School of Economics Working Paper
2022-08

University of East Anglia

SCHOOL OF ECONOMICS

A New Empirical Index to Track the Technological Novelty of Inventions: A Sector Level Analysis

Yuan Gao *
Emiliya Lazarova *

*School of Economics, University of East Anglia

School of Economics
University of East Anglia
Norwich Research Park
Norwich NR4 7TJ
United Kingdom
www.uea.ac.uk/economics

A New Empirical Index to Track the Technological Novelty of Inventions: A Sector Level Analysis

Yuan Gao[1]*, Emiliya Lazarova[1]

[1]School of Economics, University of East Anglia; Norwich, United Kingdom.

*Corresponding author. Email: y.gao4@uea.ac.uk

Abstract: We propose the Knowledge Origin Re-Combination Index (KORCI) to measure the ex-ante technological novelty of inventions at the sectoral level. The index is developed through the intertemporal comparison of a sequence of networks, which represents the complex connections between the technological components listed in subsequent cohorts of patent applications in the sector. Using patent data from three sectors, we are the first to document the cyclical nature of the evolution of ex-ante technological novelty. Further investigation into the correlation between KORCI and patent application growth rates suggests that this relation, however, is sector-specific. This suggests that the relation between the degree of ex-ante technological novelty and invention activities depends on the specific drivers of innovation in the sector – whether it is process-based or application-based.

One-Sentence Summary: Using our new Knowledge Origin Re-combination Index, we document the cyclical nature of technological novelty in inventions.

In recent quantitative assessments of technological innovation patent data have been widely used due to their direct relation to inventions - the outcome of scientific and technological research and development (R&D) activities - and their growing availability. Most existing studies employ patent data to construct simple counts and capture inter-temporal relations through patent citations. As Schumpeter (*3*) noted, however, innovation is not merely contained in inventions, and it is defined by the scientific or technological novelty embedded in the inventions and combined with their applications. Thus, one can think of innovation as being comprised of ex-ante technological novelty and ex-post impact on technological developments and wider socio-economic life. In this context, our work contributes to developing a quantitative measure -- *Knowledge Origin Re-Combination Index* (KORCI) -- of the ex-ante novelty content in a cohort of inventions.

The view that the origins of any novel ideas lie in existing knowledge is most often attributed to a famous quote by Isaac Newton "If I have seen further, it is by standing on the shoulders of giants." found in a letter he wrote in 1675. These words have been used to develop an understanding of ex-ante innovation as a complex process of combining existing knowledge in novel ways and incorporating radical new ideas. To distinguish between existing and new knowledge, empirical studies of innovation are usually defined on a specific field of technology. With KORCI, we capture two aspects of ex-ante novelty in a sector: new knowledge origins which can be classified as technological components that have not been used in the sector to date; and knowledge re-combination that captures new ways of combining technological components compared to existing practice in the sector. Our work builds on existing patent-level indicators for the novel combinations of knowledge origins that are based on pairwise links between technological components listed in a patent (*4*, *5*). By tracking only pairwise links, these studies greatly simplify the complex interconnectedness among technological components in a snapshot of the population of patents and would therefore present a biased measure of the degree of re-combination of knowledge origins that occurs in a sector. Instead, we adopt a network approach to represent the current state of usage of technological components. This allows us to detect complex changes in the co-existence of technological components used to classify patents across time and thus to develop a measure of the degree of ex-ante technological novelty of inventions that captures both the degree of re-combination of existing knowledge origins and the new knowledge origins that are introduced in the field of technology.

KORCI has a significant value as a tool for those studying innovation as it allows to build intertemporal trends and conduct historical and cross-sectoral analysis of the process of innovation. We evidence the informativeness and versatility of our index through the discussion of its application to three technology-driven sectors: artificial intelligence (AI), pharmaceutical (PHARM) and computer technology (COMP). The latter two are chosen for the large volumes of patent applications according to the World Intellectual Property Indicators (2020), (*1*), and the former for its fast-growing importance in the world of technology according to the 2019 report of the World Intellectual Property Organization, (*2*). The choice of three distinct sectors allows us to detect technology-specific differences in trends of re-combination intensity. Globally, AI has seen several "winters" and "booms" since 1950s, and has most recently re-emerged circa mid-1990s. With COMP and PHARM being mature yet actively evolving technologies, the comparison across these three fields may also reveal differences in trends due to stage of technological development.

Methodology

We identify an empirical counterpart of a knowledge origin in the primary units in the International Patent Classification (IPC) system and employ those to record the technological components that make up a patent. The IPC scheme is a hierarchic system used by patent authorities to assign technical fields as a patent attribute.[1] In each year, we identify a cohort of patents in a specified field of technology using the WIPO sector definition of associated IPC knowledge origins.[2] We note that the IPC classification is updated periodically by the WIPO to reflect changes brought by technological development[3]. However, any potential inconsistency in how the mapping of technologies onto IPC codes due to the IPC version updates is minimized for two reasons: Firstly, patent documents are reclassified according to the amendment of each revision and all our data is downloaded as a single batch. Secondly, a new version release occurs on January 1 each year since 2010, thus, patents filed in the same year are subject to the same version of IPC. Since we take year of application to define a cohort of patents, our methodology is consistent with the WIPO classification process.

Given a population comprised of cohorts of patents with their associated IPC knowledge origins, KORCI is developed in three stages: network construction, clusters identification, and computation.

Stage 1: Network Construction

We represent every cohort of patents as a network of connected technological components. As in previous works (*6*, *7*), we use the 4-digit level IPC codes, known as subclasses, to identify the nodes of the network. In each network, any two subclasses that are assigned to the same patent are linked through an edge. Unlike other authors, we construct the weight of the edges at the subgroup level which is given at the 8-11-digit level IPC codes: The weight of the edge between any two nodes (subclasses) equals the total number of pairwise links between any subgroups listed under the subclasses in the patent and aggregated over all patents where these two subclasses co-exist. An example of the IPC classification scheme is provided in the Supplementary Material where we also present an illustration of how we compute the weight of the edges in Fig. S1. Thus, the weight of an edge between two subclasses indicates the strength with which these two technological components coexist in the cohort of patents. If two subclasses are not associated with the same patent across the whole cohort, then the edge between the two respective nodes has a weight of 0, i.e., there is no link between the two technological components. By computing the weight of the edges at the subclass level, we can increase the network density and retrieve more accurate information on the concurrent usage of technological components within a cohort of patents compared to the methods used by other

[1] The classification scheme is accessible at https://www.wipo.int/classifications/ipc/en/ (last accessed, X 2022).

[2] Sector definition for Pharmaceuticals and Computer Technology can be found at https://www.wipo.int/edocs/mdocs/classifications/en/ipc_ce_41/ipc_ce_41_5-annex1.doc. Sector definition for the Artificial Intelligence field are described here: https://www.wipo.int/tech_trends/en/artificial_intelligence/patentscope.html (last accessed x 2022).

[3] We use the 2006.01 release for Pharmaceuticals and Computer Technology, and the more recent 2021.01 release for Artificial Intelligence to incorporate the latest updates.

researchers. In our sample, the number of unique subclasses of a patent is between 2 and 3, while the number of unique subgroups of a patent varies from 4 to 6.

This stage results in a complete map of the technological components listed in the patent applications in a specific period in a sector and the strength of their concurrent use across the cohort.

Stage 2: Clusters Identification

We use Piccardi's lumped Markov chains network cluster identification method (*8*) to partition the network of patent applications in a cohort into clusters. These clusters identify groupings of technological components present in a cohort of patents such that the technological components within the same group are more likely to be listed in the same patent than to be listed with one from the other groups. With sufficient network density to form a network partition, each cluster is assigned a *persistence probability,* $\alpha \in [0,1]$, which is related to the weight of the edges among the nodes within the cluster.

