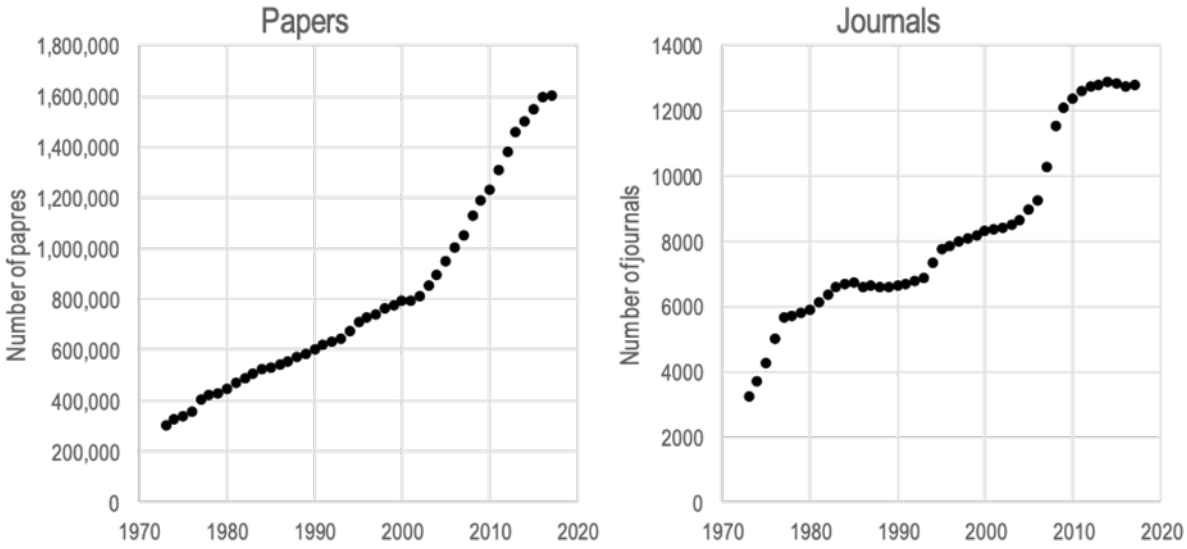


Figure 2. ***Coverage of the Web of Science over time****. Number of papers (articles, notes and reviews) and number of journals indexed within the Web of Science database, for each year between 1973 and 2017.*

