# Ex-ante Novelty and Invention Quality: A Cross-country Sectoral Empirical Study

Yuan Gao*[1] and Emiliya Lazarova†[2]

[1,2]School of Economics, University of East Anglia, Norwich Research Park, Norwich, NTR4 7TJ

October, 2023

Keywords: Ex-ante novelty, Disruptive innovation, Patent, Network

## Abstract

The research on measuring technological innovation quality has evolved with our understanding of the origin of novelty. Patents have been widely used in such studies because they are a form of copyright-protected outcome of inventions deemed to be valuable. The quality of technological innovation can be measured in multiple dimensions. In this paper, we make a methodological contribution to the literature on ex-ante technological novelty and propose two new

*Corresponding author: y.gao4@uea.ac.uk
†e.lazarova@uea.ac.uk

indices based on a network approach: the Inverse Recombination Intensity Index (IRII) to capture the extent to which an invention is the outcome of a novel combination of pre-existing technological components; and the New Technology Ratio (NTR) to measure the share of new knowledge elements in the invention. Through an in-depth empirical study of patents filed in the Pharmaceuticals and Computer Technology sectors, we show that our proposed indices are correlated with some of the conventional patent quality indicators and go beyond that to reveal previously unnoticed features of the inventions process, of which some are sector-specific. Moreover, through our regression analysis, we demonstrate that IRII and NTR are important predictors of a patents' potential impact on future inventions, which confirms the ex-ante nature of our indices. In the regression analysis we also include sector-country-specific R&D input variables as controls to test the robustness of our results. Our analysis suggests that the distinct characteristics of each sector affect how the quality of innovation is related to the ex-ante measures of technological novelty. We argue, therefore, that future analysis of the link between ex-ante novelty and ex-post quality of innovation needs to take into consideration the recombinant content of the invention and account for sectoral characteristics.

# 1 Introduction

Innovation is a widely acknowledged driving force for economic growth and advances of the society. Researchers have been studying innovation in order to gain insights into the status of technology development and its relationships with socio-economic changes. In quantitative assessment of technological novelty

and values, patent information has been a widely used data source because patents are directly associated with inventions - the outcome of scientific and technological research and development (R&D) activities. As we entered the digital age, patent data has become particularly popular with the availability of electronic patent database through the Internet, and was embraced by a series of important works by the NBER (National Bureau of Economic Research) researchers [Griliches et al., 1986, Fleming, 2001, Jaffe and Trajtenberg, 2002, Hall et al., 2005]. Most literature uses relatively simple counts statistics, such as the number of patent publications or citations. Some took a step further to develop composite indices based on such basic counts, like the patent quality index by Squicciarini et al. [2013]. However, as Schumpter noted in his original German book in 1911, innovation goes beyond invention - it is more about the scientific or technological novelty embedded in the inventions combined with the application Schumpeter [2017]. Simple counts are insufficient to establish a sound understanding of the degree or nature of novelty. For example, the number of patent applications does not capture the quality in terms of their economic values and impact on the following technological advances. It is also ill-equipped to reveal whether the inventions are driven by similar adaptations of the existing technologies or more ground-breaking methods. Other researchers have also questioned the reliability of conventional patent measures Griliches [1989], J. Acs and Audretsch [1989], Griliches [1998], Shepherd and Shepherd [2003]. Researchers have therefore proceeded to a search of other measures to assess the level of technology development or innovative activities.

One type of innovation that has particularly attracted researchers' interest is *disruptive innovation*; often associated with other descriptions like radical,

unconventional, ground-breaking. These types of innovations have the potential to "disrupt" the existing industry and market by destructing the established products or operations while replacing with new ones, or changing the course of current technology development rather significantly. Such changes typically happen quickly in a powerful way, rather than in a gradual progress, and have a fundamental impact within and sometimes out of the original sector. Indeed, among the earliest works, Schumpeter described "Creative Destruction" as a process that "incessantly revolutionizes the economic structure from within, incessantly destroying the old ones, and incessantly creating a new one" Schumpeter [1942]. These ideas were later summarised as the "Innovation Trilogy": invention, innovation and diffusion Kaya and Joseph [2015]. Bower and Christensen [1995] influenced the direction of further research into disruptive technologies by establishing the idea that they are distinguished by a difference in performance attributes that rapidly improve to penetrate established markets rather than by technological complexity or novelty. This view of disruptive innovation served as a theoretical foundation for subsequent researchers to take a posterior perspective and identify and discuss disruptive technologies exclusively based on their commercial applications. While we agree with Bower and Christensen's view that technological novelty and complexity do not necessarily contribute to disruptiveness, we would argue that it is essential to pursue an ex-ante perspective and study disruptiveness at the invention stage. An ex-ante assessment of the degree of technological disruptiveness has the potential to provide an early indication of the potential of a patent to induce technological change. It also allows to normalize the measure of novelty at the time of invention irrespective of the geographical origin of the invention and does not

suffer from biases linked to varying by geography socioeconomic factors that impact on diffusion and adoption. Finally, the R&D phase of innovation is where government expenditure on innovation is concentrated and where national-level policies have most impact. With our work, we therefore focus attention on the technological novelty from an ex-ante perspective.

