Journal of Evolutionary Economics
https://doi.org/10.1007/s00191-024-00872-8

REGULAR ARTICLE

A new empirical index to track the technological novelty of inventions: A sector-level analysis

Yuan Gao[1] · **Emiliya Lazarova**[1]

Accepted: 6 September 2024

Abstract

We introduce 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. This allows us to quantify the intensity of the recombination of components and the introduction of new ones at the frontier of technological knowledge. Using patent data from three sectors – artificial intelligence, computer technology, and pharmaceuticals – we are the first to document the cyclical nature of the evolution of ex-ante technological novelty of inventions across all three sectors. These evolutionary cycles, however, are not synchronized, and therefore it is unlikely that they are driven by a common innovation engine. Further investigation into the correlation between KORCI and patent growth rates reveals other differences among the sectors in both direction and strength. We conjecture that the relation between the degree of ex-ante technological novelty and invention activities depends on the specific innovation environment of the sector – whether these are process-based or product-based. Our new tool opens opportunities for new empirical research into the evolution of innovation at the sectoral level.

Keywords Knowledge Origin Re-Combination Index · Ex-ante novelty · Cyclicity · Technological evolution · Patents · Sectoral analysis

JEL Classification O31 · O33 · O34

One-sentence summary We make a methodological contribution by introducing our Knowledge Origin Re-combination Index to measure ex-ante novelty at the sectoral level, which allows us to document the cyclical nature of the evolution of technological novelty of inventions and uncover differences in trends across technological fields.

✉ Yuan Gao
y.gao4@uea.ac.uk

[1] School of Economics, University of East Anglia, Norwich, UK

Published online: 10 October 2024

1 Introduction

Technological innovation, one may argue, is the evolutionary engine of technological knowledge. As Schumpeter (1934) noted, innovation is not merely contained in inventions; it is defined by the scientific or technological novelty *embedded* in the inventions, on the one hand, and the new value generation through the adoption of the inventions, on the other. Following in Schumpeter's footsteps, scholars distinguish between the *ex-ante* technological novelty given purely by the technological knowledge contained in an invention and the ex-post impact of an invention on the wider socio-economic life. In this dichotomy, 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.

We would argue that the study of the evolution of technological knowledge depends on the availability of appropriate tools to map existing technological space and identify patterns of change through time. The objective of our work is to offer one such methodological tool that is designed to enable researchers to gain a comprehensive view of the complex relations between technological components at the frontier of knowledge and track the changes to these relations brought about by inventions, known as *ex-ante technological novelty.*

Gaining such tools is an essential first step in the study of frontier knowledge in a technological field. Having a comprehensive, coherent, and aggregate measure of technological novelty content in the cohort of inventions related to a sector enables a researcher to conduct cross-sectoral analysis that uses a common denominator in the measure of novelty. It opens opportunities for the study of the impact of policy and industrial strategy at the national level and detects induced changes to technological change or differential impact across technological sectors. Our approach allows us both to capture the complex relations in a static technology space and to extract intertemporal changes to consecutive technological frontiers.

In recent quantitative assessments of innovation and the evolution of technology spaces, 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 like applications growth rates or patent citations (Griliches et al. 1986; Fleming 2001; Hall et al. 2000; Jaffe and Trajtenberg 2002).

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. We aim to provide an empirical counterpart to this conceptual framework.

We build KORCI as a three-step quantitative method to measure the intensity of the re-combination of technological components and new technological

components at a technological frontier given to us by a cohort of inventions in a sector. The first two steps utilize tools from network theory. These tools are valuable in themselves, as they allow researchers to visualize the complex relations between technological components in a cohort of inventions. First, using data on patent applications filed during a period in a certain technology sector, we construct a network based on the technological subclasses used to categorize the cohort of patents. The connections in the network are based on the co-assignment of subclasses to patents, where the weight of any connection is the result of an aggregation over all the patents in the cohort. Second, we identify clusters in the network consisting of strongly connected technological subclasses that are more likely to be co-assigned to the same groups of patents. In the third step, we define KORCI as an empirical measure of the differences in the network clusters of two consecutive time periods. Our measure tracks changes in the compositions of clusters (knowledge recombination) and the introduction of new technological components that were not used in the previous period (new knowledge origin).

KORCI has a significant value as a tool for those studying innovation, as it allows for building intertemporal trends and conducting historical and cross-sectoral analysis of the process of innovation. Empirical studies of innovation are usually defined on a specific field of technology. This allows researchers to clearly identify when a technological component is introduced in the technological knowledge production of a new sector. In line with these studies, we design KORCI as a sectoral measure.

We evidence the informativeness and versatility of our index through the discussion of its application to three technology-driven sectors: artificial intelligence (AI), pharmaceuticals (PHARM), and computer technology (COMP). The latter two are chosen for the large volumes of patent applications according to the World Intellectual Property Indicators (WIPO) (2020), and the former for its fast-growing importance in the world of technology according to the 2019 report of the World Intellectual Property Organization (2019). 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 the 1950s and has most recently re-emerged circa the 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 the stage of technological development.

The remainder of the paper is organized as follows. In the next section, we discuss the literature background of our research and compare our work to others' work. Next, we present our methodological contribution in the form of a novel empirical measure to capture the degree of re-combinational and novel origins at the technological frontier of a sector. After that, we illustrate the use of KORCI by computing it with data on the AI, COMP, and PHARM sectors. We compare the information that is provided by KORCI with that of traditional measures of innovation activities such as patent counts, growth in application volume, and number of unique subclasses through presenting time trends and simple time-series regression analysis. We conclude that KORCI is a valuable new measure of ex-ante novelty at the sector level, as it captures the cyclical nature of the evolution of technological knowledge at the sectoral frontier and has a non-trivial association with the established invention activities measures.

2 Literature review

We are not the first to utilize network methods in technological innovation studies using patent data. Researchers built networks based on connections defined by various relationships among different technologies to study the structure and changes in innovation. We summarize these works by their network construction methods into two main categories: (1) connections directly based on the referencing relationships, such as patent citations and scientific publication citations (Chang et al. 2009; Fontana et al. 2009; Érdi et al. 2013); and (2) connections defined by shared properties, such as patents ownership (Trapper et al. 2012; Guan and Liu 2016), co-authorship or R&D collaboration (Beaudry and Schiffauerova 2011; Li et al. 2014), and the co-occurrence of technological classification, which is used in this paper. While citation represents one way of technology diffusion, it is reliant on self-reporting and rarely provides an exhaustive list of related works. These are less reliable, however, when the study's objective is to capture patterns of usage of knowledge among innovations in a technological ecosystem. This is also why we choose a different approach. We use the technological classification information as listed on patent applications to retrieve the interconnectedness between inventions. In doing so, we not only represent the knowledge spread through innovation but also how the knowledge origins are used in combination to generate novelty. Moreover, as we explore the technological evolution patterns within a sector, we take inventions and technologies – rather than their owners or inventors – as the fundamental building blocks of our network.

Given the rich information contained in and related to patenting, some researchers combine multiple networks and propose more complex approaches. One example is to analyze the relationship between distance among technologies and geographical proximity to study regional knowledge spillover and R&D collaboration (Rodríguez-Pose and Crescenzi 2008; Colombelli et al. 2014; Gao and Zhu 2022). Broekel (2019), for example, presents the technological complexity trends over a time-series empirical analysis with the complexity of each temporal interval independent of the others. While this is a valuable use of the network complexity measure, we differ from this branch of the literature in that our analysis is interested in measuring novelty, and our method captures the definition of novelty as the new versus the old by temporal comparison.

Among the research works more similar to ours, based on technology co-classification in patents, Verhoeven et al. (2016) and others (Strumsky et al. 2011; Silvestri et al. 2018) developed indicators of novel knowledge origins combinations based on *pairwise links* formed in various manners, such as technological classes co-assigned to the same patent, or assigned to a patent and its prior art, or in patents filed by the same applicant. Our index is most closely inspired by the work of Verhoeven et al. (2016). These authors distinguish between two dimensions of ex-ante technological novelty in a field of research: new knowledge

origins which can be classified as technological components that have not been used in the field to date; and knowledge recombination that captures new ways of combining technology components compared to existing practice in the field.

What distinguishes our work from theirs is that we design our measure as a tool to capture the evolution of technological knowledge at the sectoral level while they focus on the technological novelty in a single patent. Methodologically, we adopt a clustering method to identify patterns of co-assignment of technological components, while they identify re-combination based on pairwise connections only. We argue that tracking only pairwise links is inadequate as far as the characterization of technology space is concerned, as this would greatly simplify the complex interconnectedness among technological components. Firstly, those studies that rely exclusively on the existence of new pairwise links between technology components as an indicator of novelty do not consider the frequency of occurrence of such connections, as shown in Verhoeven et al. (2016). While such an approach can be justified when applied to the individual patent analysis, it will miss out on important information on the aggregate thickness of links at the sector level. Our method measures the strength of connections between technological components through the likelihood of such connections evaluated on the frequency of connections in a group of technological components. This allows us to identify clusters of technologies that are interconnected via stronger ties amongst them compared to their ties with technologies outside the clusters. Secondly, pairwise links would not identify indirect strong connections between subclasses. For example, a technological component C1 can be co-assigned with technological components C2 and C3 with higher frequency but never in the same patent. Thus, a pairwise link between C2 and C3 may not exist in the data and therefore the technological connection between these two clusters will be missed. A clustering algorithm, instead, may assign these two technological components to the same cluster due to their thick connection to C1.