Through the comparison of the network partitions of two consecutive cohorts of patents, we aim to identify changes in the clusters of technological components and the introduction of new technological components. If we impose a threshold value of α on the clusters, this may result in a different number of clusters in the networks of subsequent cohorts, and, as a result, overestimate the degree of recombination in the use of technological components. Thus, for intertemporal consistency and comparability, we choose to partition each cohort into a fixed number of clusters. Since clusters may vary in size and associated persistence probability, we include these statistics in the definition of the index. Fig. S2 in Supplementary Material provides an example of the networks constructed in two consecutive periods using data from the PHARM sector.

In a cohort of patents, it is possible to observe subclasses that are uniquely and singularly attributed to a patent application. Such subclasses constitute unconnected nodes in the network, i.e., nodes with edges of weight 0. Such nodes are collected in a cluster of its own to which we refer as c*luster zero* in the cohort network. Given the nature of nodes in cluster zero, the persistence probability of such cluster is 0.

At this stage, we identify clusters of technological components based on the frequency of their concurrent use in the cohort of patent applications. The changes to the groupings of subclasses into clusters over time are complex. We provide a visual example of the evolution of the largest cluster in the PHARM sector in Fig.S3 in Supplementary Material along with descriptive comments.

Stage 3: Computation

For a field of technology, we quantify the technological novelty in a cohort of patents using the structural changes to the network constructed in Stage 1 and the clusters identified in Stage 2 compared to the preceding (reference) cohort. Intuitively, in a network cluster KORCI measures the concentration of subclasses that are clustered together in the reference window. Clearly, if all the subclasses in a cluster of the current network also belong to one cluster in the reference

period, no recombination has occurred. Conversely, if no two subclasses in a current cluster were attributed to the same cluster in the network partition of the reference window, the current cluster represents an entirely novel combination of knowledge origins.

Before we formally define KORCI, we need to introduce some auxiliary notations. We denote by $K = \{K_1, K_2, \dots, K_T\}$ the population of patents filed between the first and last, T, period, where each K_t denotes the set of patents in cohort t. Each patent, $k \in K_t$ is associated with a list of IPC subclasses. Let s be the length of the window $\{t-s, t-s+1, \dots, t\}$ that is used in Stage 1 to construct the network of subclasses and let n be the number of clusters identified in Stage 2. We denote by $N_{t,s} = \{N_{t,s}(0), N_{t,s}(2), \dots, N_{t,s}(n)\}$ the set of $n+1$ clusters in the window $t-s$ to t, where cluster $N_{t,s}(0)$ is cluster zero and contains all nodes that have only 0-weight edges, i.e., the unconnected subclasses. The persistence probability associated with the i[th] cluster in the network partition $N_{t,s}$ is denoted by $p_{t,s}(i)$. Equipped with this notation, we define the KORCI of period t as

$$KORCI_{t,s} = \frac{1}{|N_{t,s}|}\sum_{i=1}^{n} \frac{|N_{t,s}(i)|p_{t,s}(i)}{\sum_{j=0}^{n}\left(\frac{|N_{t,s}(i)\cap N_{t-1,s}(j)|}{|N_{t,s}(i)|}\right)^2} \quad (1)$$

The operator $|\ \ |$ denotes the cardinality of the set, as measured by the number of nodes (subclasses). We follow the convention that the cardinality of the empty set is set equal to 0. Although it is theoretically possible for the denominator to be equal 0, this is not an empirically relevant case given our sector-level focus, as it would require that none of the subclasses associated with the current cohort of patents was associated with a patent in the reference window. Since every field of technology is mapped onto a well-defined subset of IPC codes, the probability that sector-specific patent applications in two consecutive periods do not contain any common IPC codes is 0. Section 3 in Supplementary Material provides an example to demonstrate the application of KORCI computation.

It is important to highlight some important features of KORCI. Firstly, the index is increasing in the intensity of re-combination. This is because the denominator, $\sum_{j=0}^{n}\left(\frac{|N_{t,s}(i)\cap N_{t-1,s}(j)|}{|N_{t,s}(i)|}\right)^2$, is monotonically decreasing in the number of re-combinations that occur in the current window in reference to the preceding one. This is easy to see when one considers the extreme case of no new combinations of subclasses: let cluster i from the partition of the current window, t, be a subset of the subclasses attributed to some cluster, j, from the reference window, $t-1$; then the value of the denominator for cluster i equals 1 as $N_{t,s}(i) \cap N_{t-1,s}(j) = N_{t,s}(i)$. Clearly, the presence of any subclasses in i that are not present in cluster j, which constitute a novel combination of knowledge origins, would result in a value of the denominator smaller than 1 as $N_{t,s}(i) \cap N_{t-1,s}(j) \subsetneq N_{t,s}(i)$.

Similarly, KORCI increases with the introduction of technological components that are new to the sector, i.e. those that have not been used in the patents filed the reference window. This is evident in that the denominator $\sum_{j=0}^{n}\left(\frac{|N_{t,s}(i)\cap N_{t-1,s}(j)|}{|N_{t,s}(i)|}\right)^2$ for cluster i of the current window is lower, the larger the proportion of subclasses in i that are not attributed to any cluster in the reference window is.

Next, KORCI is increasing in the persistence probabilities associated with each cluster in the current partition with the effect being stronger for the larger clusters.

Overall, a larger value of KORCI indicates that more new knowledge origins have been introduced in a sector or existing clusters of technological components in the reference window have been more vigorously re-combined to form more persistent clusters in the current window.

Finally, we note that the value of the index is standardized by using the total number of unique subclasses in the current cohort. As some cohorts of patents are associated with a considerably larger number of subclasses than others, our approach ensures comparability of KORCI across time periods.

We also recommend that the choice of the length of the window, s, in the computation of KORCI depends on the scope of one's study. A longer length (e.g., $s = 5$) is recommended for the analysis of long-term trends to average out short-lived fluctuations. Conversely, where the focus is on temporal variations, a one-period length window can be deemed appropriate. The choice of the number of clusters, n, on the other hand, should be done based on the computed persistence probability such that the smallest cluster in every cohort is still associated with a sufficiently high persistence probability. We illustrate graphically in Fig. S5 in Supplementary Material the changes in the behavior of KORCI under different window lengths (s =1, 3 and 5 years) and number of clusters (n =8, 12, 16 and 20) using PHARM patent data for the period 1980 – 2017.

Empirical Case Study

We use patent application data for the three sectors: AI, COMP, and PHARM. We firstly compare the time trends of KORCI with other established measures of technological novelty within sectors. Next, we discuss the correlations between the series in each sector and demonstrate differences in behavior.

Data

Patent application data from the AI, COMP, and PHARM are sourced from the REGPAT database[4] (*9*) that contains information on patents filed from 1978 to 2019. From the raw data we construct the following data series: patent applications volume and the total number of unique IPC subclasses listed in patent applications in each year in each sector. We note that the correlation between the volume of applications and the number of unique subclasses is very high for all sectors: 0.927 for AI; 0.938 for COMP; and 0.801 for PHARM. We therefore choose to utilize in our case study the series on patent applications volume to construct annual growth rates in applications and use the number of unique IPC subclasses in levels.

We employ the information on IPC subclasses and IPC subgroups in the three-step methodology outlined above to compute sector-specific KORCI. As our methodology requires a large number

[4] We use the REGPAT database released in January, 2020, accessible upon request from the OECD MSTI data dissemination service at: https://www.oecd.org/sti/msti.htm.

of observations with a sufficient number of connections in each cohort to construct a network, we choose the initial period based on the filings being consistently above a certain threshold: 500 for COMP and PHARM and 100 for AI. This allows us to calculate $KORCI_{1,1}$ starting from 1980 for PHARM; 1981 for COMP, and 1982 for AI patents, respectively. We choose 2017 as the end period due to the drop in the number of patent applications in subsequent years in the dataset. We suspect that the drop is due to a delay in processing the application data rather than a decrease in activity.