In this respect, we contribute to the line of research that aims to develop measures of the disruptive technological content of an invention by linking it to the degree to which an invention is the result of recombination of technological components using international patent data [Fleming, 2001, Arts and Veugelers, 2015, Kaplan and Vakili, 2015, Verhoeven et al., 2016, Silvestri et al., 2018]. By its very nature a patented invention presents a novel technology, the challenge authors in this branch of the literature aim to address is to quantify the degree of technological novelty on which it is built. To date, there is no recognised single best measure. Authors put forward statistical tools that vary in their versatility, computational complexity, and the empirical counterparts for foundational concepts. Verhoeven et al. [2016] offer a comprehensive overview of these measures and illustrate this point. For example, the empirical counterpart of a technological component in the development of an invention has being derived from a textual analysis of the abstract on a patent application [Kaplan and Vakili, 2015]; the technological fields of prior inventions cited on the patent's application, e.g. Dahlin and Behrens [2005]; or the author-reported attributes of the patent under the established technological classification system [Fleming, 2001, Strumsky and Lobo, 2015, Arts and Veugelers, 2015]. Arguably, the first two methods are more susceptible to biases: the use of language and choice of algorithm may be culture and norm dependent while the choice of references

may be strategic as this information is used by patenting agency to evaluate the originality of the patent application. We, therefore, adopt the approach of using the author-reported technological fields in the description of the patent on its application. The additional advantages of using this information are that it is based on a globally adopted classification system; it is directly linked to the technological content of the invention; and, in accordance with gaining full protection, applicants have the incentive to provide a comprehensive description of their invention.

Even when authors use the same empirical counterpart for a technological input, they may employ it differently in the development of their index of ex-ante novelty with methods varying from simple counts of each component in isolation [Rosenkorf and Nerkar, 2001], distance between technological fields in a classification system of pairs of components [Trajtenberg et al., 1997], or tracking the changes in combinations of components usage [Fleming, 2001]. In our view, focusing on a single or a pair of components at a time, carries the potential to underestimates the complex way in which an invention combines multiple technological components of possible a range of technological fields. In the example of the "Oncomouse" patent application used by Verhoeven et al. [2016], there are eight technological components. If a researcher restricts attention to only pairwise combinations, then they will be looking at 56 possible pairwise combinations of technological components rather than $2^8 - 1 = 255$ possible subsets of any size of the same eight components. On the other hand, judging the ex-ante novelty based on the uniqueness of the entire list, may over-state the recombinant content of a patent if overlapping subsets of the complete list of technological components have been frequently used in previous cohorts of

patents.

Moreover, authors also differ in the reference group against which novel use of technological components are identified with many of them referencing the whole historic records [Fleming, 2001, Arts and Veugelers, 2015, Verhoeven et al., 2016] and a handful using a shorter window, e.g. previous give years Silvestri et al. [2018].

What sets our measures apart is that they are grounded in time-indexed comprehensive maps of linkages between technological components that capture the complex multilateral combinations of these components that are used in the development of patents of a specific cohort. Rather than tracing if a combination of technological components has been utilised or not in the past, we develop a statistical measure of the likelihood of the combination combining information on pairwise relations between components and the formation of clusters of components based on the frequency of their use on patent applications. We evaluate the degree of technological recombination based on the whole network of connections rather than the relatedness of two nodes. We would argue that our measures offer a more comprehensive network perspective on the process of re-combination. Furthermore, we utilise a finer level of the classification of technological components and can therefore measure the degrees of recombination and use of novel components with greater accuracy.

In addition, Verhoeven et al. [2016] measure ex-ante novelty of a patent by the use of new components that previously have not been used in a technological field. In the same vein, we put forward two distinct measures of the ex-ante novelty of a patent: the Inverse Recombination Intensity Index (IRII) and the

New Technology Ratio (NTR).