We acknowledge that there are other algorithms to identify network clusters, for example, the modularity optimization method (Blondel et al. 2008). There is no established standard to determine which clustering algorithm is the best. The contribution of our work is not aimed at network cluster detection either. For the purposes of developing our index, we have chosen to use one possible algorithm that meets our needs. We refer interested readers to the robustness checks in this regard presented in the work of Gao (2018).

Therefore, we believe that the network approach proposed in our methodological contribution is more suitable to represent the evolution of technological knowledge in a sector compared to two-dimensional pairwise measures introduced by previous authors. It allows us to detect complex changes across time in a more stereoscopic way and form 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.

To identify an empirical counterpart of a knowledge origin in a sector with technological classifications, we use 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] Different authorities may have their own classification systems, such as the Cooperative Patent Classification (CPC) scheme of the United States Patent and Trademark Office and the File Index (FI) of the Japanese Patent Office. We choose the IPC system because our empirical analysis is carried out using a global patent dataset in which IPC is the international standard. Class sizes are not set as part of the IPC scheme: the number of subclasses in each section varies, and so does the number of subgroups in each subclass. The number of lower hierarchical levels subordinated to a higher hierarchical level depends on the amount of content and extent of segmentation of the technology represented by the higher level. Reclassifications and variances in the distribution of technology category size are prevalent across patent classification systems. As Lafond and Kim (2019) observe in their study on the U.S. patent classification system, these variations are required to accurately categorize inventions according to the up-to-date actual technology spaces.

Choosing the appropriate hierarchical level of the IPC classification in an empirical analysis is a challenge, which has been recognized by other authors. Sasaki and Sakata (2021) construct co-classification networks at different IPC hierarchical levels – subclass, group, and subgroup – and study how upper-level connectedness correlates with lower-level connectedness. They identify the lower level of IPC as richer in information and therefore a more appropriate tool in studies that involve classifications based on technological connectedness. Kay et al. (2014) also recognize the different degrees of informativeness of different hierarchical levels. Due to the significant variation in the number of patents with attributes at each level, these authors suggest that an appropriate hierarchical level should be chosen for each technological field to ensure that it is sufficiently well represented when taking a network approach to mapping technology connections. Consistent with these findings, Souza et al. (2019) and Choi and Yoon (2022) choose to use the subgroup level of the IPC in their network approach to patent data analysis. Informed by these authors' work, we also choose to use the information at the IPC subgroup level to derive a reliable estimate of the connectedness across the knowledge origins coded in IPC subclasses.

In the empirical time-series analysis, we also employ the number of patent applications and associated unique subclasses at the sector level. The number of patents is a natural measure of the volume of innovation activities and as such it has been widely used in the literature. For example, it is used in both the 2019 and 2020 WIPO reports as a key indicator of sectoral innovation performance. The number of unique subclasses at the patent level is used in the literature under the label "Patent Scope". It is often considered to be associated with the technological and economic value of inventions (see, for example, Squicciarini et al. 2013). Similarly, measured at a cohort level, the total number of unique subclasses of a group of patents is taken

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

to indicate the technological breadth. Lerner (1994) found that at the firm level, a broader scope of subclasses is positively associated with the firm value. In our sectoral analysis, we adopt a similar approach and include the collective unique subclass number in our study. We are not aware of any other works that use the number of unique subclasses at a sector level. There is no direct comparability between firm-level and sector-level analysis, as multiproduct firms usually operate across multiple sectors and sector-specific differences may be lost in such analysis.

3 Methodology

We regard inventions as the basic carriers of new knowledge in a technological sector. We take an annual snapshot of patent applications to represent the stock of technological knowledge at the frontier. To develop our empirical measure, we first identify a cohort of patents in a specified field of technology as defined by the WIPO. We note that the IPC scheme is updated periodically by the WIPO to reflect changes brought by technological development.[2] However, any potential discrepancy 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. By downloading the data as a single batch, a researcher can ensure consistency in the applied classification system. Secondly, a new version release occurs on January 1 of each year since 2010, thus, patents filed in the same year are subject to the same version of IPC. Since we take the year of application to define a cohort of patents, our methodology is consistent with the WIPO classification process.

We further note that the differences in the size of technological categories should not introduce biases to the strengths of connections between subclasses. As elaborated in Stage 2 below, we identify network clusters based on the relative probability of a group of subclasses to be co-listed on patent applications in a cohort. The strength of the connections of one subclass to another does not depend on how many different subordinate subgroups are co-assigned with other subclasses in a cohort of patents. Instead, it depends on the frequency with which subordinate subgroups are co-listed. Thus, a subclass with one subgroup can be identified as the element with the strongest connections in a cluster if this subgroup is co-assigned with a large number of subgroups of other subclasses. Equally, a subclass may contain many subgroups but it could be the case that only a few are co-assigned with subgroups of other clusters in a cohort of patent applications. Such a subclass will exhibit weaker connections.

Given a population of patent applications with their associated IPC knowledge origins and indexed by time period, we develop our index – KORCI – in three stages: network construction, cluster identification, and computation. The first two stages are designed to capture the knowledge landscape in a specific period. The

[2] 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.

final stage measures the technological novelty of the frontier, i.e., the data in the most recent annual data, vis-à-vis a well-specified historical benchmark. We present the three stages in detail below.

We denote by $P = \{P_1, P_2, \ldots, P_T\}$ the population of patents filed between the first and last, T, period of annual data, where each P_t denotes the set of patents in the cohort of year t. Each patent, $i \in P_t$ is associated with a list of four-digit level IPC codes called *subclasses* and, to the finer level of classification, a list of 8 to 11-digit level IPC codes known as *subgroups*. The collection of all unique IPC subclasses listed on patent applications filed in the period from $t - s$ to t is denoted $C(P_{t-s}, \ldots, P_t)$; similarly, all the associated unique subgroups are $G(P_{t-s}, \ldots, P_t)$, for $s = 0, \ldots, t-1$ and $t = 1, \ldots, T$.

3.1 Stage 1: Network construction

We represent the technological frontier encoded in a collection of patent applications as a network of connected technological components. As in previous works (Gao 2018; Gao et al. 2018a, b), we use subclasses to mark the nodes of the network[3] and define the weight of the edges using the information gathered at the subgroup level. Thus, the technological frontier derived from the cohorts of patents$P_{t-s}, \ldots, P_t$, with $s = 0, \ldots, t-1$ and t= $1, \ldots, T$, is represented by a network whereby the set of nodes is equivalent to the set of subclasses$C(P_{t-s}, \ldots, P_t)$; and edges exist between any two nodes if the subclasses they represent are co-listed on a patent application in this set. We denote the resulting network $\Gamma(C(P_{t-s}, \ldots, P_t), G(P_{t-s}, \ldots, P_t)) \equiv \Gamma^t_{t-s}$, with $s = 0, \ldots, t-1$ and t = $1, \ldots, T$; Γ^t_{t-s} is a shorthand notation. We clarify that a network, as defined hereby, may be derived from an annual range as wide as the whole historical population ($s = t-1$ and$t = T$) or as narrow as a single yearly cohort of patent applications ($s = 0$ and$t = 1, \ldots, T$).

We construct the weight of the edges at the subgroup level using the information in the set $G(P_{t-s}, \ldots, P_t)$, for patents in the sequence of cohorts $P_{t-s}, \ldots, P_t$ with $s = 0, \ldots, t-1$ and t = $1, \ldots, T$. The weight of the edge between any two nodes (subclasses) equals the total number of pairwise links between any subgroups listed under the subclasses on an application and aggregated over all patents where these two subclasses are co-listed.

We use the patent application data presented in Fig. 1 to illustrate how we construct the network and compute the weight of the edges between any two nodes. The corresponding network representation is seen in Fig. 2. Figure 1 presents a cohort of four patents all filed in a year t: $P_t = \{P1, P2, P3, P4\}$; four

[3] Here we demonstrate the IPC classification scheme using an example. The top level of the hierarchic structure is Section. There are eight Sections. As an example, we take Section A which is labeled "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".