Time-series analysis

Figure 1 presents trends in patent applications volume, patent application annual growth rate, total number of unique IPC subclasses included in a cohort, and $KORCI_{1,1}$ with a fixed number of clusters n=8 for each of the three sectors under investigation (AI, COMP, and PHARM). The y-axis on each panel measures number of patent applications in hundreds and number of unique IPC subclasses and the secondary y-axis to the right measures growth rates and $KORCI_{1,1}$. Sample period differs for the three sectors and is determined by the data availability on KORCI. Thus, Fig 1 presents data for 1980-2017 for PHARM, 1981-2017 for COMP, and 1982-2017 for AI.

In comparison to the other three series - patent volume, unique IPC subclasses, and growth rates - $KORCI_{1,1}$ exhibits more pronounced cyclical behavior of alternating periods of heightened re-combination and novelty in technological components followed by periods of lower levels in all three sectors. This cyclical feature of the process of ex-ante innovation is therefore omitted in studies that rely on one of the traditional measures of innovation. Although the cycles are not synchronized, the average $KORCI_{1,1}$ across the three sectors are quite similar with COMP having the highest value of 2.002, followed by PHARM with 1.848 and AI with 1.710. The higher degree of re-combination on average observed in the COMP sector versus the PHARM can be explained by the wider range of applications COMP patents have in other technological fields, which is also reflected in the higher number of unique IPC subclasses listed on the COMP patents. Given the growing adoption of AI technologies in many fields, one would expect KORCI for AI to exhibit a similarly high value. The reason AI average KORCI is the lowest may be explained by the fact that AI is a relatively new sector with the penetration of AI technologies in other sectors being relatively recent (see (*2*)).

While KORCI captures a qualitatively different aspect of ex-ante innovation, there are visible co-movements across the other three series plotted in Fig 1. We already reported the high correlation between patent applications quantities and the number of unique IPC subclasses. The high correlation is intuitive as by the very nature of patents as proof of invention, one would expect that they build on unique technological components. Moreover, the correlation is stronger for AI and COMP where inventions are more likely to spill onto other sectors via novel applications compared to PHARM. The wide spread of applications associated with patents in COMP and AI can also explain why the highest numbers of unique IPC subclasses in COMP and AI in a year (362 in COMP and 300 in AI, both in 2016) is higher than that in PHARM (240 in 2002) even though the number of patent applications in PHARM is higher than that of AI in every single year apart from 2017 and the average number of patent applications in COMP and PHARM are comparable: 7,545 (COMP) and 7,335 (PHARM); while the average for AI is considerably lower (1,967). Interestingly, the period in the first half of 2000s, when PHARM

IPC subclasses exhibit a plateau coincides with a wider range of technological components listed in PHARM applications as shown in Fig S3. The latter observations on patent volumes are also reflected in the annual growth rate series. On average, in each sample period, patent applications grow slower in COMP and PHARM (7.88% and 6.38%, respectively) and faster in AI (13.70%).

The cyclical behavior of KORCI exhibited in Fig 1clearly indicates that our index captures a distinct feature of the evolution of technological innovation compared to the other three. It is still unclear, however, how ex-ante novelty is related to the level of innovation activity, if at all. To investigate the co-movements between KORCI and traditional measures of innovation, we proceed to conduct a regression analysis.

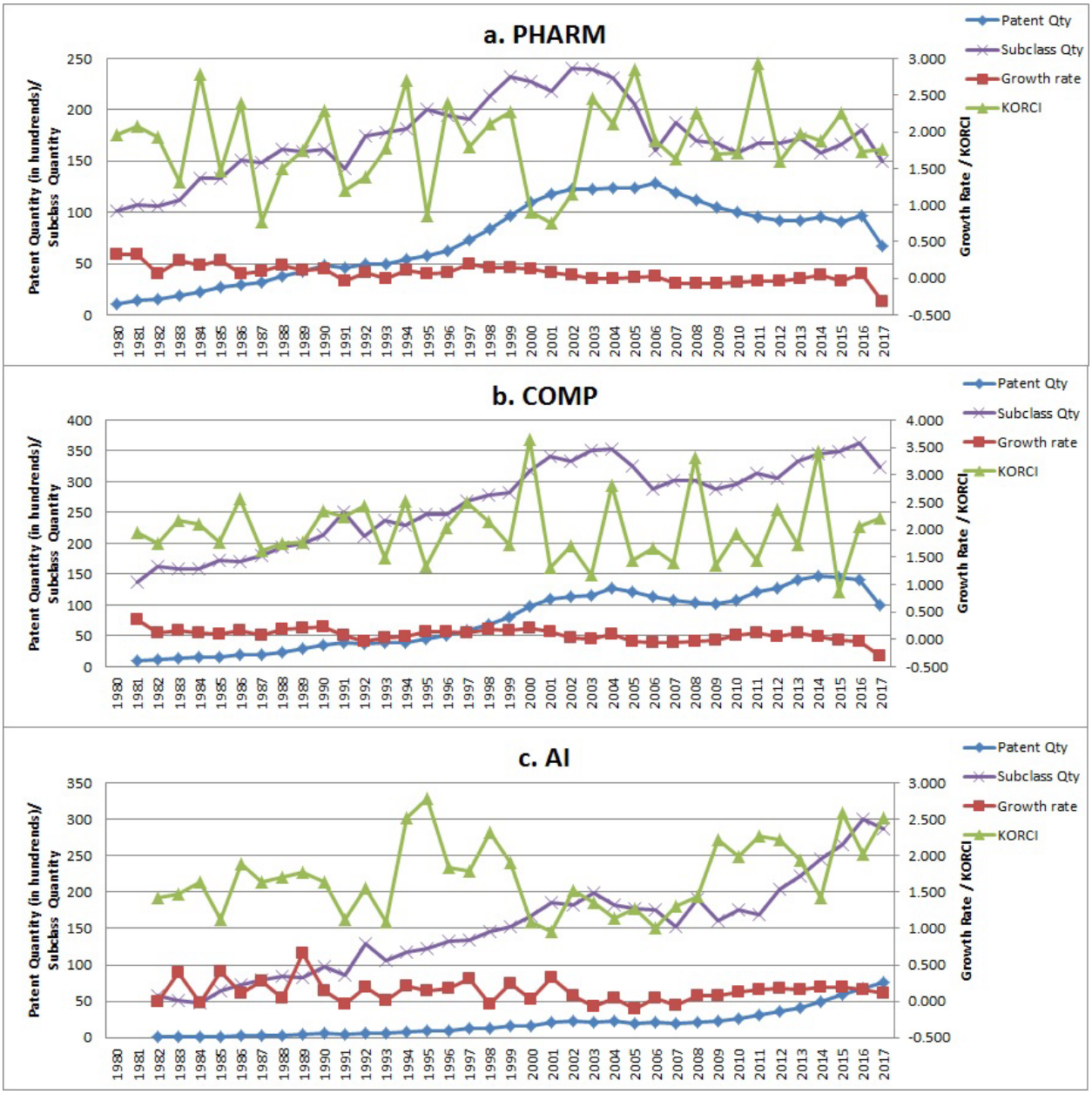


Fig 1 Each panel presents trends for four data series: patent application volume measured in hundreds (blue diamond); number of unique IPC subclasses listed on patent applications (purple

cross); patent application annual growth rate (red square); and $KORCI_{1,1}$ with number of clusters fixed to 8 (green triangle). Panel a. presents the data for PHARM for 1980-2017; Panel b. presents the data for COMP for 1981-2017; and Panel c. presents the data for AI for 1982-2017. The y-axis on each panel measures application volume (in hundreds) and unique IPC subclasses quantity and the secondary y-axis to the right measures growth rates and $KORCI_{1,1}$.