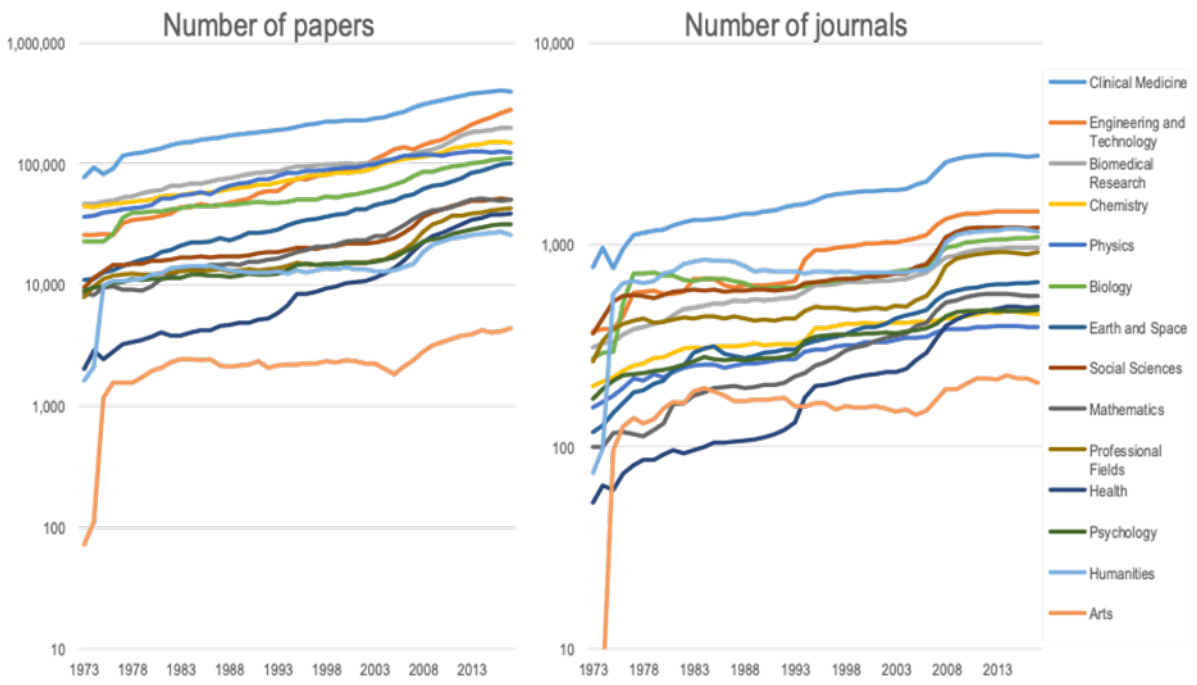


Figure 3. ***Disciplinary coverage of the Web of Science over time***. *Number of papers (articles, notes and reviews), and number of journals indexed, by each of 14 high-level disciplinary categories, for each year between 1973 and 2017.*

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(3). 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(4, 5) 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(6, 7). 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. 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-

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 2. 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

Table 3 Number of countries in income group during time period.

	H	UM	LM	L
1988-1992	35	27	71	52
1992-1993	40	26	62	64
1998-2002	42	33	54	62
2003-2007	45	39	56	53
2008-2012	54	53	52	36
2013-2017	60	56	50	31

Disciplinary Relatedness Network

The aggregated disciplinary proximity matrix is showed in supplementary figure 4. 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 supplementary figure 5).