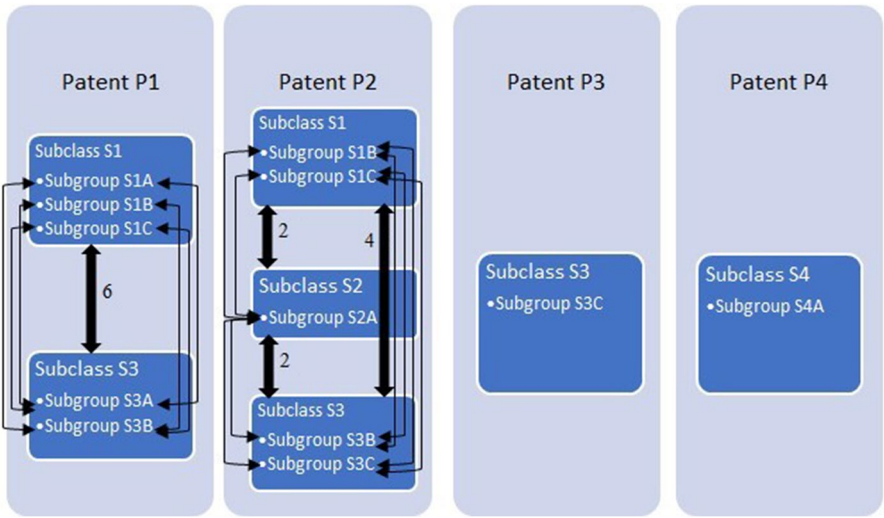

Fig. 1 An example of a cohort consisting of four patents $\boldsymbol{P_t} = \{\boldsymbol{P1}, \boldsymbol{P2}, \boldsymbol{P3}, \boldsymbol{P4}\}$; four unique subclasses $\boldsymbol{C(P_t)} = \{\boldsymbol{S1}, \boldsymbol{S2}, \boldsymbol{S3}, \boldsymbol{S4}\}$; and eight unique subgroups; $\boldsymbol{G(P_t)} = \{\boldsymbol{S1A}, \boldsymbol{S1B}, \boldsymbol{S1C}, \boldsymbol{S2A}, \boldsymbol{S3A}, \boldsymbol{S3B}, \boldsymbol{S3C}, \boldsymbol{S4A}\}$. The *thin arrows* point to all pairwise links between subgroups of distinct subclasses listed under a patent. The *bold arrows* point to pairwise links between subclasses listed under a patent and the numbers next to them provide the total pairwise connections between these subclasses (the number of thin arrows that go between the two subclasses)

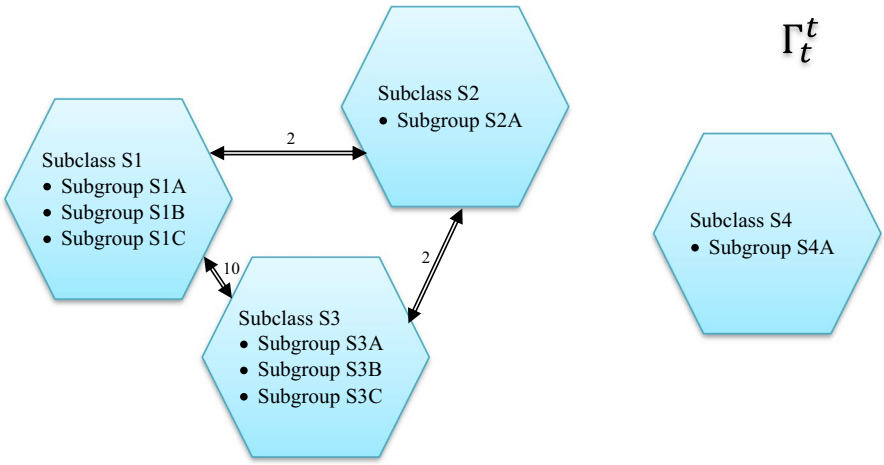

Fig. 2 The network corresponding to the patent cohort in Fig. 1. Each *hexagon* represents a node in $\boldsymbol{C(P_t)} = \{\boldsymbol{S1}, \boldsymbol{S2}, \boldsymbol{S3}, \boldsymbol{S4}\}$. The *arrows* represent the edge between two nodes that are connected in the network, and the *numbers* next to them indicate the weight of the edge

unique subclasses $C(P_t) = \{S1, S2, S3, S4\}$; and eight unique subgroups; $G(P_t) = \{S1A, S1B, S1C, S2A, S3A, S3B, S3C, S4A\}$. This is represented in Fig. 2 by a network with four nodes corresponding to the four unique subclasses:

$S1, S2, S3, S4$. The following pairs of subclasses are co-listed on at least one patent: $S1$ and $S3$ (co-listed on $P1$ and $P2$);$S2$ and $S3$ (co-listed on $P2$). Subclass $S4$ is listed only on patent $P4$ where it is the only subclasses listed on the application. Therefore, in Fig. 2, there are edges only between nodes $S1$ and $S3$, $S1$ and $S2$, and between $S2$ and $S3$; and node $S4$ is a singleton, i.e., unconnected. In Fig. 1, the thin arrows point to all pairwise links between subgroups of distinct subclasses listed under a patent. The bold arrows point to pairwise links between subclasses listed under a patent and the number next to them provides the total pairwise connections between the subgroups listed under these subclasses (number of thin arrows). Starting with $P1$, we note that there are three subgroups ($S1A, S1B, S1C$) listed under subclass $S1$ and two subgroups ($S3A, S3B$) listed under subclass $S3$; therefore, there are six pairwise links between subclasses $S1$ and $S3$ associated with patent application $P1$. To these six pairwise links, we need to add the pairwise links between $S1$ and $S3$ listed on patent application $P2$ where there are two subgroups ($S1B, S1$C) listed under subclass $S1$ and two subgroups ($S3$B, $S3$C) listed under subclass $S3$; therefore, there are four pairwise links between subclasses $S1$ and $S3$ on patent application $P2$. In Fig. 2, these calculations are reflected by allocating weight ten to the edge between the nodes $S1$ and $S3$, calculated as six pairwise links associated with patent application $P1$ plus four pairwise links associated with patent application $P2$. The edge between the nodes $S1$ and $S2$ is based on the pairwise links between these two subclasses listed on patent application $P2$: there are two subgroups ($S1B, S1$C) listed under subclass $S1$ and one subgroup ($S2A$) listed under subclass $S2$, hence, there are two pairwise links between these two subclasses. In Fig. 2, therefore, the weight of the edge between nodes $S1$ and $S2$ is two. Similarly, the weight of the edge between nodes $S2$ and $S3$ is based on the pairwise links between these two subclasses on patent application $P2$, which are two due to one subgroup $S2A$) listed under subclass $S2$ and two subgroups ($S3B, S3$C) listed under subclass $S3$. In Fig. 2, therefore, the weight of the edge between nodes $S2$ and $S3$ is also two. Finally, we note that data on patent $P3$ do not contribute to the computation of the weight of any of the edges because on that application subclass $S3$ is the only one listed and therefore there are 0-pairwise links between subgroups under distinct subclasses.

3.2 Stage 2: Clusters identification

We use Piccardi's lumped Markov chains network cluster identification method (Piccardi 2011) to partition the network, Γ^t_{t-s}, derived from patent applications filed in the window from $t-s$ to t, with $s = 0, \ldots, t-1$ and $t = 1, \ldots, T$, into clusters of technological components such that the technological components (subclasses) within the same cluster are more likely to be listed on the same patent application than to be listed on an application with alongside a subclass from any other cluster. With sufficient network density to form a network partition, the algorithm assigns to each cluster a *persistence probability,* $\alpha \in \lceil 0,1 \rceil$, which is related to the weight of the edges among the nodes within the cluster. Within the cluster identification method, there is the choice of exogenously setting the number of clusters into which the network must be partitioned or setting limits

of the persistence probability of each cluster. If we set a threshold value of α, this is likely to result in a different number of clusters in the networks of subsequent cohorts. Comparing coarser to finer partitions may therefore 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, denoted as n. We recognize that by imposing the same number of clusters on all temporal network partitions, we introduce variations in cluster size and associated persistence probability; we thus include these statistics in the definition of our recombination index. We denote the partition of the network Γ^t_{t-s}, into clusters as $N(\Gamma^t_{t-s}) \equiv \left\{N_{t,s+1}(0), N_{t,s+1}(1), \ldots, N_{t,s+1}(n)\right\}$ where $\left\{N_{t,s+1}(0), N_{t,s+1}(1), \ldots, N_{t,s+1}(n)\right\}$ is the set of $n+1$ clusters with clusters $(1) - (n)$ identified through the Piccardi's lumped Markov chains network cluster identification method with a fixed number of clusters and cluster $N_{t,s+1}(0)$, defined as the set of all nodes that have 0-weight edges, i.e., the unconnected subclasses. In the notation of each cluster, $N_{t,s+1}(i)$, i is the cluster identifier while t and $s+1$ indicate the window of data used in the construction of the network: t is the last year of data and $s+1$ is the number of consecutive years of data, i.e. the network is constructed using patent applications data from year $t-s$ to t.

Figure 3 provides an example of the networks constructed using data from 1999 and 2000 in the PHARM sector.

We can draw several insights from Fig. 3. First, the largest cluster does not 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 number of edges 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. Subclasses 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 type of recombination across cohorts of patent applications. This example also shows that the distribution of cluster size varies. The smallest cluster in 1999 is much smaller than the largest, while in 2000 the cluster sizes are more evenly distributed. This observation justifies our choice to include cluster size into the definition of KORCI that is presented in Stage 3.