Regression Analysis

Given that our time series are relatively short, we can only estimate very parsimonious models. Based on the figures presented in the previous section and the high correlation between patent volumes and number of unique IPC subclasses, we choose an Autoregressive Distributive Lag type specification as presented below.

$$g_t = \beta_0 + \beta_1 \ln q_{t-1} + \beta_2 KORCI_{1,1,t} + \varepsilon_t , \qquad t = 2, \dots, T \qquad (2)$$

$$\sum_{k=0}^{2} \frac{g_{t-k}}{3} = \beta_0 + \beta_1 \ln q_{t-3} + \beta_2 \sum_{k=0}^{2} \frac{KORCI_{1,1,t}}{3} + \varepsilon_t \;\; t = 4, \dots, T \quad (3)$$

where g_t denotes annual growth rate in patent application numbers and q_t is the number of unique IPC subclasses listed in patent applications in period t. With model (2) we aim to test for short-term correlations between KORCI and the growth rate, In model (3) we use three-year rolling averages of the growth in patent numbers ($\sum_{k=0}^{2} \frac{g_{t-k}}{3}$) as dependent variable, and, correspondingly, we use the three-year rolling average of our recombination index ($\sum_{k=0}^{2} \frac{KORCI_{1,1,t}}{3}$) as regressor, with the aim to identify long-term correlations within this relatively short sample period. We include the lagged log of the number of unique IPC subclasses to capture convergence in applications, thus, we expect the estimate of β_1 to be negative and strongly statically significant. We are agnostic about the sign and significance of β_2. A statistically significant positive estimate of this coefficient would indicate that a higher degree of ex-ante technological novelty in a sector is associated with an expansion in the invention activities. Conversely, a statistically significant negative estimate would provide evidence that when a sector's innovation activities rely more heavily on introducing new knowledge origins or discovering new ways of combining them, the growth in patenting is lower.

We estimate these two models separately for PHARM, COMP, and AI. The regression results are reported in Table 1 where the first two columns present the results for the PHARM sample (1980-2017), the middle two columns present the results for the COMP sample (1981-2017), and the last two columns present the results for the AI sample (1982-2017).

Across the three sectors, we detect stronger statistical significance of the correlation between growth rates in patent applications and KORCI in the model containing three-year rolling annual averages. This is consistent with our graphical examination of the trends that suggested little co-movements in the cyclical components of the series. The results reveal evident sectoral differences. We observe a statistically significant and positive longer-term correlation between growth rates in patent applications and KORCI in PHARM (see column 2) but negative correlation in COMP and AI. For AI the estimated coefficient is strongly significant (see column 6) but for COMP, it is only marginally significant at the 89.5% confidence level (see column 4). For each sector, the correlation signs between growth rates and KORCI in the short-term and long-term models are consistent. The negative correlation in PHARM suggests that exploration

into novel use of knowledge origins is correlated with a decrease in innovation activities. In COMP and AI, on other hand, the positive correlation suggests that a more intense usage of established knowledge origin combinations, i.e. lower KORCI, is correlated with lower growth in patent applications.

Except for the model presented in column 5, across all other estimation we find strong evidence for convergence in growth rates. Interestingly, the rates of converge for PHARM and COMP are very similar which suggests that the two sectors are at the similar stage of maturity on the evolutionary path.

Conclusion

We develop a methodology to track re-combination of existing technological components and the introduction of new ones through the cohorts of patent applications in a sector. We build on that to propose a new tool to measure the ex-ante technological novelty in inventions in a sector that captures the degree of novel use of knowledge origins in inventions. Through the application of our index to data on AI, COMP, and PHARM, we are the first to empirically document the cyclicity in the ex-ante technological novelty in the evolution of innovation. It is notable that the cyclical patterns are robust to the choice of reference window length and the number of clusters in which each period is partitioned. Moreover, a similar cyclical pattern in the KORCIs is evident in all three technological sectors AI, COMP, and PHARM, which suggests that this is not a sector-specific phenomenon. This feature is compatible with alternating periods of intensive technological innovation (when inventions occur through the further exploration of existing knowledge clusters) and of extensive innovation (when novel (combinations of) knowledge origins underpin inventions).

There are important sectoral differences, however, in the co-evolution of KORCI and patent applications growth as revealed through our regression analysis. In PHARM the results suggest that a high degree of re-combination is associated with a lower growth in patent applications and in AI and COMP the correlation is positive. The results may be driven by the different nature of innovation in these sectors. Despite these differences, we observe a similar rate of convergence between COMP and PHARM suggesting that these sectors are of similar maturity; as expected AI is distinct in this respect. The documented differences across the three sectors suggest a promising research agenda on the factors that drive these patterns using KORCI.

References

(1) World Intellectual Property Organization, (WIPO), *World Intellectual Property Indicators 2020*. (Geneva, World Intellectual Property Organization, 2020; https://www.wipo.int/edocs/pubdocs/en/wipo_pub_941_2020.pdf).

(2) World Intellectual Property Organization (WIPO), *WIPO Technology Trends 2019: Artificial Intelligence.* (Geneva, World Intellectual Property Organization, 2019; https://www.wipo.int/edocs/pubdocs/en/wipo_pub_1055.pdf) .

(3) J.A. Schumpeter, *The Theory of Economic Development* (Routledge, New York, ed. 1, 2017).

(4) D. Verhoeven, J. Bakker, R. Veugelers, Measuring technological novelty with patent-based indicators. *Research Policy.* **45**, 707–723 (2016).
(5) D. Silvestri, M. Riccaboni, A. Della Malva, Sailing in all winds: Technological search over the business cycle. *Research Policy*. **47**, 1933–1944 (2018).
(6) Y. Gao, Z. Zhu, M. Riccaboni, Consistency and trends of technological innovations: A network approach to the international patent classification data. *International Conference on Complex Networks and their Applications*, (Springer, 744–756 2017).
(7) Y. Gao, Z. Zhu, R. Kali, M. Riccaboni, Community evolution in patent networks: technological change and network dynamics. *Applied Network Science*. **3** 26 (2018). https://doi.org/10.1007/s41109-018-0090-3
(8) C. Piccardi, Finding and testing network communities by lumped Markov chains. *PloS ONE* **6** e27028. https://doi.org/10.1371/journal.pone.0027028
(9) S. Maraut, et al., The OECD REGPAT Database: A Presentation, *OECD Science, Technology and Industry Working Papers*, No. 2008/02 (2008), OECD Publishing, Paris, https://doi.org/10.1787/241437144144.

Acknowledgements:

We thank Prof. Corrado Di Maria (University of East Anglia) for critical review and comments on the manuscript.

Data and Materials Availability:

Original patent data used in this manuscript is available upon request from the OECD MSTI REGPAT database. Data used to generate Fig. 1 and Fig. S5 and for the regression analysis are included in the supplementary materials.