We illustrate the value of using our two ex-ante technological novelty measures to look into the process of disruptive innovation by using them to study inventions in two established technological sectors: pharmaceuticals and computer technology over a period of about 37 years. We choose these sectors because, for one, they represent a significant contribution to the volume of patenting activities by investors from a wide range of countries around the globe WIPO [2022]. Our choice of sectors is also because of their distinct processes of innovation. Saha and Bhattacharya [2011], alongside others, point to the extraordinary large share of sales of pharmaceuticals, that the R&D expenditure constitutes. These authors add that the competitive edge in the pharmaceutical sector is predicated on the advancement of scientific knowledge rather than technological know-how. Our ex-ante technological novelty measures, indeed, capture systematic differences between the two sectors that are consistent with these insights from the literature. We observe high rates of recombination of technological components among patents attributed to the pharmaceutical sector while patents attributed to the computer technologies contain a higher ratio of technological components that are new among patents in this sector.

We explore these differences between the two sectors through studying the correlations between the ex-ante and ex-post indicators of technological novelty among the patents. A strong link between the ex-ante and ex-post measures is important to make a successful case in favour of the use of the ex-ante measures as an early indicator of the performance potential of a patent. We benchmark our IRII and NTR against several measures of patent quality and technologi-

cal value described in Squicciarini et al. [Squicciarini et al., 2013] and widely used in studies on patenting and innovation. When correlated to their ex-ante indicators - patent scope, family size, backward citations, and originality - our proposed measures exhibit low to moderate degrees of linear relation. More significant is the observation that our two measures capture differences in the relation between the ex-ante and ex-post measures of patent novelty between the pharmaceuticals and computer technology sector that remain undetected when using only the established measures of ex-ante indicators.

During this benchmark endeavor, the ex-post indicators are considered measures of the outcome of invention. Meanwhile, there is an equally understandable interest on the input end. [Evenson, 1993, Kim and Marschke, 2004, Singh, 2008, Arts and Veugelers, 2015, Briggs, 2015, Fink et al., 2016], argued that the intensity and performance of innovation dependent on the regional investment and resources on R&D, such as the availability of funds, specialist skill resources and the shares of public and private input, as these are closely related to the economic status and policies of the region where the R&D activities occur. Therefore, we proceed to a country-sector analysis to look into the relationship between IRII and NTR and the output of innovation, considering national and sectoral R&D inputs.

In the following sections, we will first present an overview of the existing research on ex-ante and ex-post measures of disruptive innovation using patent data. We next describe the methodology we use to construct the network, identify technology cohorts clusters, and the development of our proposed indices: IRII and NTR. We will then present the descriptive statistics of these indices

through empirical analysis, by using patent data from the pharmaceuticals and the computer technology sectors. Lastly, through correlation study and regression estimation models, we demonstrate results on the relationship between our proposed metrics with the conventional patent indicators at patent level with and without control for country-sector-specific investment and resources in R&D.

## 2 Literature Review

In the recent literature, researchers use a variety of metrics to capture disruptive innovation through proposed empirical counterparts to novelty, unconventionality, and commercialization potential. There is not yet a consensus on a "best" measure as the proposed indicators capture different stages of the innovation cycle, depend on data availability, or a specific to sector or a selected group of inventions.

Among the indicators that focus on the first stage of innovation, that involves production of new knowledge and inventions, the most widely used is the *originality* index which was developed by Trajtenberg et al. [Trajtenberg et al., 1997] as a measure of the extent of the diversity of the knowledge sources that form the foundation of a patent by being cited by it. The concept behind is that absorbing knowledge from a wide range of technological fields is presumably a contributing factor to original innovation. The index has been used by researchers in studying the patenting agencies' decisions and economic performance of invention enterprises[Gompers et al., 2005, Harhoff and Wagner, 2009,

Stahl, 2010]. It is worth contrasting our IRII and the *originality* index as both measures use the degree of concentration of "same group" technological components inferred from a patent application. The main difference is that while the *originality* index refers to a 'group' as the classification codes belonging to the same subclass among the set of patents cited by the reference one, in the IRII, the 'group' is defined by all subclasses that belong to the same cluster based on the frequency of their co-assignment to a patent in the reference cohort of patent applications. Thus, our index is designed to measure the degree of *novelty in the combination* of technological components compared to the mere breadth of usage, which is why our approach is better tailored to capture the destructive nature of innovation. We measure the novelty vis-à-vis the whole cohort of patent applications, while the breadth is derived from the backward citations on the specific patent application.