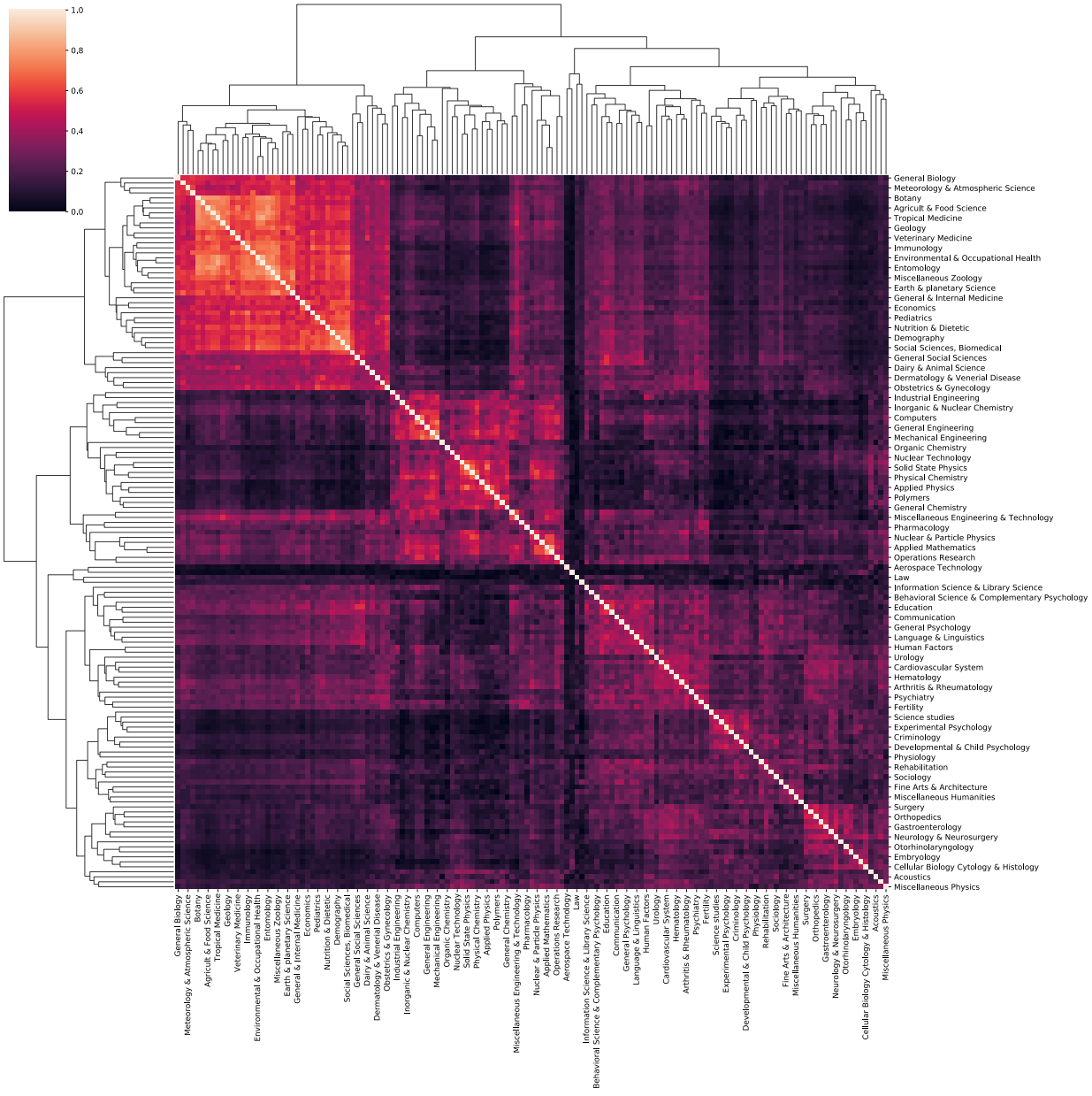


Figure 4. ***Disciplinary proximity matrix****. A 143 x 143 matrix records the pairwise similarity between disciplines. Discipline similarity is calculated by conditional co-occurrence. The hierarchical clustering shows the clustered structure with 3 clusters.*

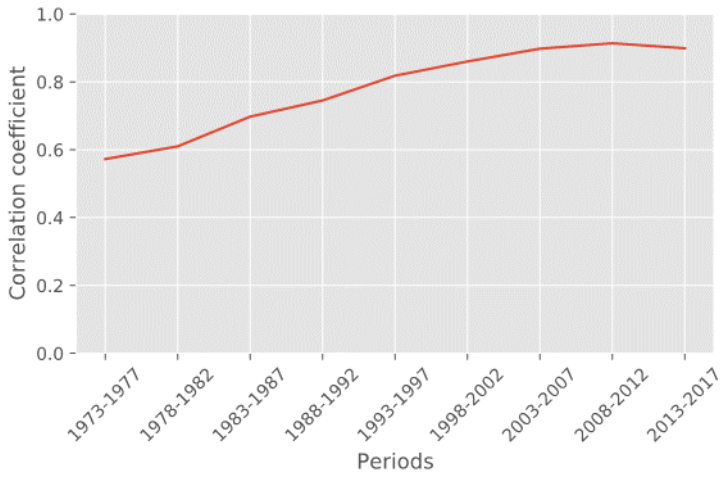


Figure 5. Similarity between the aggregated network and temporal snapshots.

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*(8)* 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 supplementary figure 6, 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.

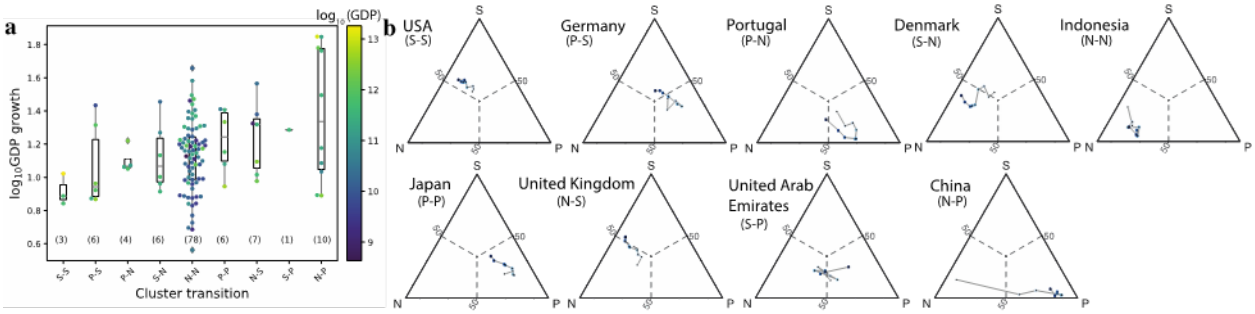


Figure 6. ***National scientific evolution****. (a) cluster transition distribution by aggregating with their initial cluster specialization and the most recent cluster level specialization. Countries are color coded based on their latest GDP value. N, P, S stand for Natural cluster, Physical cluster 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. The center line indicates the median value, upper bound and lower bound of the boxplot represent the 75th percentile and 25th percentile of the data. Upper whisker and lower whisker represent the maximum and minimum value in the dataset. Points locate outside whiskers are outliers which they are lower than the minimum or greater than the maximum. (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 supplementary figure 7, 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.