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

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. 4.

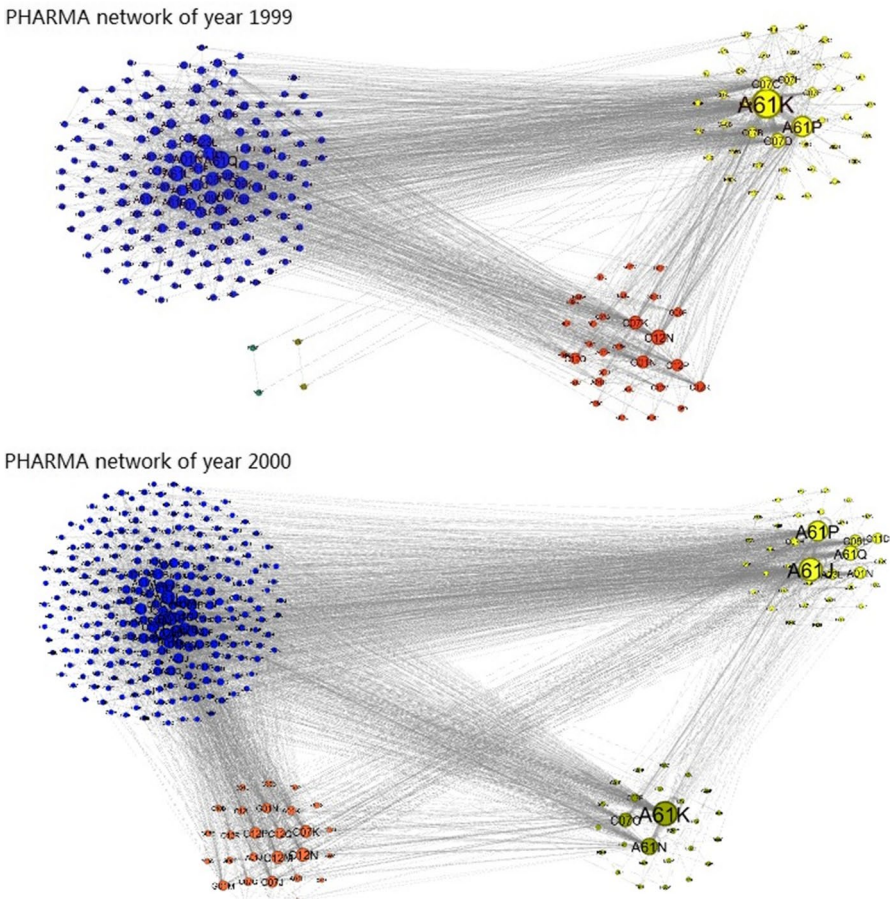

Fig. 3 Network partition visualization using data of the PHARM patents filed in the years 1999 and 2000. The network of each year is constructed using the method introduced in Stage 1, and then divided into four clusters using the network partitioning method described in Stage 2. *Different colors* are used to indicate the cluster size in each network: *blue* for the largest cluster, *yellow* for the second largest, *orange* for the third largest, and *olive green* for the smallest. Node size is in proportion to node degree, i.e. 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

Figure 4 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

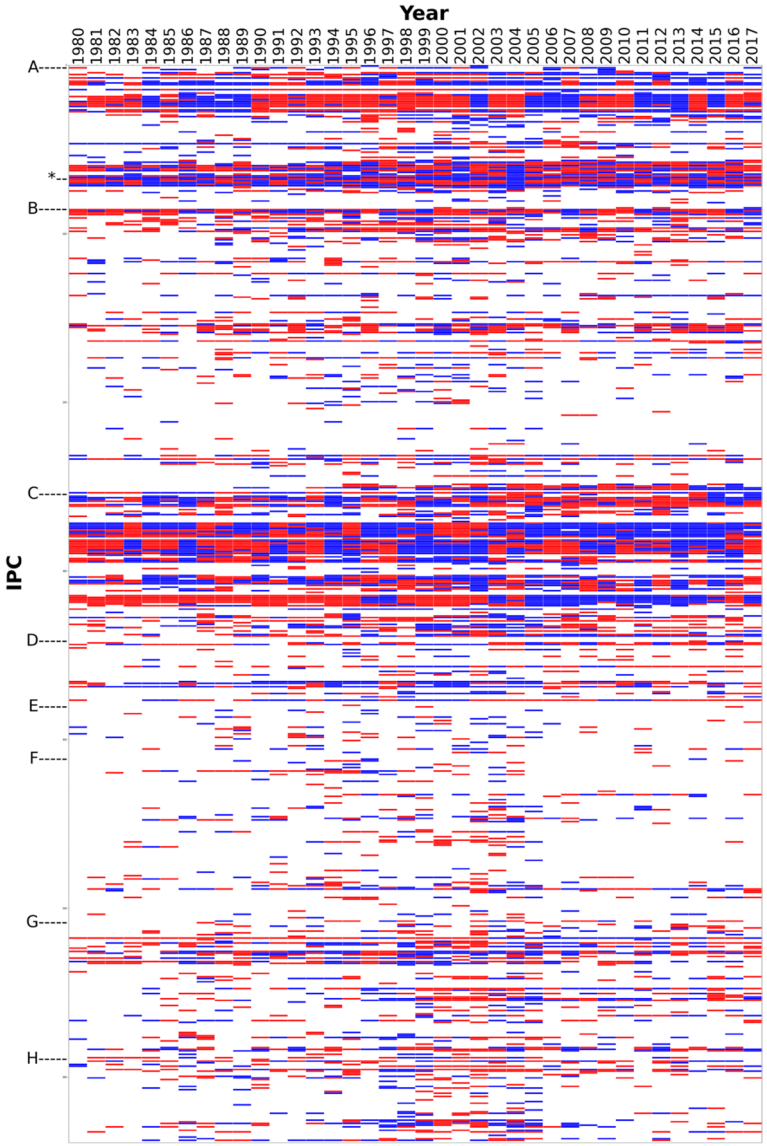

Fig. 4 Visualization of the network partition changes over time using data of PHARM patent applications for the period 1980–2017. All the IPC subclasses in the IPC scheme at the time of data extraction (The classification scheme is accessible at https://www.wipo.int/classifications/ipc/en/ (last accessed, December 2022).) 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 and with an asterisk (*) we indicate the IPC subclasses that define the sector – in this case A61K for PHARM according to the sector definition (Sector definition for Pharmaceuticals can be found at 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 subclass is not attributed to any PHARM patent application made in the corresponding year

in PHARM patent applications in this period. In addition, we can identify years with a wider spread of colored spaces (late 1990s to early 2000s) indicating exploration across more diverse knowledge origins. Conversely, there are periods when colored spaces are concentrated along fewer and adjacent lines suggesting that inventions use fewer technological components.

3.3 Stage 3: Computation

For a field of technology, we quantify the technological novelty in a cohort,t, of patents by comparing the structural changes to the network constructed in Stage 1, as Γ_t^t and its partitioning into $n+1$ clusters,$N(\Gamma_t^t,)$, derived in Stage 2 using patent applications filed in period t (for any t= $2,\ldots,T$) to the network and cluster partitioning (Γ_{t-1-s}^{t-1}, $N(\Gamma_{t-1-s}^{t-1})$) with $s = 0,...,t\text{-}2$[4] and t= $1,\ldots,T$, derived from a reference period of patent applications data. Note that we measure ex-ante technological novelty of a cohort of patent applications submitted in a single year by benchmarking it to data derived from a reference window that may be constructed based on 1, 2, or more preceding years of applications whereby the length of the reference window is given by$s+1$.

Intuitively, our ex-ante technological novelty index, KORCI, measures the concentration of subclasses in the current network that are also clustered together in the network derived from data in the reference window. Clearly, if all the subclasses in a cluster of the current network also belong to one cluster in the benchmark network, 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 technological components. As noted above in the discussion of Stage 2, the index also includes the size of the clusters and the associated persistence probability that controls for the strength of connections between the nodes within and across clusters. We recall that KORCI is defined with respect to a given cohort of patent applications P_t (for $t = 2,\ldots,T$); a fixed number of clusters in the partition network, $n+1$ (with *cluster 0* containing all unconnected nodes and clusters $1,\ldots,n$ identified via the Piccardi's lumped Markov chains network cluster identification method); and a length of the reference window, $s+1$ (for $s = 0,\ldots.,t-2$). It should be noted that when the reference window is of length longer than a year (for $s > 0$), the computation uses a rolling reference window: in every t the network composition is benchmarked to a network constructed using data from the previous $s+1$ periods. Our index is formally defined below:

$$KORCI_{t,n,s+1} = \frac{1}{\left|C(P_t)\right|}\sum_{i=1}^{n}\frac{\left|N_{t,1}(i)\right|\alpha_{t,1}(i)}{\sum_{j=0}^{n}\left(\frac{\left|N_{t,1}(i)\cap N_{t-1,s+1}(j)\right|}{\left|N_{t,1}(i)\right|}\right)^2} \tag{1}$$

[4] As a counter, the value of s may differ depending on context. Note that here the range of s is *0, … t–2* as it indicates the number of subsequent periods in the reference window. By definition, the reference window refers to years prior to the current period, *t.* which dictates the maximum value of *s*.