Other measures, similar to *originality*, used in the literature include *patent scope*, *backward citations*, and *family size*. Compared to *originality*, these are simple counts derived directly from information listed on patent filing documents. *Patent scope* is defined as the number of distinct subclasses that identify the technologies included in a patent. The intuition behind this measure is that a greater number of subclasses is indicative of a wider range of technological components being used and therefore more complex invention or far-reaching impact. The indicator has been used in the literature as a measure for potential of a patent to generate higher technological value as in fundamental invention or economic value through a market return, for example, [Lerner, 1994, Régibeau and Rockett, 2010].

*Backward citations* is the number of prior art (such as other patents or scientific work) cited by a patent. Backward citations are used by the patenting agency to assess the technological novelty of an invention. It should be noted, however, that agencies differ in disclosure rules and therefore comparison across of this variable across patent applications filed in different countries maybe problematic. In the literature [Criscuolo and Verspagen, 2008] backward citations are used to study knowledge transfer and the dynamics of invention within a firm or sector. While some authors find evidence that a large number of backward citations is negatively related to the degree of technological novelty of a patent [Criscuolo and Verspagen, 2008, Lanjouw and Schankerman, 2001, 2004], others find it being positively correlated to the invention's market value [Harhoff et al., 2003].

Lastly, *family size* is the number of patent authorities located in different jurisdictions that a same invention has been filed to for intellectural property protection. This indicator is linked to the rights of patent applicants to seek wider geographical protection for their invention via related filings to multiple patenting offices within 12 months of the first priority filing. In the literature, a larger patent family size is found to be positively correlated with the invention's potential to generate economic value via wider geographical market capture. [Lanjouw et al., 1998, Harhoff et al., 2003].

Two sets of authors, as we discuss above, Verhoeven et al. [2016] and Silvestri et al. [2018] develop ex-ante technological novelty indicators that are specifically designed to capture the disruptive nature of innovation. We re-call the earlier discussion from the Introduction about the distinct features of our

methodology that aims at higher accuracy and more comprehensive approach to measuring re-combination and novel use of technological components. Verhoeven et al. [2016] demonstrate the validity of their measures by finding a strong positive correlation with the likelihood that a patent is in a group of award-winning patents, on the one hand, and a strong negative correlation with the likelihood that a patent is refused by the European Patent Office, on the other. Silvestri et al. [2018], instead, offer a time-series analysis of the correlation between business-cycle fluctuations and fluctuations in the degree of unconventionality and do not conduct any analysis of the link to between their ex-ante measure of technological novelty and any ex-post indicator.

There are several established ex-post indicators introduced in Squicciarini et al. [2013]: *forward citations*, *generality*, and *breakthrough*. *Forward citations*, similar to *backward citations*, is the number of citations made by subsequent patent applications that a patent receives within five to seven years after its publication date. Intuitively, it is thought to reflect the foundational value of a patent in developing new technologies. Several authors have indeed found a positive correlation between the number forward citations and the economic value of a patent [Trajtenberg, 1990, Hall et al., 2005, Harhoff et al., 2003]. [Lanjouw and Schankerman, 2004].

*Generality* is the ex-post counter part of the ex-ante indicator, *originality*, by using *forward citations* to capture the scope and degree of general-purpose technology that a patent enables. In the literature, this index has been utilised to understand the commercialization potential of inventions and how innovation meets the market [Henderson et al., 1998, Layne-Farrar and Lerner, 2011],

Galasso, 2011].

*Breakthrough* is also derived from the number of forward citations a published patent received: it is an indicator variable which equals 1 for patents in the the top 1% by the number of forward citations among those filed in the same year; and 0 otherwise. It was first put forward by Ahuja and Morris Lampert [2001] to identify inventions that have a significant impact on future technological development. In their seminal work Ahuja and Lampert found that familiarity, maturity and propinquity are three "traps" that could hinder the creation of a breakthrough invention in firm organizations [Ahuja and Morris Lampert, 2001]. More recently, Srivastava and Gnyawali [2011] found that the quality and diversity of a firm's portfolio of technological resources have a positive impact on the probability of a breakthrough innovation. Kerr [2010], Popp et al. [2013] provided evidence that the occurrence of breakthrough innovations could stimulate subsequently regional and sectoral innovation activities.