List of Supplementary Materials:

Supplementary Materials:
Supplementary Text
Materials and Methods
Fig. S1 to S5
Table S1

Auxiliary Supplementary Materials:

Dataset for figures (Fig. 1 and Fig. S5), available at:
https://www.dropbox.com/scl/fi/i2io2ha7qvwhmjltqh82r/dataset-for-figures.xlsx?dl=0&rlkey=ol5npnythf4oeaga1oj43plhc

Dataset for regression, available at:
https://www.dropbox.com/s/dexwntak4qgstg8/dataset%20for%20regression.dta?dl=0

Table 1. Regression results of the correlation between KORCI and patent application growth rates

	PHARM		**COMP**		**AI**	
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
	$\boldsymbol{g_t}$	$\sum_{k=0}^{2} \frac{g_{t-k}}{3}$	$\boldsymbol{g_t}$	$\sum_{k=0}^{2} \frac{g_{t-k}}{3}$	$\boldsymbol{g_t}$	$\sum_{k=0}^{2} \frac{g_{t-k}}{3}$
$\ln q_{t-1}$	-0.224***		-0.229***		-0.093	
	[0.001]		[0.000]		[0.141]	
$\ln q_{t-3}$		-0.183***		-0.179***		-0.112***
		[0.000]		[0.000]		[0.000]
$KORCI_{1,1,t}$	-0.017		0.022		0.0.63	
	[0.464]		[0.362]		[0.156]	
$\sum_{k=0}^{2} \frac{KORCI_{1,1,t-k}}{3}$		-0.069**		0.072		0.114***
		[0.031]		[0.105]		[0.000]
constant	1.239*** [0.000]	1.124*** [0.000]	1.301*** [0.000]	0.923*** [0.000]	0.482 [0.121]	0.482*** [0.000]
N	38	36	37	35	36	34
R-sq	0.229	0.354	0.366	0.502	0.103	0.485
F -stats	7.05*** [0.003]	15.17*** [0.000]	9.36*** [0.000]	19.94*** [0.000]	1.69 [0.200]	13.58*** [0.000]
OLS estimation with robust standard errors p values are presented in square brackets below the coefficient estimates; * p<0.05, ** p<0.01, *** p<0.001; PHARM regressions are estimated with sample 1980-1917 and presented in columns (1) and (2); COMP regressions are estimated with sample 1981-2017; AI regressions are estimated with sample 1982-2017 and presented in columns (5) and (6)						

5

Supplementary Materials

This PDF file includes:

Supplementary Text

An Example of IPC classification

Here we demonstrate the IPC classification scheme using an example.

The top level of the hierarchic structure is Section. There are eight of Sections. As an example, we take Section A which is labelled “Human Necessities; Agriculture”.

Under each Section, there are Classes. For example, in Section A, Class A61 is for “Medical or Veterinary Science; Hygiene”.

The next level consists of Subclasses. Continuing with our example, A61K is a subclass covering “Preparations for Medical, Dental, or Toilet Purposes”.

And finally, the bottom level, Subgroup. As an example, we take Subgroup A61K 48/00 which refers specifically to “Medicinal preparations containing genetic material which is inserted into cells of the living body to treat genetic diseases; Gene therapy”.

Materials and Methods

1. Network Construction

The section contains materials to help illustrate the first stage of our methodology: patent network construction.

Fig. S1.

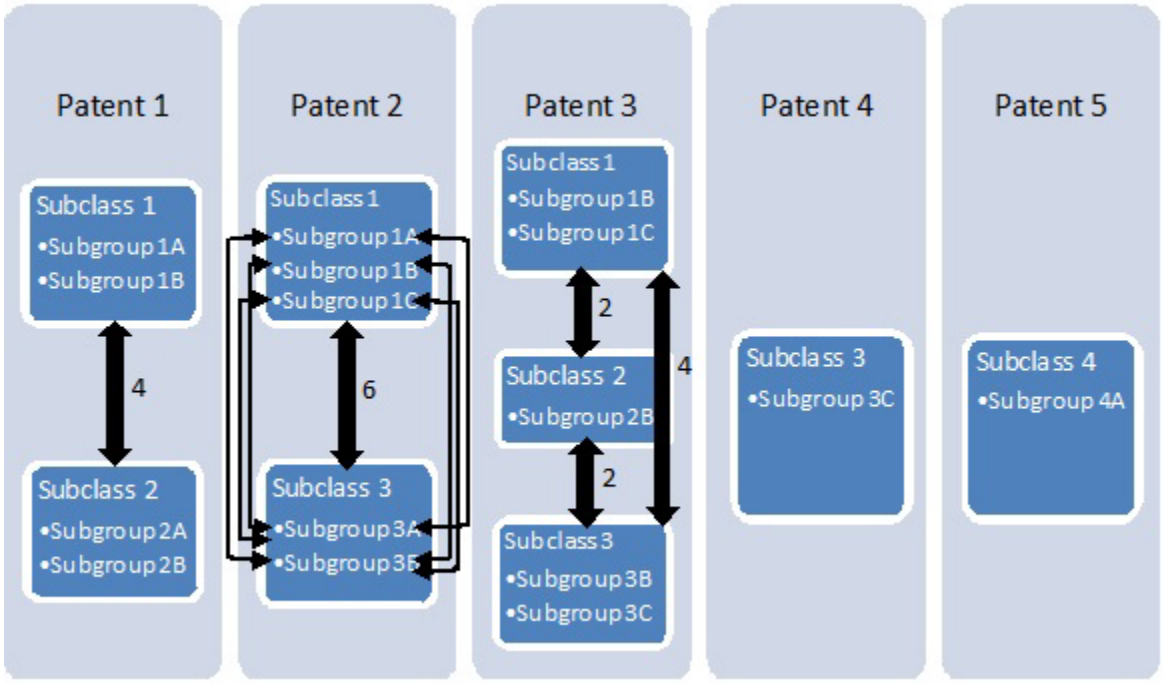


Fig. S1 shows how the weight of edges between two nodes (subclasses) is calculated, using an example with a cohort of five patents. In Patent 2 the arrows illustrate how the links between the subgroups of Subclass 1 and Subclass 3 are added to sum up to 6 - the weight of the edge between these two subclasses. Following the same method, in Patent 1 there are four pairwise links between the subgroups of Subclasses 1 and 2; in Patent 3, there are two pairwise links between the subgroups of Subclasses 1 and 2 and between the subgroups of Subclasses 2 and 3, and 4 pairwise links between the subgroups of Subclasses 1 and 3. Patens 4 and 5 contain only one subclass each, and therefore, there are no additional links between any subclass generated from these two patents. We list the subgroup-level edge weights within each patent as a complete network matrix in Table S1. For example, the weight of the edge between Subclass 1 and Subclass 3 based on the links existing in Patent 2 is 6, and the weight based on the links in Patent 3 is 4. Therefore, the aggregated weight of the links between these two nodes in this cohort is 10. This example also demonstrates some nodes in isolation from the others. This happens when a given subclass is not attributed to any patents in the network concurrently with any other subclasses. An example is Subclass 4. Subclass 3, however, is not isolated because although it is the only Subclass of Patent 4, it is connected to Subclass 1 in Patent 2.

Table S1.

		Subclass 1			Subclass 2		Subclass 3		
	Subgroup	1A	1B	1C	2A	2B	3A	3B	3C
Subclass 1	1A				1 (1,0,0,0)	1 (1,0,0,0)	1 (0,1,0,0)	1 (0,1,0,0)	0 (0,0,0,0)
	1B				1 (1,0,0,0)	2 (1,0,1,0)	1 (0,1,0,0)	2 (0,1,1,0)	1 (0,0,1,0)
	1C				0 (0,0,0,0)	1 (0,0,1,0)	1 (0,1,0,0)	2 (0,1,1,0)	1 (0,0,1,0)
Subclass 2	2A	1 (1,0,0,0)	1 (1,0,0,0)	0 (0,0,0,0)			0 (0,0,0,0)	0 (0,0,0,0)	0 (0,0,0,0)
	2B	1 (1,0,0,0)	2 (1,0,1,0)	1 (0,0,1,0)			0 (0,0,0,0)	1 (0,0,1,0)	1 (0,0,1,0)
Subclass 3	3A	1 (0,1,0,0)	1 (0,1,0,0)	1 (0,1,0,0)	0 (0,0,0,0)	0 (0,0,0,0)			
	3B	1 (0,1,0,0)	2 (0,1,1,0)	2 (0,1,1,0)	0 (0,0,0,0)	1 (0,0,1,0)			
	3C	0 (0,0,0,0)	1 (0,0,1,0)	1 (0,0,1,0)	0 (0,0,0,0)	1 (0,0,1,0)			

This table provides a matrix representation of the network corresponding to the cohort of patents in Fig. S1. The table contains a detailed breakdown showing the weight of the edge between each pair of subgroups from two different subclasses. In each cell, the value outside of the brackets is the number of links between the two subgroups which equals the sum of all the values inside the brackets. The values inside the brackets represent the number of links attributed to each patent listed in order from Patent 1 to 4. The table doesn't include Subclass 4 because no other subclasses co-exist with it in any other patent in the cohort.