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 equal to zero. Although it is theoretically possible for the denominator to be equal to zero, 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 actual probability that sector-specific patent applications in two consecutive periods do not contain any common IPC codes is nil.

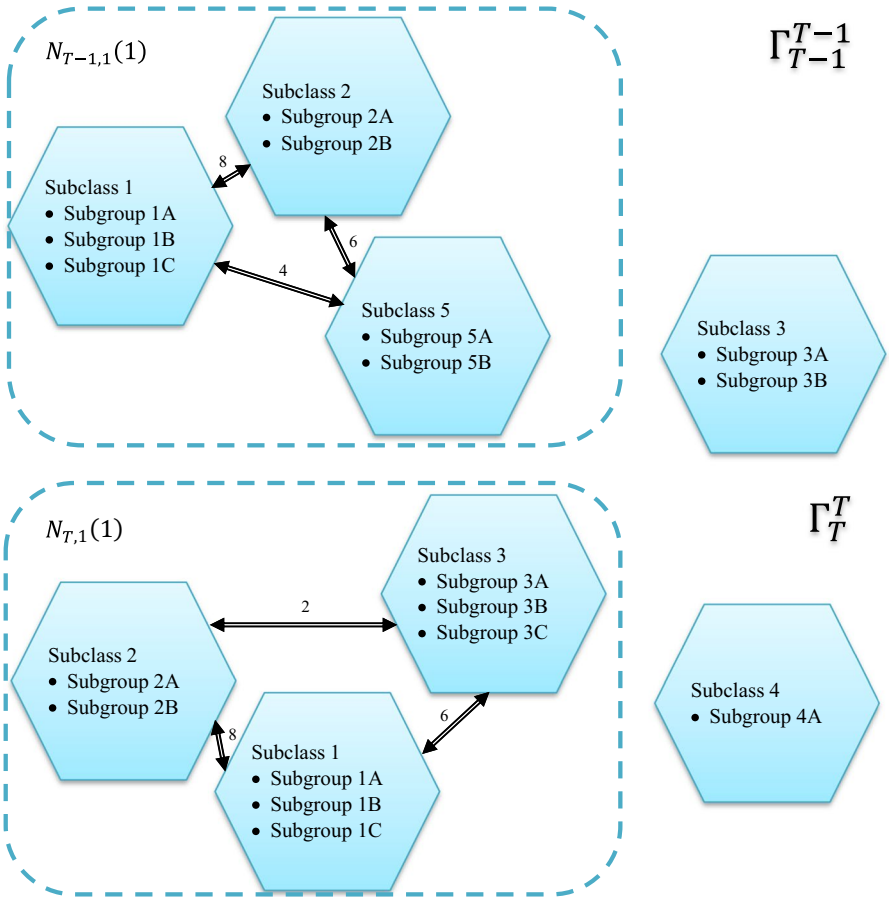

Fig. 5 Example of measuring cluster recombination of a network constructed from a cohort of patents in period T to the network constructed from patents of the previous period, T–1, as described in Eq. (1) in Stage 3 of the Methodology section. $\boldsymbol{N}(\boldsymbol{\Gamma}_{\boldsymbol{T}}^{\boldsymbol{T}})$ represents the network partition of the 1-year time window, T, including four nodes out of which one cluster $\boldsymbol{N}_{\boldsymbol{T},1}(1)$ is identified through Stage 2. $\boldsymbol{N}(\boldsymbol{\Gamma}_{\boldsymbol{T-1}}^{\boldsymbol{T-1}})$ represents the reference network partition of the previous time window, T–1, consisting of one cluster $\boldsymbol{N}_{\boldsymbol{T-1},1}(1)$ containing four nodes. Each network partition includes an unconnected singleton node that is attributed to a cluster 0

With the help of Fig. 5, we go through an example of the computation of $\text{KORCI}_{\text{T,1,1,}}$ i.e., KORCI that measures the ex-ante technological novelty of patent applications filed in period T when benchmarked to those filed in period T–1 and when the patent application network is partitioned into one cluster of connected nodes and one cluster containing non-connected nodes. Firstly, note that there are four subclasses listed in the network of period T, therefore $\left|\boldsymbol{C}(\boldsymbol{P_T})\right| = 4$. Next, we note that there is only one cluster that contains connected nodes in the network $\boldsymbol{N_{T,1}}(1) = \{$ Subclass 1, Subclass 2, Subclass 3$\}$, hence, $\left|\boldsymbol{N_{T,1}}(1)\right| = 3$. To calculate the denominator of KORCI in this example, we observe that Subclass 5, which is represented by a connected node in the cluster $\boldsymbol{N_{T-1,1}}(1)$ of network $\boldsymbol{\Gamma}_{\boldsymbol{T-1}}^{\boldsymbol{T-1}}$, is not among $\boldsymbol{C}(\boldsymbol{P_T})$ in network$\boldsymbol{\Gamma}_{\boldsymbol{T}}^{\boldsymbol{T}}$. Moreover, since Subclass 4 is not co-listed on a patent application with any other subclass based on data from period T, it is allocated to *cluster 0* $\boldsymbol{N_{T,1}}(0)$, and, by definition, it is not included in the computation of $\text{KORCI}_{\text{T,1,1}}$. Likewise, Subclass 3 in $\boldsymbol{\Gamma}_{\boldsymbol{T-1}}^{\boldsymbol{T-1}}$ is allocated to *cluster 0* in the reference period T-1, $\boldsymbol{N_{T-1,1}}(0)$. Comparing network $\boldsymbol{\Gamma}_{\boldsymbol{T}}^{\boldsymbol{T}}$ to the network using data from the reference period $\boldsymbol{\Gamma}_{\boldsymbol{T-1}}^{\boldsymbol{T-1}}$, we see that Subclasses 1 and 2 remain closely connected and that their respective links with Subclass 3 have become strong enough to result in a new network partition where Subclass 3 has replaced Subclass 5 to join Subclass 1 and 2, forming cluster $\boldsymbol{N_{T,1}}(1)$. This observation is summarized as: $\boldsymbol{N_{T,1}}(1) \cap \boldsymbol{N_{T-1,1}}(1) = \{\boldsymbol{Subclass}1, \boldsymbol{Subclass}2\}$ and$\boldsymbol{N_{T,1}}(1) \cap \boldsymbol{N_{T-1,1}}(0) = \{\boldsymbol{Subclass}3\}$. Plugging in the cardinality of these two subsets into the denominator of KORCI, we get:

$KORCI_{T,1,1} = \frac{1}{4}\left(\frac{3\alpha_{T,1}(1)}{\left(\frac{1}{3}\right)^2+\left(\frac{2}{3}\right)^2}\right)$ where $\alpha_{T,1}(1)$ is the persistence probability of $N_{T,1}(1)$ given in Stage 2.

Next, we highlight some important features of KORCI. Firstly, the index is increasing in the intensity of re-combination. This is because the denominator, $\sum_{j=0}^{n}\left(\frac{\left|N_{t,1}(i)\cap N_{t-1,s+1}(j)\right|}{\left|N_{t,1}(i)\right|}\right)^2$, 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 partition derived from the reference window of patent applications filed between periods $t-1-s$ and $t-1$; then the value of the denominator for cluster i equals 1 as $N_{t,1}(i) \cap N_{t-1,s+1}(j) = N_{t,1}(i)$. Clearly, the presence of any subclasses in i that are not present in cluster j, which constitute a novel combination of technological components, would result in a value of the denominator smaller than 1 as $N_{t,1}(i) \cap N_{t-1,s+1}(j) \subseteq N_{t,1}(i)$.

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 in the reference window. This is evident in that the denominator $\sum_{j=0}^{n}\left(\frac{\left|N_{t,1}(i)\cap N_{t-1,s+1}(j)\right|}{\left|N_{t,1}(i)\right|}\right)^2$ 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.