Other authors, Arts and Veugelers [2015] and Briggs [2015] modify the standard definition of a breakthrough innovation and introduces an endogenous threshold of citations that depends on the observed distribution among the patents in each cohort to allow for a time-varying sharing of patents in each cohort to be breakthrough [Arts and Veugelers, 2015, Briggs, 2015]. Briggs [2015] further motivate this methodology by referring to sectoral differences in the volume of citations and show that co-ownership of a patent is an important factor in determining the breakthrough potential of the patent as defined in their work. Given its well-recognized indicative significance as a ex-post patent quality indicator, *breakthrough* is also used in our study to validate the power of

our *ex-ante* novelty measures, IRII and NTR, in predicting future technological novelty impact.

## 3 Measures of Ex-ante Technological Novelty

To measure the extent of technological novelty embedded in a single patent, we must first establish what is the existing state of technological knowledge in the sector to which the patent belongs. In this respect, we build on our previous work, [Gao and Lazarova, 2022], where as part of a framework for quantifying the technological evolution at sectoral level, we offer a methodology for mapping the frontier of current technological knowledge as captured by a cohort of patent applications. In our work, we represent the frontier as a network of technological components, their interconnectedness and the strength of pairwise connections. Within this complex diagrammatic representation of the state of technological knowledge, we proceed to identify patterns of combinations of technological components usage which occur with a high frequency. Equipped with this information, we are able to gauge the degree of novelty in the combination of technological components listed on a new patent application as compared to those present in a cohort of patent application from a most recent reference period. In addition, we can identify among the patent characteristics any technological components which have not been listed in an application in the reference cohort.

In summary, our methodology consists of two stages: mapping of technological knowledge use and identifying high frequency patterns of usage in a sector;

and, measuring the ex-ante technological novelty of a patent in the sector. The first stage uses information from the whole cohort of patent applications. The second stage quantifies two distinct aspects of technological novelty: the intensity of novel combinations in the use of established technological components in a patent (IRII) and the proportion of technological components in the patent application that are new to the sector (NTR). As we have discussed in detail the network-level analysis that constitutes stage one in our previous work[Gao and Lazarova, 2022], we only provide an overview of these parts of the methodology below. The novel part of this methodology is in the patent level measures of ex-ante technological novelty that follows from that.

### 3.1 Network Construction and Clusters Identification

We follow the network construction method developed in [Gao and Lazarova, 2022] to build a map of the frontier of technological knowledge in a sector in a given time period. This exercise uses information from the set of all patent applications filed in the cohort linked to a specific sector. In the description of an invention, a patent application contains a list of technological components on which the invention is built. These technological components are well-defined categories in technological classification systems published by patenting authorities. Here we adopt the International Patent Classification (IPC) scheme, a hierarchic system assigning technical fields as a patent attribute developed and released by the World Intellectual Property Organization (WIPO)[1]. We employ two tiers of the IPC scheme: the first tier is the 4-digit level IPC codes, known as

[1]The IPC scheme can be accessed at https://www.wipo.int/classifications/ipc/en/

*subclasses*; and the 2nd-tier is the 8 to 11 digits IPC codes labeled as *subgroups*. In the network representation of the technology encoded in the patents, we use the subclasses listed on all patent applications in a cohort as the network nodes. We define a link between two nodes to exist in the network if the two subclasses corresponding to these nodes are co-listed on at least one patent application in the cohort. The weight of the link between the two nodes is calculated by using the 2nd-tier IPC codes at the subgroup level and aggregating this information across the whole cohort. In particular, We take the strength of the technological complementarity between any two subclasses in the development of a patent to be proportional to the number of pairwise combinations of subgroups listed under each subclass. For example, consider two patents, A and B, that both list subclassess 1 and 2 as their technological attributes. In patent A, subclass 1 lists one subgroup and in patent B subclass 1 lists three subgroups. Let subclass 2 list two subgroups as attributed to both patents A and B. Then, the strength of the complementarity between subclasses 1 and 2 is calculated to be two in the development of patent A (there are only two distinct pairs of subgroups between the two subclasses) and six in patent B (there are 6 distinct pairs). If two subclasses are not co-listed on a given patent, then their technological complementarity in the development of that patent is zero. So to derive a measure of the strength of the technological link between any two subclasses present in a cohort of patents, we sum up the number of pairwise combinations of subgroups listed under these two subclassess for all patent applications in the cohort.[2]

[2]We acknowledge that the IPC scheme is imbalanced in the sense the number of subgroups listed under each subclass varies. This implies that subclasses with a smaller number of subgroups, theoretically, can form fewer pairwise links. However, empirically, we do not observe that subclasses with a larger number of subgroups list more subgroups on a patent.**Can we add some data to support this statement?** Since our method tracks only the number of subgroups and not the variety of subgroups, we think there is no empirical bias that underestimates the degree of complementarity between subclasses with a smaller number of subgroups

The resulting weighted network provides a comprehensive snapshot of the interconnectedness between the subclasses used in the development of the cohort of inventions. We will use that as a benchmark against which we aim to measure the technological novelty of an invention that occurs in the future period. Our task is to measure how close the technology use in a new patent is to those which were used in the development of all the patents filed in the reference window. The next step towards answering this question, given the complexity of the information on technology use captured by the network, is to identify groupings of technological components based on the high frequency of their co-usage in the cohort.