Fig. S2.

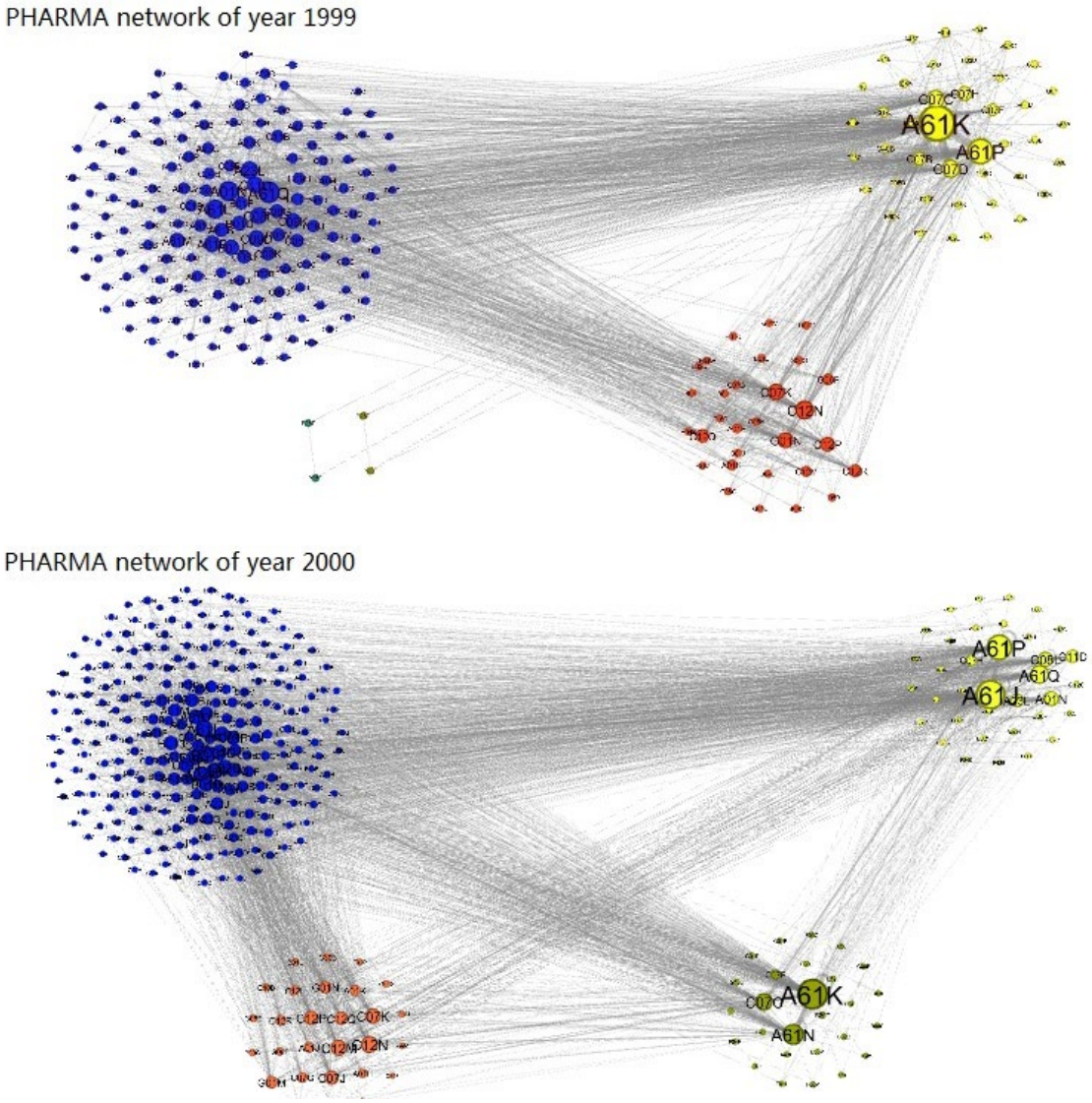


Fig. S2 is a visualization of the network clusters using data of the PHARM patents filed in years 1999 and 2000. The network of each year is constructed using the method introduced in Stage 1, and then divided into 4 clusters using the network partitioning method describes in Stage 2. We use colors to indicate the cluster size: blue for the largest cluster, yellow for the second largest, orange - the third largest, and olive green - the smallest. Node size is in proportion to node degree, the number of edges the node has to other nodes. Edges in the diagram only indicate the existence of a link between two nodes and do not represent the weights.

We can draw several insights from Fig. S2. First, the largest cluster doesn't necessarily contain the most well-connected nodes. Subclass A61K is the defining subclass of the PHARM sector, which means any patents in this sector must contain at least one subgroup under A61K. It therefore has the highest degree by default. However, grouping A61K into the largest cluster does not necessarily yield the highest *persistence probability*. Our clustering method focuses on the combined usage of technological components in inventions and not the ranking of core technologies by any simple measures. Second, we can see the changes in clustering from 1999 to 2000. Subclass A61K and A61P are both in the yellow cluster in 1999, but in the next year they are found in two different clusters. The clustering method allows us to capture such recombination. This example also shows that the clusters size distribution varies. The smallest cluster in 1999 is much smaller than the largest, while in 2000 the clusters are more evenly distributed. This is why we take cluster size into consideration when computing KORCI in Stage 3.

Fig. S3.