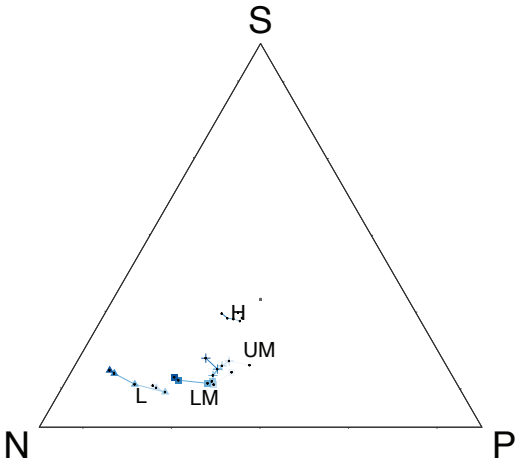


Figure 7. ***Average evolution trajectory of countries across income groups***. *Colors are color coded by time which recent time is represented by dark color.*

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 (see supplementary figure 8). 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.

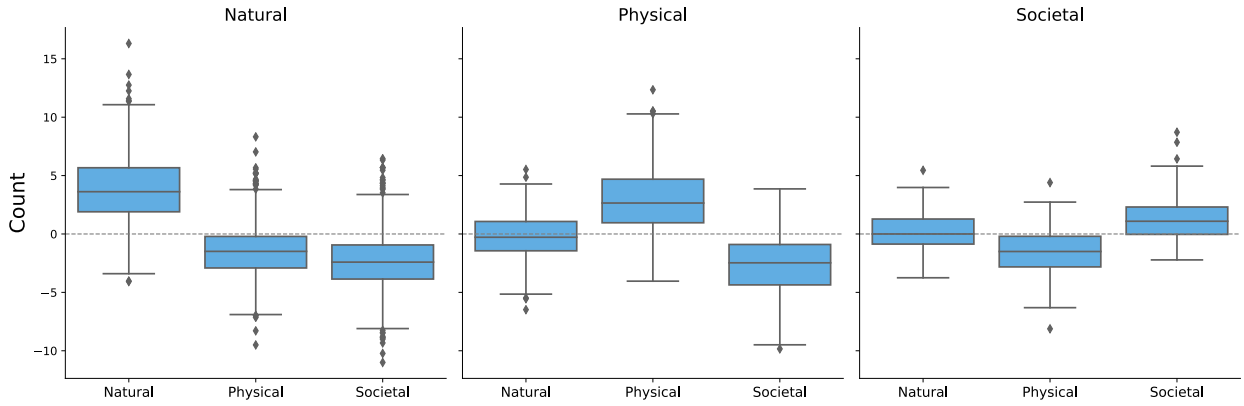


Figure 8. ***Difference between the number of actual activated disciplines and predicted activated disciplines in each cluster****. Countries are aggregated by the cluster-level classification. In total, 212 countries are included in the analysis. The center line indicates the median value, upper bound and lower bound of the boxplot represent the 75th percentile and 25th percentile of the data. Upper whisker and lower whisker represent the maximum and minimum value in the dataset. Points locate outside whiskers are outliers which they are lower than the minimum or greater than the maximum.*

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 supplementary figure 9, the simulated portfolio is far more skewed than the actual portfolio. A single-sided t-test is performed to test whether the actual GINI value is smaller than the simulated GINI value at country level (t=17.02, P=0). 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.

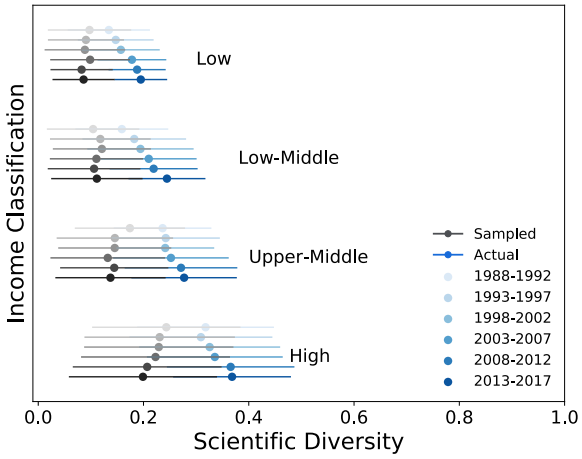


Figure 9. ***Difference of scientific diversity between the simulated research profile and the actual research profile***. *Countries are aggregated by the income-level classification. Point represents the Gini mean value of each income group during each period. Error bars represent the 95% confidence interval of the mean value drawn from bootstrapping. The number of countries in income group during time period is presented in Supplementary Table S3. 1000 times of iterations are used to compute the confidence interval.*