As in [Gao and Lazarova, 2022], we use Carlo Piccardi's lumped Markov chains network community identification method [Piccardi, 2011] to identify naturally formed clusters in the network. When the sample data size is sufficiently large, the clustering method results in a distinguishable network partition with the definition of each cluster being directly related to the strength of links between any two nodes within the cluster. As technologies evolve in every new cohort of patent applications, the composition of clusters, their size, and connectedness strength vary. From a pure probability point of view, where the reference network is partitioned into more clusters, a new patent application is more likely to utilize a combination of subclasses that spans different clusters compared to a reference network partition with fewer clusters. To configure the networks of consecutive cohorts in a temporally comparable way, we choose to fix the number of clusters in the partition of each network.[3]

in the IPC scheme.

[3]In [Gao and Lazarova, 2022] we examine different values for a fixed number clusters as part of a robustness check in the construction of the technological frontier.

The resulting output from the stage of the methodology is a partition of the set of subclasses listed on all patent applications in a given cohort based on the frequency and strength of their co-listings based on the cluster identification method employed. We capture this output in the following notation, which we will subsequently use in the formal definition of our patent-level technological novelty indices. We will denote a generic patent as $k$ and the set of all patents filed in a period $t$ as $N_t$. We will denote the set of subclasses listed on a patent application $k$ as $S_k$ and the collection of all subclasses listed on all patent applications filed in a reference window of size $s$ time periods, i.e., from year $t-s+1$ up to year $t$ as $\mathcal{C}_{s,t} = \cup_{j \in N_{t-s+1} \cup \ldots \cup_t} S_j$. We denote the resulting partition of the set $\mathcal{C}_{s,t}$ into $n$ clusters as $P_{s,t} = \{C_{s,t}(1), \ldots, C_{s,t}(n)\}$.

## 3.2 Inverse Recombination Intensity Index

Our Inverse Recombination Intensity Index (IRII) characterises a patent application by the degree to which the grouping of technological components on which it is based presents a novel way of combining these IPC subclasses compared to their mode of usage in the preceding cohort of patent applications in the same sector. The index is designed to measure the degree of a radical ex-ante technological innovation carried by an individual invention which is benchmarked against the sector-wide practice. We first introduce some notation that we will use in the formal definition of the index.

For a patent $k$, we recall that the collection of subclasses listed on the $k$'s application is debited as $S_k$. Then, the IRII of a patent $k$ filed in a period $t$

vis-à-vis the reference period $t-s,\ldots,t-1$ is formally defined as:

$$\mathrm{IRII}_k = \sum_{j=1}^{n} \left( \frac{\mid C_{s,t-1}(j) \cap S_k \mid}{\mid S_k \mid} \right)^2 \tag{1}$$

where $C_{s,t-1}(j)$ is the $j$th cluster of the partition $P_{s,t-1}$ of subclasses listed on patent applications that were filed in the period from year $t-s$ to $t-1$ in the same sector. First, we note that there is at least one subclass which is listed in common on $k$ application and the application of patents filed in the reference window since these patents below to the same technology sector. It follows that the lowest value that IRII can attain is bounded by $\frac{1}{|S_k|}$. This is obtained when all subclasses listed on $k$'s application but the sector-definition one are not elements of the set $\mathcal{C}_{s,t-1}$. The maximum value that the index can obtain, instead, is 1. This is when all the subclasses listed on $k$'s applications belong to the same cluster in the reference window, i.e. there has been no radical recombination in the use of technological components used in the development of $k$ compared to those combination used in the development of the cohort of patents in the reference window.