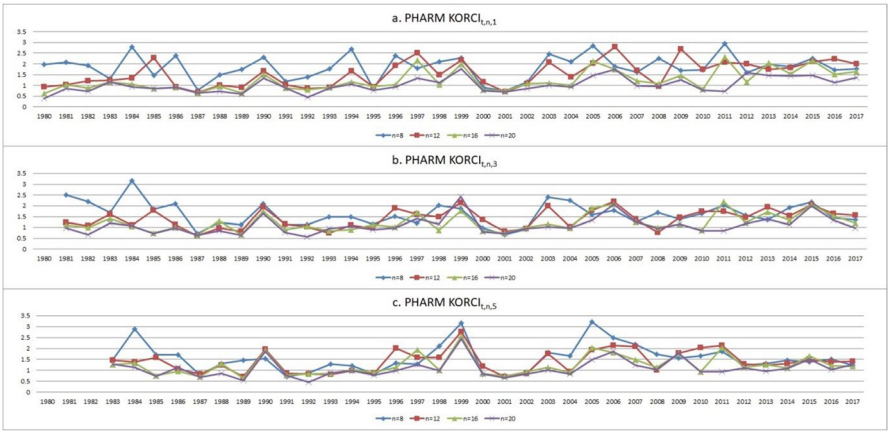

Fig. 6 KORCI calculated with PHARM data 1980–2017 using four different cluster levels $n=8$ (*marked with blue diamond*); $n=12$ (*red square*); $n=16$ (*green triangle*); and $n=20$ (*purple cross*). Panel a. presents $KORCI_{t,n,1}$ 1980–2017; panel b. presents $KORCI_{t,n,3}$ 1981–2017; panel c. presents $KORCI_{t,n,5}$ for 1983–2017 (For the time period, please refer to Subsection Data in the next section. With 1980 being the first year when above 500 patents were filed in the PHARM sector, KORCI with 1-year reference window starts from 1980. For $KORCI_{t,n,3}$, however, the 3-year reference window means a sufficient number of patent filings in 1978–1980 to allow 1981 to be the starting year, and $KORCI_{t,n,5}$ is available from 1983 with 1978–1982 as the first reference time period.)

Table 1 PHARM KORCI average statistics by number of clusters and reference window length

	Period	*n*	Mean	Std. Dev	Min	Max
$KORCI_{t,n,1}$	1980–2017	8	1.848	0.564	0.762	2.932
		12	1.549	0.587	0.673	2.788
		16	1.229	0.483	0.610	2.332
		20	1.025	0.341	0.404	1.768
$KORCI_{t,n,3}$	1981–2017	8	1.603	0.518	0.664	3.178
		12	1.410	0.445	0.634	2.213
		16	1.225	0.420	0.657	2.209
		20	1.100	0.410	0.559	2.432
$KORCI_{t,n,5}$	1983–2017	8	1.558	0.643	0.703	3.214
		12	1.402	0.518	0.701	2.763
		16	1.233	0.470	0.631	2.570
		20	1.088	0.416	0.466	2.449

Next, KORCI is increasing with 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 recombined 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 reference window, $s + 1$, in the computation of KORCI should depend on the scope of one's study. A longer length (e.g., 5 years) is recommended for the analysis of long-term trends to average out short-lived fluctuations. Conversely, where the focus is on temporal variations, a 1-year reference 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 lowest persistence probability attributed to a cluster within the partitioning is still high enough to strong connections between the nodes in this cluster. Using Fig. 6, we illustrate empirically the behavior of KORCI under different reference window lengths (1, 3, and 5 years) and a different number of clusters ($n = 8$, 12, 16, and 20) using PHARM patent data for the period 1980–2017. Accordingly, the average KORCI value of each specification is provided in Table 1.

Across all three panels, we observe that a smaller number of clusters, n, tends to result in a higher index of knowledge recombination as the blue-diamond curves lie above the others at most data points. Indeed, as shown in Table 1, 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 $\text{KORCI}_{t,n,1}$; 1.602 vs. 1.010 for $\text{KORCI}_{t,n,3}$; and 1.558 vs. 1.088 for $\text{KORCI}_{t,n,5}$). 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 of the 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 $\text{KORCI}_{t,n,1}$ and the lowest for $\text{KORCI}_{t,n,5}$ irrespective of the number of clusters because a shorter reference window has a smaller knowledge origin base and fewer nodes. On the other hand, a longer reference window is more likely to result in a different clustering of the technological components base to the current one and thus we may observe a higher average value for $\text{KORCI}_{t,n,5}$. In fact, we do not detect a regular pattern in our data; we observe the following orderings of the average values over the sample period $\text{KORCI}_{t,8,1} > \text{KORCI}_{t,8,3} > \text{KORCI}_{t,8,5}$; $\text{KORCI}_{t,16,4} > \text{KORCI}_{t,16,1} > \text{KORCI}_{t,16,3}$; and average $\text{KORCI}_{t,20,3} > \text{KORCI}_{t,20,5} > \text{KORCI}_{t,20,1}$.)

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 time-trend exhibited by the index is quite robust to changes in these parameter values. Across these different specifications, KORCI rises to a high in 1999, followed by a downward trend before reaching a low around 2000–2002, and then starts to rise again around 2003.

4 Empirical case study of cross-sector comparison

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

Table 2 Sector-specific descriptive statistics

	Period	Variable	Mean	Std. Dev	Min	Max
PHARM	1980–2017	q_t	7334.526	3767.826	1084.000	12832.000
		C_t	172.289	36.645	102.000	240.000
		g_t	0.064	0.121	– 0.310	0.322
		$KORCI_{t,8,1}$	1.849	0.564	0.762	2.932
COMP	1981–2017	q_t	7544.595	4631.046	1014.000	14678.000
		C_t	265.649	67.428	137.000	362.000
		g_t	0.079	0.113	– 0.293	0.359
		$KORCI_{t,8,1}$	2.002	0.620	0.873	3.653
AI	1982–2017	q_t	1966.778	1893.150	101.000	7581.000
		C_t	150.056	65.397	47.000	300.000
		g_t	0.137	0.154	– 0.108	0.655
		$KORCI_{t,8,1}$	1.711	0.493	0.966	2.789

4.1 Data

Patent application data from the AI, COMP, and PHARM are sourced from the REGPAT database[5] that contains information on patents filed from 1978 to 2019. Patents are identified as pertaining to a sector using the WIPO sector definition.[6] From the raw data, we construct the following data series: patent applications volume (q_t) and the total number of unique IPC subclasses listed in patent applications in each year (C_t) in each sector. We also compute the annual growth rate in application quantity (g_t) for each sector.

We employ the information on IPC subclasses and IPC subgroups in the three-stage methodology outlined above to compute sector-specific KORCI. As our methodology requires a large number of observations with a sufficient density of connections in each cohort of applications to construct a network and run a clustering algorithm, we choose the initial period based on the applications volume being consistently above a certain threshold: 500 for COMP and PHARM and 100 for AI. This allows us to calculate $KORCI_{t,8,1}$ starting from 1980 for PHARM; 1981 for COMP, and 1982 for AI patents, respectively. We choose to fix the number of clusters to $n = 8$ and the length of the reference window to 1 ($s = 0$) for presentation purposes. We have done the descriptive statistics with other values of the parameters n and s and we find the results to be robust. These estimations are available from the authors upon request. We choose 2017 as the end period due to the drop in the number of

[5] 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.

[6] 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 December 2022).

Table 3 Pairwise correlation coefficients within and between sectors as indicated in the subscript of each variable label. Pairwise correlations between sectors are shared in *color*

Period: 1982–2017	$q_{t,PHARM}$	$q_{t,COMP}$	$q_{t,AI}$	$C_{t,PHARM}$	$C_{t,COMP}$	$C_{t,COMP}$	$g_{t,PHARM}$	$g_{t.COMP}$	$g_{t.AI}$	$KORCI_{t,n,1,PHARM}$	$KORCI_{1,1,COMP}$	$KORCI_{1,1,AI}$
$q_{t,PHARM}$	1											
$q_{t,COMP}$	0.888	1										
$q_{t,AI}$	0.511	0.798	1									
$C_{t,PHARM}$	0.698	0.466	0.126	1								
$C_{t,COMP}$	0.902	0.957	0.763	0.628	1							
$C_{t,AI}$	0.726	0.908	0.941	0.392	0.908	1						
$g_{t,PHARM}$	– 0.423	– 0.548	– 0.634	– 0.042	– 0.547	– 0.611	1					
$g_{t.COMP}$	– 0.347	– 0.433	– 0.611	0.040	– 0.413	– 0.581	0.745	1				
$g_{t.AI}$	– 0.317	– 0.230	– 0.062	– 0.219	– 0.264	– 0.146	0.374	0.290	1			
$KORCI_{7t,8,1,PHARM}$	0.083	0.131	0.089	– 0.002	0.064	0.080	–0.113	–0.067	– 0.261	1		
$KORCI_{t,8,1,COMP}$	– 0.040	– 0.025	– 0.008	– 0.025	– 0.013	0.025	0.114	0.133	0.001	–0.008	1	
$KORCI_{1t,8.1,AI}$	–0.119	0.077	0.354	–0.050	0.056	0.251	–0.229	–0.102	0.142	0.218	–0.174	1