Regression Models

Table 4 Regression result of predicting growth rate of publications

	Dependent variable:			
	Publication growth (log-ratio)			
	(1)	(2)	(3)	(4)
Log GDP	0.10**		0.29***	0.30***
	(0.003, 0.20)		(0.21, 0.38)	(0.21, 0.38)
	p = 0.05		p = 0.00	p = 0.00
ECI	-0.06**		-0.01	-0.01
	(-0.11, -0.01)		(-0.06, 0.03)	(-0.06, 0.03)
	p = 0.03		p = 0.50	p = 0.52
Log Population	0.03		0.46***	0.45***
	(-0.17, 0.23)		(0.29, 0.63)	(0.28, 0.62)
	p = 0.79		p = 0.0000	p = 0.0000
Log no.Pub		-0.41***	-0.41***	-0.40***
		(-0.45, -0.38)	(-0.45, -0.37)	(-0.45, -0.35)
		p = 0.00	p = 0.00	p = 0.00
Diversity				-0.10
				(-0.36, 0.16)
				p = 0.44
Observations	837	1,503	837	837
R^2	0.01	0.29	0.34	0.34
Adjusted R^2	-0.17	0.17	0.22	0.22
F Statistic	2.80** (df = 3; 706)	527.92*** (df = 1; 1290)	92.42*** (df = 4; 705)	74.01*** (df = 5; 704)
Note:	The P-value is derived from a two-sided t-test. *p<0.1;**p<0.05;***p<0.01			

Table 5 Regression result of predicting growth rate of GDP without data of China

	Dependent variable:						
	GDP growth (log-ratio)						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Log GDP	-0.44***	-0.45***	-0.45***	-0.44***	-0.43***	-0.44***	-0.45***
	(-0.50, -0.38)	(-0.51, -0.39)	(-0.51, -0.39)	(-0.50, -0.38)	(-0.49, -0.37)	(-0.50, -0.38)	(-0.51, -0.39)
	p = 0.00	p = 0.00	p = 0.00	p = 0.00	p = 0.00	p = 0.00	p = 0.00
ECI	0.01	0.003	0.003	0.01	0.01	0.01	0.001
	(-0.02, 0.04)	(-0.03, 0.03)	(-0.03, 0.03)	(-0.02, 0.04)	(-0.02, 0.04)	(-0.02, 0.04)	(-0.03, 0.03)
	p = 0.73	p = 0.87	p = 0.83	p = 0.69	p = 0.59	p = 0.69	p = 0.96
Log Population	0.17***	0.14**	0.15**	0.16**	0.18***	0.16**	
	(0.05, 0.29)	(0.02, 0.27)	(0.03, 0.27)	(0.03, 0.28)	(0.05, 0.30)	(0.03, 0.28)	
	p = 0.005	p = 0.03	p = 0.02	p = 0.02	p = 0.01	p = 0.02	
Log no.Pub		0.03					0.04**
		(-0.01, 0.06)					(0.01, 0.08)
		p = 0.12					p = 0.03
Log no.Natural			0.02			0.01	
			(-0.01, 0.05)			(-0.04, 0.05)	
			p = 0.14			p = 0.84	
Log no.Physical				0.02		0.02	
				(-0.01, 0.05)		(-0.03, 0.06)	
				p = 0.16		p = 0.45	
Log no.Societal					0.01		
					(-0.03, 0.04)		
					p = 0.74		
Diversity							-0.06
							(-0.25, 0.13)
							p = 0.55
Observations	828	828	828	820	819	820	828
R^2	0.23	0.23	0.23	0.23	0.22	0.23	0.23
Adjusted R^2	0.09	0.09	0.09	0.08	0.08	0.08	0.08
F Statistic	70.00*** (df = 3; 698)	53.25*** (df = 4; 697)	53.15*** (df = 4; 697)	51.19*** (df = 4; 689)	49.79*** (df = 4; 688)	40.91*** (df = 5; 688)	51.69*** (df = 4; 697)