### 3.3 New Technology Ratio

Our New Technology Ratio (NTR) characterises a patent application by the degree to which it employs technological components that have not been listed on any patent applications in the previous cohort in the sector. Patent data analysts have used IPC classifications at different hierarchic levels for their purposes. Since our aim is to detect any new technological elements compared to the previous cohort, we define the ratio at the subgroup level of the IPC

hierarchy as this is a more refined measure with a higher degree of variability compared to a similar measure based on subclasses. Like Verhoeven et al. [2016] who used the 7-digit subgroups to identify novelty in technological knowledge origins, we take a further step from there to measure the intensity of such novelty. Given a patent $k$, we denote the set of subgroups listed on $k$'s application as $G_k$. We denote the set of all subgroups listed on a patent application in a given sector in the period $s-t+1$ to $t$ as $\Gamma_{s,t} = \cup_{j \in N_{t-s+1} \cup \ldots \cup_t} G_j$. By exclusion, the subset of subgroups listed on $k$'s application filed in period $t$ which had not been referenced on a patent application in the $k$'s sector during the reference window is given by $G_k \setminus \Gamma_{s,t-1}$.

$$\mathrm{NTR}_k = \frac{\mid \{G_k \setminus \Gamma_{s,t-1}\} \mid}{\mid G_k \mid}. \tag{2}$$

Notice that NTR is higher as a patent uses new subgroups within the hierarchy of the technological classes on which the sector is defined. It may also increase if new subgroups under the hierarchical classification of other sectors are employed in the development of the patent. The maximum value of NTR is 1; this is when none of the subgroups listed on a patent application are used by any patent in this sector in the reference window. The minimum value, conversely, is 0; this is when all the subgroups listed on a patent application are in the set of subgroups from the reference window.

# 4 Data and Empirical Statistics

## 4.1 Patent-level Data

To illustrate the use of our novel measures, we obtain data on patent applications from the REGPAT database [Maraut et al., 2008] in the COMP and PHARM sectors. Measures of patent quality are taken from the OECD's Patent Quality Indicators database [Squicciarini et al., 2013]. We use the February 2022 release by OECD for both datasets. Based on the information in the REGPAT dataset, we can identify and select all patent applications that can be attributed to each of the two sectors. As per established practice in technological field studies, [Fink et al., 2016], patents are classified into these sectors in accordance with the definition of the WIPO.[4] We note that the IPC scheme has undergone regular updates to keep up with the latest scientific and technological development. With each reform, WIPO re-classify patent files to reflect the changes made to the IPC scheme through the revision. By downloading the data in one batch, we ensure that the IPC classification information is consistent and coherent across cohorts of patent applications.

While longer time-series are available in these databases, we select the sample period from 1980 to 2018 for patents in PHARM, and from 1980 to 2018 and from 1981 to 2018 for the COMP. The samples are selected on the basis that the volume of applications is consistently above 500 in each consecutive year. We need such large cohorts of patents in order to construct a network with a

[4] 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`. (last accessed December 2022).

sufficiently large number of nodes and high enough density of connections to identify persistent clusters at the first stage of our methodology in each consecutive year. We therefore select the samples for which we can derive reliable values for IRII and NTR.

As an illustration of our two-stage methodology, we provide a series of graphs that allow us to visualise the network clustering, identification and distribution of new subgroups, and the recombination process. In all network graphs, we choose to present the network partitions of the 2005 and 2006 cohorts because a new version of the IPC scheme (the eighth edition) was released on January 1, 2006 which presented a major revision.[Makarov, 2006]. This allows us to detect a possible impact of the process of revision on our results.

We start with Figure 1, where panels (a) and (b) show the 8-cluster network partitions constructed based on PHARM patents filed in years 2005 and 2006, respectively. Figure 2 shows the partitions of the COMP patent networks in the same two consecutive years. Nodes highlighted in blue color are the subclasses containing subgroups that are not present among patents of the sectoral cohort of the previous year, i.e., the new subclasses used in the calculations of the NTR. We note that in neither Figures 1 or 2, the network partitions show a structural change between the two years. The distribution and portion of these nodes are also similar in the temporally consecutive network partitions. While Figures 1 and 2 provide snapshots, in Figure 3 we present the temporal trend by sector of the share of new subgroups in the total number of unique subgroups, and the share of patents containing such new subgroups, whereby the new subgroups are identified using a 1-year reference window. Both sectors show a decreasing

trend of the two metrics. Similar to the cluster structure in Figures 1 and 2, the major IPC scheme update in 2006 does not appear to introduce a structural break in these series.