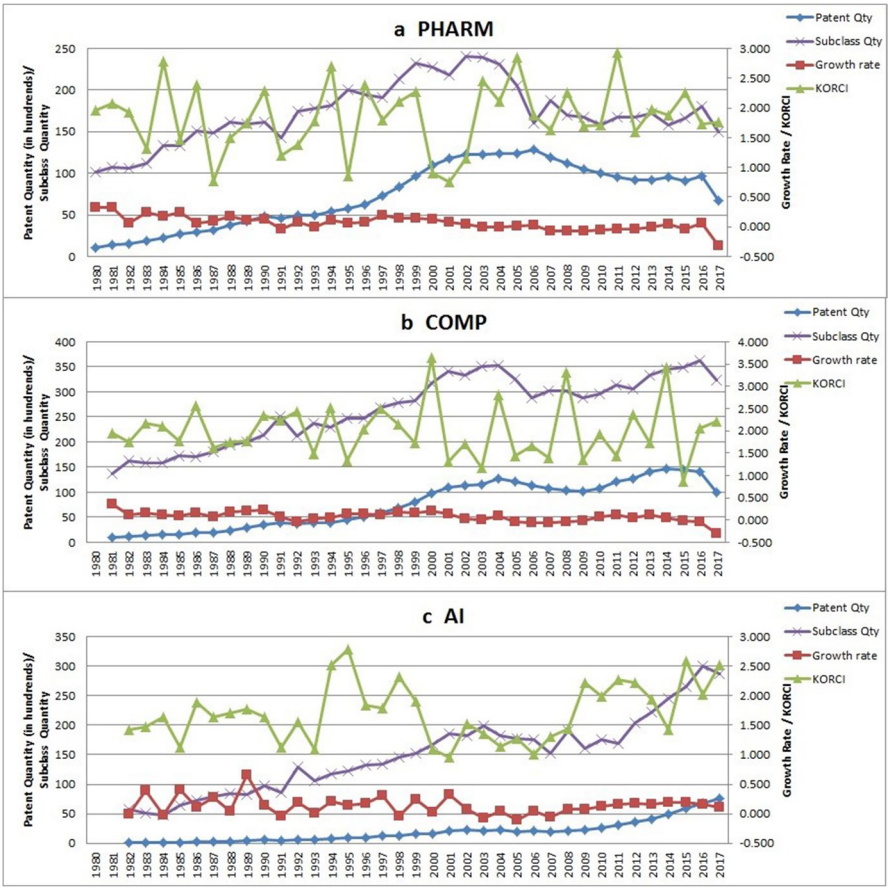

Fig. 7 Each panel presents trends for four data series: patent application volume measured in hundreds (*blue diamonds*); number of unique IPC subclasses listed on patent applications (*purple cross*); patent application annual growth rate (*red square*); and $\text{KORCI}_{t,8,1}$ (*green triangle*). 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 $\text{KORCI}_{t,8,1}$

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.

We present the sector-specific descriptive statistics of the data series in Table 2 and pairwise correlations between the variables within and between sectors in Table 3. We refer to these statistics in greater detail alongside our discussion of the time trends below.

4.2 Time-series analysis

Figure 7 presents trends in patent application volume, patent application annual growth rate, total number of unique IPC subclasses included in a cohort, and

$KORCI_{t,8,1}$ for each of the three sectors under investigation (AI, COMP, and PHARM). The *y*-axis on each panel measures the number of patent applications in hundreds and the number of unique IPC subclasses and the secondary *y*-axis to the right measures growth rates and $KORCI_{t,8,1}$. The sample period differs for the three sectors and is determined by the data availability on KORCI. Thus, Fig. 7 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 – $KORCI_{t,8,1}$ 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. The cycles are not synchronized across the sectors as we observe by comparing the three panels in Fig. 7 and note from the very low pairwise correlation coefficients for $KORCI_{t,8,1,}$ between each pair of sectors as reported in Table 3.

Nonetheless, the average values of $KORCI_{t,8,1}$ 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 (see Table 2). The higher average degree of recombination in COMP versus 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 (WIPO 2019).

While KORCI captures a qualitatively different aspect of ex-ante innovation from the other three data series, there are visible co-movements of the other three series both within sectors and across the three sectors as plotted in Fig. 7 and reported in Table 3. We firstly note the high correlation between patent application quantities and the number of unique IPC subclasses within sectors: 0.927 for AI; 0.938 for COMP; and 0.801 for PHARM (see Table 3). 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 number 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: 7545 (COMP) and 7335 (PHARM); while the average for AI is considerably lower (1967) (see Table 2). 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. 4. 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%).

Across the three sectors, we also observe a relatively higher correlation between patent volume, number of unique IPC subclasses, and growth rates as reported in Table 3. The number of unique IPC subclasses across COMP and AI have a pairwise correlation coefficient of 0.91. The high time-series correlation is suggestive of an alignment of periods of expansion of application-relevant innovations in these technological sectors. Similarly, we observe a very high pairwise correlation coefficient of 0.89 for the volume of patent applications between COMP and PHARM; and slightly lower but still high, 0.80, for the same variable, between COMP and AI. The co-movement in the volume of patenting activity across the three sectors may be underpinned by common drivers such as R&D investment, global economic conditions, patent policy, and governance.

The cyclical behavior of KORCI exhibited in Fig. 7 and the low correlation of KOCIs between sectors clearly indicates that our index captures a distinct feature of the evolution of technological innovation compared to what is measured by the other three variables. 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 the traditional measures of innovation, we proceed to conduct a regression analysis.

4.3 Regression analysis

Given that our time series are relatively short, we can only use a very parsimonious framework for the dynamic process of the evolution of the innovation frontier and this is why we choose the Autoregressive Distributed Lag (ARDL) regression model. At the foundation of our ARDL regression model is a process that relates the ex-ante technological novelty and quantity of patent applications in the current period to past quantities of patent applications and associated unique subclasses according to the following formula:

$$e^{\widetilde{\beta}_1 KORCI_{t,8,1}} q_t = \widetilde{\beta}_0 \left(q_{t-1} \left| C_{t-1} \right|^{\beta_2} \right) \tag{2}$$

Conceptually, we stipulate that ex-ante technological novelty of the frontier is encapsulated in the left-hand-side of the equation and that it is generated by the existing knowledge base captured by the right-hand-side of Eq. (2); a combination of volume applications and breadth of technological components. Note that we do not assume that there is a causal link between ex-ante novelty, KORCI, and the volume of patent applications, q_t; instead, our aim is to test for their correlation. To arrive at a regression model that we can estimate, we take the natural log transformation of both sides of Eq. (2) and obtain:

$$\widetilde{\beta}_1 KORCI_{t,8,1} + \ln q_t = \ln \widetilde{\beta}_0 + \ln q_{t-1} + \beta_2 \ln \left| C_{t-1} \right|$$

Table 4 Regression results of the correlation between patent application annual growth rates, KORCI, and unique subclass quantity

	PHARM		COMP		AI	
	(1)	(2)	(3)	(4)	(5)	(6)
	g_t	$\sum_{k=0}^{2}\frac{g_{t-k}}{3}$	g_t	$\sum_{k=0}^{2}\frac{g_{t-k}}{3}$	g_t	$\sum_{k=0}^{2}\frac{g_{t-k}}{3}$
$\ln\lvert C_{t-1}\rvert$	− 0.224***		− 0.229***		− 0.093	
	[0.001]		[0.000]		[0.141]	
$\ln\sum_{k=0}^{2}\frac{C_{t-k-1}}{3}$		− 0.188***		− 0.176***		− 0.110***
		[0.000]		[0.000]		[0.000]
$KORCI_{t,8,1}$	− 0.017		0.022		0.063	
	[0.464]		[0.362]		[0.156]	
$\sum_{k=0}^{2}\frac{KORCI_{1,1,t-k}}{3}$		− 0.072**		0.078*		0.155***
		[0.028]		[0.091]		[0.001]
constant	1.239***	1.161***	1.301***	0.900***	0.482	0.477***
	[0.000]	[0.000]	[0.000]	[0.000]	[0.121]	[0.001]
N	38	36	37	35	36	35
R-sq	0.229	0.286	0.366	0.447	0.103	0.450
F -stats	7.05***	9.49***	9.36***	16.61***	1.69	11.20***
	[0.003]	[0.000]	[0.000]	[0.000]	[0.200]	[0.000]

OLS estimation with robust standard errors p values are presented in square brackets below the coefficient estimates;

* $p<0.1$, ** $p<0.05$, *** $p<0.01$; PHARM regressions are estimated with sample 1980–1997 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)

Re-arranging by moving $\mathrm{KORCI}_{t,8,1}$ to the right-hand-side; moving $\ln q_{t-1}$ to the left-hand-side and utilizing the logarithmic approximation of the annual growth rate$g_t = \ln q_t - \ln q_{t-1}$,, we arrive at the equation:

$$g_t = \beta_0 + \beta_1 KORCI_{t,8,1,} + \beta_2 \ln\lvert C_{t-1}\rvert + \varepsilon_t, t = 2, \ldots, T \tag{3}$$

where $\beta_0 \equiv \ln\widetilde{\beta}_0$ is a constant and $\beta_1 \equiv -\widetilde{\beta}_1$ and β_2 are slope parameters and ε_t is a stochastic shock. We would expect the estimate of β_2 to be negative, otherwise, we would observe an exponential growth in the volume of patent applications over time.