Note: The P-value is derived from a two-sided t-test. *p<0.1;**p<0.05;***p<0.01

Table 6 Regression result of predicting growth rate of GDP per capita without data of China

	Dependent variable:						
	GDP per capita growth (log-ratio)						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Log GDP per capita	-0.29***	-0.30***	-0.30***	-0.30***	-0.29***	-0.30***	-0.25***
	(-0.34, -0.24)	(-0.35, -0.25)	(-0.34, -0.25)	(-0.35, -0.26)	(-0.34, -0.24)	(-0.35, -0.26)	(-0.29, -0.20)
	p = 0.00	p = 0.00	p = 0.00	p = 0.00	p = 0.00	p = 0.00	p = 0.00
ECI	0.01	0.01	0.01	0.01	0.01*	0.01	0.01*
	(-0.003, 0.03)	(-0.005, 0.03)	(-0.004, 0.03)	(-0.004, 0.03)	(-0.001, 0.03)	(-0.004, 0.03)	(-0.002, 0.03)
	p = 0.12	p = 0.18	p = 0.16	p = 0.14	p = 0.08	p = 0.14	p = 0.09
Log Population	-0.13***	-0.15***	-0.14***	-0.15***	-0.13***	-0.15***	
	(-0.20, -0.06)	(-0.22, -0.07)	(-0.22, -0.07)	(-0.23, -0.08)	(-0.20, -0.05)	(-0.23, -0.08)	
	p = 0.0004	p = 0.0002	p = 0.0002	p = 0.0002	p = 0.002	p = 0.0002	
Log no.Pub		0.01					0.01
		(-0.01, 0.03)					(-0.01, 0.03)
		p = 0.17					p = 0.24
Log no.Natural			0.01			0.002	
			(-0.01, 0.03)			(-0.02, 0.03)	
			p = 0.22			p = 0.89	
Log no.Physical				0.01		0.01	
				(-0.01, 0.03)		(-0.02, 0.03)	
				p = 0.28		p = 0.52	
Log no.Societal					-0.003		
					(-0.02, 0.01)		
					p = 0.70		
Diversity							-0.11**
							(-0.21, -0.01)
							p = 0.03
Observations	810	810	810	803	802	803	810
R^2	0.19	0.19	0.19	0.19	0.19	0.19	0.18
Adjusted R^2	0.04	0.04	0.04	0.04	0.03	0.04	0.02
F Statistic	52.69*** (df = 3; 681)	40.05*** (df = 4; 680)	39.93*** (df = 4; 680)	40.56*** (df = 4; 673)	39.46*** (df = 4; 672)	32.41*** (df = 5; 672)	37.09*** (df = 4; 680)

Note: The P-value is derived from a two-sided t-test. *p<0.1;**p<0.05;***p<0.01

Table 7 Regression result of predicting scientific diversity

	Dependent variable:		
	Diversity growth		
	(1)	(2)	(3)
Log GDP	-0.14	0.34**	0.43***
	(-0.48, 0.20)	(0.03, 0.65)	(0.12, 0.74)
	p = 0.43	p = 0.04	p = 0.01
ECI	-0.08	0.02	0.04
	(-0.25, 0.09)	(-0.13, 0.18)	(-0.11, 0.19)
	p = 0.38	p = 0.78	p = 0.62
Log no.Pub		-1.04***	-0.80***
		(-1.19, -0.89)	(-0.98, -0.62)
		p = 0.00	p = 0.00
Diversity			-2.33***
			(-3.27, -1.39)
			p = 0.0000
Observations	837	837	837
R^2	0.002	0.20	0.23
Adjusted R^2	-0.18	0.06	0.09
F Statistic	0.80 (df = 2; 707)	60.48*** (df = 3; 706)	52.70*** (df = 4; 705)
Note:	The P-value is derived from a two-sided t test. *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. V. Larivière, S. Haustein, P. Mongeon, The Oligopoly of Academic Publishers in the Digital Era. *PLOS ONE* **10**, e0127502 (2015).

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

4. 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).

5. 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).

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

7. , 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.

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