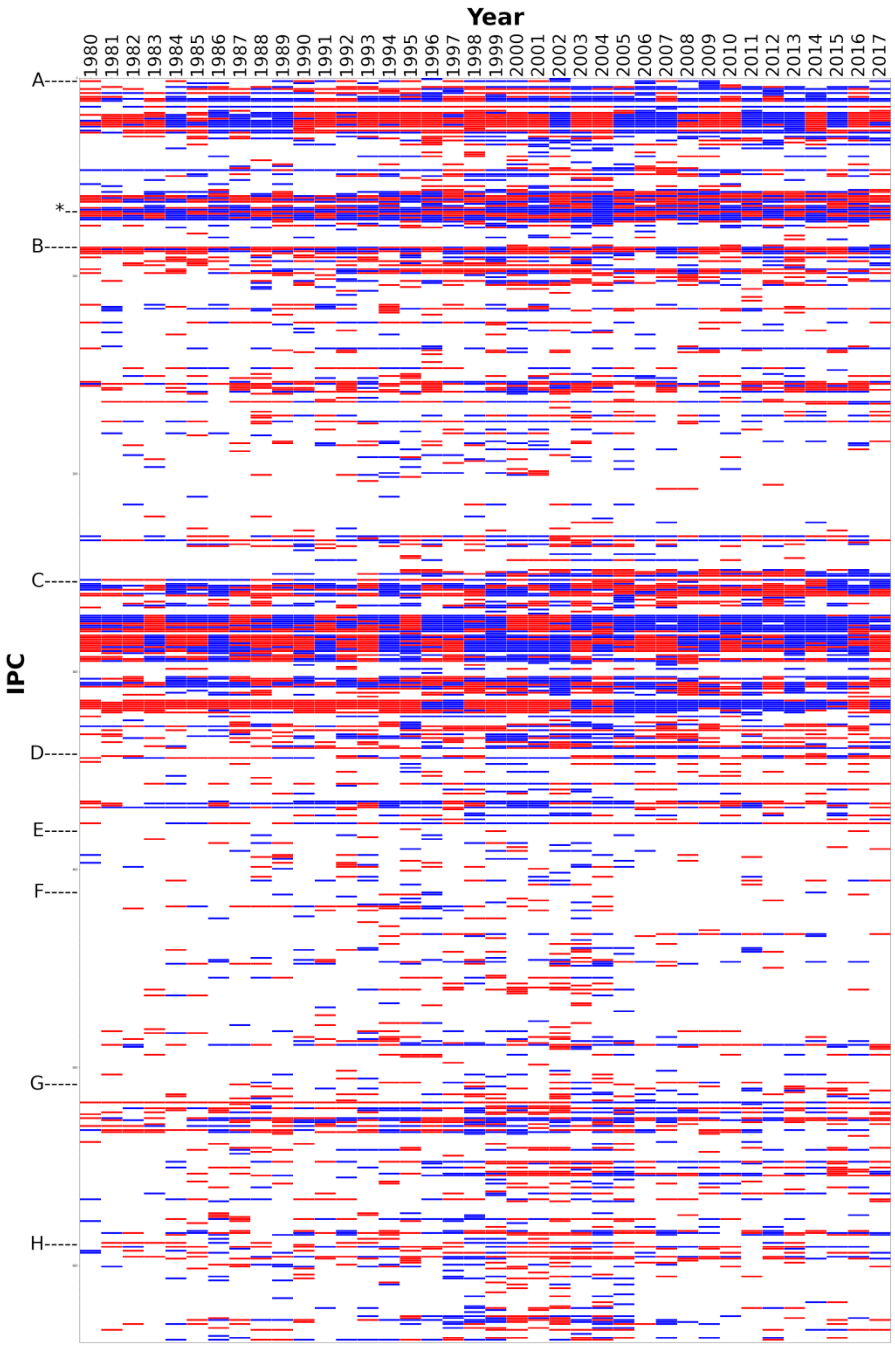

Year
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
IPC
A
*
B
C
D
E
F
G
H

Fig. S3 presents a different view of the network partition changes over time. Here we use data of PHARM patent applications for the period 1980-2017 to illustrate the evolution of network re-combinations by focusing on the composition of the largest cluster in the network of any given year and all other subclasses in the cohort. All the IPC subclasses in the IPC scheme at the time of data extraction are listed on the vertical axis where we use markers A through to H to indicate the technological sections as defined in the IPC scheme (https://www.wipo.int/classifications/ipc/en/) and with an asterisk (*) we indicate the IPC subclasses that define the sector – in this case A61K for PHARM according to the sector definition (https://www.wipo.int/edocs/mdocs/classifications/en/ipc_ce_41/ipc_ce_41_5-annex1.doc). The year when the application of the patent took place is listed on the horizontal axis. In the graph, an IPC subclass is colored in red in any given year when this subclass is associated with the largest cluster identified in Stage 2 (network clustering) of our methodology. An IPC subclass is colored in blue in any given year when this subclass is a member of another cluster different from the largest one in the same network. A blank space indicates that a given IPC subgroup is not attributed to any PHARM patent application made in the corresponding year.

Fig. S3 clearly shows that while there are certain IPC subclasses that are present in every cohort of PHARM patent applications, these "permanent" subclasses spread across different technological sectors: human necessities including agriculture, foods, and health and life saving (Sector A), performing operations and transporting (Sector B), chemistry and metallurgy (Sector C), and physics including optics, computing and checking instruments (Sector G). Moreover, among the permanent subclasses, there is not even one that is constantly associated with the largest cluster throughout the entire period. The changing red-blue-blank pattern provides a visual representation of the re-combination activities in the evolution of ex-ante innovation in PHARM patent applications in this period. In addition, we can identify years with wider spread of colored spaces (late 1990s – early 2000s) indicating exploration across more diverse knowledge origins; and, other years when colored spaces are concentrated along fewer and adjacent lines suggesting that innovation more intensively use fewer knowledge origins of a smaller collection of different categories.

2. KORCI Computation

In Stage 3 of our methodology, we compute KORCI as the quantitative measure of technological novelty in a cohort of patents by comparing its network partition to that of the previous cohort, with the partitions identified through Stage 1 and 2. This section provides a simple example to illustrate the computation of KORCI.