Estimating model (3) empirically, we are able to test for short-term correlations between KORCI and the quantity of patents at the frontier. As shown in Fig. 7, all three series – KORCI, growth rate, and unique subclass quantity – exhibit fluctuations. In order to smooth out these variations and look for persistent correlations between KORCI and patent quantity, we also estimate the following regression model:

$$\sum_{k=0}^{2}\frac{g_{t-k}}{3} = \beta_0 + \beta_1\sum_{k=0}^{2}\frac{KORCI_{t-k,8,1}}{3} + \beta_2\sum_{k=0}^{2}\frac{\ln\lvert C_{t-k-1}\rvert}{3} + e_t t = 4, \ldots, T \tag{4}$$

Model (4) is derived as 3-year rolling average of Eq. (3). That is in (4), we use 3-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 3-year rolling average of our recombination index ($\sum_{k=0}^{2} \frac{KORCI_{,t,8,1}}{3}$) and the lagged log of the 3-year rolling average of the number of unique IPC subclasses as regressors, with the aim to identify persistent correlations within this relatively short sample period. We expect the estimate of β_2 to be negative and strongly statically significant. We are agnostic about the sign and significance of β_1. 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 4 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 3-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 negative longer-term correlation between growth rates in patent applications and KORCI in PHARM (see column 2) but positive 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 90.1% confidence level (see column 4). For each sector, the correlation signs between growth rates and KORCI are consistent across both estimations. 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 estimations we find strong evidence for the stationarity of growth rates. Interestingly, the estimates for β_2 for PHARM and COMP are very similar, which suggests that the two sectors are at the similar stage of maturity on the evolutionary path.

5 Conclusion

In this paper, we propose a new index to measure the ex-ante technological novelty in a sector that captures the degree of novel use of knowledge origins. In doing so, 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

the sector. 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 KORCI is evident in all three sectors, 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). For future work, it will be interesting to see that the same pattern exists throughout the technology space.

We acknowledge that our ex-ante technological novelty measure, KORCI, as currently defined, postulates that the technological frontier is measured using data from a single year, albeit it being benchmarked against historical data from potentially longer reference windows. The definition can easily be generalized to a measure by making longer the window at which the frontier is measured. The choice of methodology should be dictated by the objective of one's study. We are interested in visualizing year-on-year changes to the novelty of the technological frontier; researchers in the future to whom much longer time-series are available may be interested in matching the technological frontier to, for example, 5-year strategic plans at a national level.

With our time-series analysis, we also identified important sectoral differences in the co-evolution of KORCI and patent applications growth. 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.

Going forward, our work opens new venues for further research. A fruitful line of research is linking the sectoral level ex-ante technological novelty to national-level policy interventions and studying their impact across sectors such as industry subsidies or R&D incentives. One may also study structural breaks in the evolution of the technological frontier whereby with a sufficiently long time series of KORCI it may be possible to detect changes to the length or synchronicity across sectors of evolution cycles.

Finally, the sectoral-level measure that we present here can be modified to measure ex-ante novelty of a single invention. At the patent level, a well-defined ex-ante technological novelty measure that captures the sector-specific technological characters can then be used as a stepping stone to a study of what can be called *ex-post technological novelty* that refers to the potential an invention has on having a substantial impact on future inventions, the technological market, and beyond on economic structures. Having a comprehensive and coherent measure of an invention at the time of its launch enables a researcher to accurately study the causal link between strictly defined technological novelty of the invention and its future technological and market value.

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

Funding The authors declare that this work has not been supported by external funding.

Data availability Original patent data used in this manuscript is available upon request from the OECD MSTI REGPAT database.

Dataset for figures (Figs. 6 and 7), 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.

Declarations

Ethical conduct The authors declare that the research did not involve human subjects or animals.

Conflict of interest The authors declare that there are no financial and non-financial interests that are directly or indirectly related to the submitted work.

References

Beaudry C, Schiffauerova A (2011) Impacts of collaboration and network indicators on patent quality: the case of Canadian nanotechnology innovation. Eur Manag J 29(5):362–376

Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech: Theory Exp 2008(10):P10008

Broekel T (2019) Using structural diversity to measure the complexity of technologies. PLoS ONE 14(5):e0216856

Chang SB, Lai KK, Chang SM (2009) Exploring technology diffusion and classification of business methods: using the patent citation network. Technol Forecast Soc Chang 76(1):107–117

Choi J, Yoon J (2022) Measuring knowledge exploration distance at the patent level: application of network embedding and citation analysis. J Informet 16(2):101286

Colombelli A, Krafft J, Quatraro F (2014) The emergence of new technology-based sectors in European regions: A proximity-based analysis of nanotechnology. Res Policy 43(10):1681–1696

Érdi P, Makovi K, Somogyvári Z, Strandburg K, Tobochnik J, Volf P, Zalányi L (2013) Prediction of emerging technologies based on analysis of the US patent citation network. Scientometrics 95:225–242

Fleming L (2001) Recombinant uncertainty in technological search. Manage Sci 47(1):117–132

Fontana R, Nuvolari A, Verspagen B (2009) Mapping technological trajectories as patent citation networks. An application to data communication standards. Econ Innov New Technol 18(4):311–336

Gao Y (2018) Network analysis of international patent data: technology cohort, temporal dynamics and the role of trade. PhD thesis, IMT School for Advanced Studies Lucca

Gao Y, Zhu Z (2022) Regional industrial growth and biopharma patent networks: empirical insights from the UK. Appl Network Sci 7(1):77

Gao Y, Zhu Z, Kali R, Riccaboni M (2018a) Community evolution in patent networks: technological change and network dynamics. Appl Network Sci 3:1–23
Gao Y, Zhu Z, Riccaboni M (2018b) Consistency and trends of technological innovations: a network approach to the international patent classification data. In: Complex networks & their applications VI: proceedings of complex networks 2017 (ed), The sixth international conference on complex networks and their applications. Springer International Publishing, pp 744–756
Griliches Z, Pakes A, Hall BH (1986) The value of patents as indicators of inventive activity. In: Dasgupta P, Stoneman P (eds) Economic policy and technical performance. Cambridge University Press, Cambridge, pp 97–124
Guan J, Liu N (2016) Exploitative and exploratory innovations in knowledge network and collaboration network: a patent analysis in the technological field of nano-energy. Res Policy 45(1):97–112
Hall BH, Jaffe AB, Trajtenberg M (2000) Market value and patent citations: a first look. Development and comp systems 0201001, University Library of Munich, Germany
Jaffe AB, Trajtenberg M (2002) Patents, citations, and innovations: a window on the knowledge economy, Part 1. MIT press
Kay L, Newman N, Youtie J, Porter AL, Rafols I (2014) Patent overlay mapping: visualizing technological distance. J Am Soc Inf Sci 65(12):2432–2443
Lafond F, Kim D (2019) Long-run dynamics of the US patent classification system. J Evol Econ 29(2):631–664
Lerner J (1994) The importance of patent scope: an empirical analysis. The RAND J Econ 25(2):319–333
Li GC, Lai R, D'Amour A, Doolin DM, Sun Y, Torvik VI, Amy ZY, Fleming L (2014) Disambiguation and co-authorship networks of the US patent inventor database (1975–2010). Res Policy 43(6):941–955
Maraut S, Dernis H, Webb C, Spiezia V, Guellec D (2008) The OECD REGPAT database: a presentation. OECD science, technology and industry working papers, no. 2008/02. OECD Publishing, Paris. https://doi.org/10.1787/241437144144
Piccardi C (2011) Finding and testing network communities by lumped Markov chains. PLoS ONE 6(11):e27028
Rodríguez-Pose A, Crescenzi R (2008) Research and development, spillovers, innovation systems, and the genesis of regional growth in Europe. Reg Stud 42(1):51–67
Sasaki H, Sakata I (2021) Identifying potential technological spin-offs using hierarchical information in international patent classification. Technovation 100:102192
Schumpeter JA (1934) Theory of economic development. Routledge
Silvestri D, Riccaboni M, Della Malva A (2018) Sailing in all winds: technological search over the business cycle. Res Policy 47(10):1933–1944
Souza CM, Meireles MR, Almeida PE (2019) Clustering algorithms performance analysis applied to patent database. Methodology 1(2/365):139
Strumsky D, Lobo J, Van der Leeuw S (2011) Measuring the relative importance of reusing, recombining and creating technologies in the process of invention. SFI Working Paper 2011-02–003:23
Squicciarini M, Dernis H, Criscuolo C (2013) Measuring patent quality: indicators of technological and economic value. OECD science, technology and industry working papers, No. 2013/03, OECD Publishing, Paris. https://doi.org/10.1787/5k4522wkw1r8-en
Trappey AJ, Trappey CV, Wu CY, Lin CW (2012) A patent quality analysis for innovative technology and product development. Adv Eng Inform 26(1):26–34
Verhoeven D, Bakker J, Veugelers R (2016) Measuring technological novelty with patent-based indicators. Res Policy 45(3):707–723
World Intellectual Property Organization (WIPO) (2019) WIPO technology trends 2019: artificial intelligence. Geneva, World Intellectual Property Organization. https://www.wipo.int/edocs/pubdocs/en/wipo_pub_1055.pdf. Accessed December 2022
World Intellectual Property Organization (WIPO) (2020) World intellectual property indicators 2020. Geneva: World Intellectual Property Organization. https://www.wipo.int/edocs/pubdocs/en/wipo_pub_941_2020.pdf. Accessed December 2022