Fig. S4

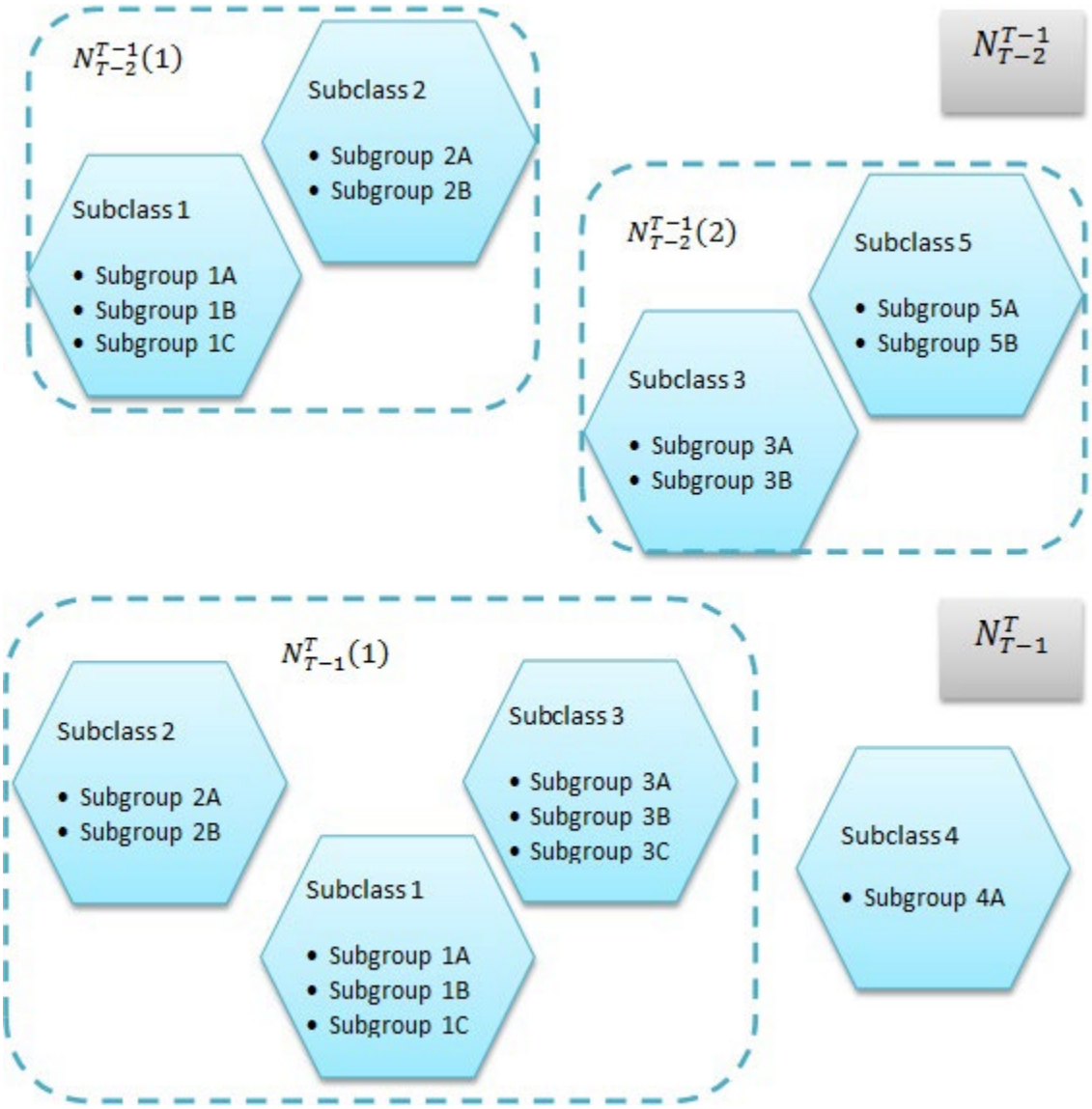


Fig. S4 demonstrates how to measure cluster re-combination of a network constructed from a cohort of patents in reference to the network constructed from patents of the previous time period, as described in Equation 1 in Stage 3 of the Methodology section. $\boldsymbol{N_{T-1}^{T}}$ represents the network partition of the 1-year time window T-1 to T, including four nodes out of which one cluster ($\boldsymbol{N_{T-1}^{T}(1)}$) is identified through Stage 2. $\boldsymbol{N_{T-2}^{T-1}}$ represents the reference network partition of the previous time window, consisting of two clusters (($\boldsymbol{N_{T-2}^{T-1}(1)}$) and ($\boldsymbol{N_{T-2}^{T-1}(2)}$) developed from four nodes. The figure shows that Subclass 5 from the reference network no longer exists in $\boldsymbol{N_{T-1}^{T}}$, whereas a new node, Subclass 4 emerges. From $\boldsymbol{N_{T-2}^{T-1}}$ to $\boldsymbol{N_{T-1}^{T}}$, Subclass 1 and Subclass 2 remain closely connected, and their links with Subclass 3 have become strong enough to result in

a new network partition where Subclass 3 has joined Subclass 1 and 2 to form $\boldsymbol{N_{T-1}^{T}(1)}$ and $\boldsymbol{N_{T-2}^{T-1}(2)}$ no longer exists. Following the definition of KORCI, the denominator of the first component in the primary sum in Equation 1 will be 5/9, shown as below:

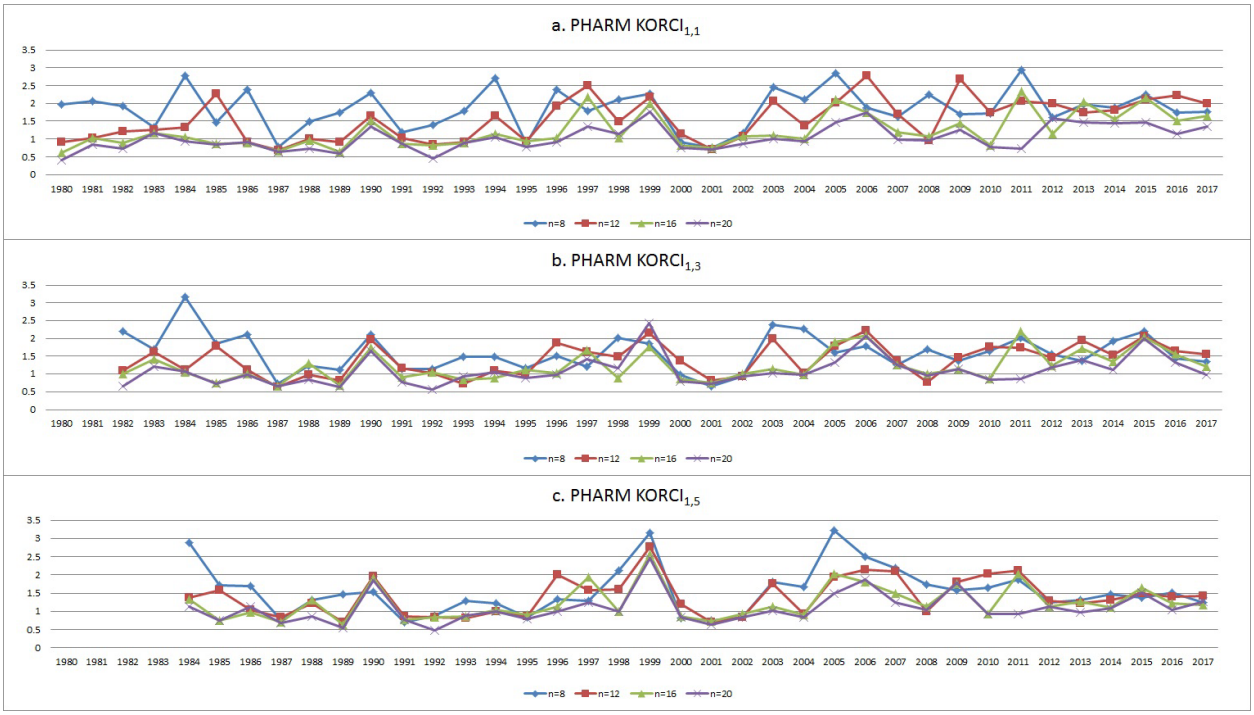
$$KORCI_{T,1} = \frac{1}{4}\left(\frac{3\,p_{T,1}(1)}{\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2} + 0\right)$$

$p_{T,1}(1)$ is the persistence probability of $N_{T-1}^{T}(1)$.

3. KORCI Robustness Analysis

When computing KORCI a researcher needs to set the values of two key parameters: the number of clusters in each partition, n, and the length of the reference window, s. Fig. S5 illustrates the behavior of KORCI under different specifications for n and s using data of PHARM patent applications for the period 1980 – 2017. Under the IPC scheme the subclasses are grouped into eight sections at the top hierarchy level. We are therefore interested in network partitions at the same level (n =8) or finer. The maximum number of clusters that we present is 20 as above this level the network partition will be too fragmented and re-combination will be over-represented. We present reference time window lengths with s =1, 3 and 5 years to illustrate that the use of longer reference windows results in smoothing-out short-term fluctuations.

Fig. S5



In Fig. S5, each panel presents KORCI calculated with PHARM data 1980-2017 using four different cluster levels n=8 (blue diamond); n=12 (red square); n=16 (green triangle); and n=20 (purple cross). Panel a. presents $KORCI_{1,1}$ 1980-2017; panel b. presents $KORCI_{1,3}$ 1982-2017; panel c. presents $KORCI_{1,5}$ for 1984-2017.

In all three panels we observe that a smaller number of clusters, n, tends to result in a higher index of knowledge re-combination as the blue-diamond curves lie above the others at most data points. Indeed, KORCI computed with n=8 has the highest average values in all three panels and the index computed with n=20 exhibits the lowest (1.848 vs 1.025 for $KORCI_{1,1}$; 1.602 vs 1.010 for $KORCI_{1,3}$; and 1.558 vs 1.088 for $KORCI_{1,5}$). Intuitively, this can be explained by the fact that a partition with a larger total number of clusters would have a lower level of persistence

probability with smaller clusters. We can also compare average KORCI values across the three panels. On the one hand, one could expect to see the highest KORCI average value for $KORCI_{1,1}$ and the lowest for $KORCI_{1,5}$ irrespective of the number of clusters because a shorter reference window has a smaller knowledge origin base ($N_{t,s}$). On the other hand, longer reference window is more likely to result in a different clustering of the origin base to the current one and thus we may observe a higher average value for $KORCI_{1,5}$. In fact we do not detect a regular pattern in our data. (For n=8 and n=16, we have average $KORCI_{1,1}$> $KORCI_{1,3}$> $KORCI_{1,5}$; for n=16 we have average $KORCI_{1,5}$> $KORCI_{1,1}$> $KORCI_{1,3}$; and for n=20 we have average $KORCI_{1,3}$> $KORCI_{1,5}$> $KORCI_{1,1}$.)

Despite differences in average values, all series exhibit similar features of highs and lows in the observed window, and we can therefore claim that the index is quite robust to changes in these parameter values. Across these different specifications, the KORCI rise to a high in 1999, followed by a downward trend before reaching a low around 2000-2002, and then start to rise again around 2003.