Next, we present Figures 4 and 5 which illustrate the process of recombination with reference windows of one year and five years. Using PHARM patent data, Figure 4-(a) shows how the network clusters of subclasses used by patents filed in 2006 are recombined versus the clusters identified through patent use among those filed in 2005. Compared to Figure 4-(c) where the reference is the network partition built on patents filed in the 5-year window from 2001 to 2005, some differences are visually detectable. For example, the largest cluster in 2006 is broken down to more evenly distributed in size sub-clusters in (c) compared to (a), which indicates that the degree of recombination of that cluster is higher against the 5-year window network partition. Since our algorithm for identifying network clusters implies a likely positive correlation between the persistence probability and the cluster size, in this example, IRII51 is likely to be smaller than IRII11 thanks to the higher extent of recombination in the largest cluster in the network partition of 2006. Similarly, Figure 5 provides the resulting networks for COMP using data for the two same years. The recombinant degree in Figure 5-(a) is not so different from the one exhibited in 5-(c). This can also be reflected by the intermediate stage in (b). In Figure 4-(b) about half of the nodes in the core blue sub-cluster within the largest cluster are in red color - i.e. they below to a different cluster in the 5-year window, while in Figure 5-(b) a much smaller number of nodes in the blue sub-cluster are red, showing that the network partition of COMP patents in 2005 is not that different from that in 2001-2005. With Figures 4 and 5, we start noting important differences in

the composition of ex-ante technological novelty between PHARM and COMP. We will explore these further throughout our empirical case study.

To ensure the robustness of our analysis, we adopt a similar approach to Gao and Lazarova [2022] and calculate IRII and NTR with three different reference windows: 1 year, 3 years and 5 years, labelled as IRII11, IRII31 and IRII51; and NTR11, NTR31, and NTR51, respectively. For IRII and NTR with three-year and five-year reference periods it is feasible to calculate IRII and NTR from 1981 and 1983, respectively, for both COMP and PHARM. In addition, to render our results less sensitive to the choice of the number of clusters at the first stage of our methodology, we construct four different network partitions using 8, 12, 16 and 20 clusters, respectively. We then obtain the average IRII value of a patent over the 4 different partition configurations.[5] We perform the same average calculation for the IRII in each of the there reference window: IRII11, IRII31, and IRII51.

In Table 1 we provide a descriptive summary of IRII and NTR calculated, as described above, using different reference windows. During the sample period PHARM and COMP are comparable in the number of patent filings. There are, however, important sectoral differences. PHARM, on the one hand, has lower average NTR, which is indicative of a lower proportion of new technological components being introduced in this sector compared to COMP. COMP, on the other hand, has higher average IRII, suggesting that the pattern of usage of established technological components is more persistent and inventions in the field are likely to rely, to a less extent than in PHARM, on a novel combination

[5]The number of clusters in a partition does not impact on the definition of NTR.

of the established technologies. We have already seen an indication of this observation in the two-year snapshots presented in Figures 4 and 5, which is now demonstrated again in the summary statistics of the entire sample. Meanwhile, the minimum values of IRII in PHARM are larger than those in COMP for all indices irrespective of the reference window. This indicates that the patents with the lowest IRII in COMP have a higher level of technological recombination than those in PHARM. There are six COMP patents with IRII11 equal to zero, the minimum value. They all have NTR11 equal to one, the maximum value. Each of these patents has only one or two subclasses which are all new technological components compared to the previous year.

Figure 6 shows how the annual average IRIIs and NTRs, when calculated using different reference windows, change over time. For both sectors, IRIIs fluctuate around a constant level and NTRs exhibit a decreasing trend. We deduce, therefore, that while the recombination of technologies is a permanent feature of inventions, the introduction of new technologies diminishes as a sector matures; an observation which is consistent with Figure 3. Using different reference windows for calculating the IRII and NTR result in similar values and trends. This supports the robustness of our method. Comparing the two sectors, we note that the IRIIs of COMP not only have the highest all-time average values, but also exhibit the lowest level of fluctuations, which indicate a more stable rate of technology recombination in this sector.

We further assess IRII and NTR in comparison with several measures of patent quality and technological value described in Squicciarini et al. [Squicciarini et al., 2013]. For the sample of patents for which we have calculated IRII

and NTR, we retrieve individual patent data from the OECD Quality Indicators dataset using the unique patent identifiers.[6] We discussed these indicators in our literature review. In Table 2 we provide their definitions. Some of these patent quality indicators are designed to measure ex-ante technological novelty (patent scope, family size, backward citations, originality) similar to IRII and NTR. Others - generality, breakthrough rate, and forward citations[7] - are measures of ex-post quality.

We will further investigate the statistical power of correlation between the ex-ante and ex-post measures in the next sections. Here we tale the opportunity to illustrate that this link is not trivial using the first stage of our methodology and the data on breakthrough rate. In Figures 7 and 8, for PHARM and COMP, respectively, once again we present the network clusters in years 2005 and 2006, however, in the clusters we highlight the nodes with subclasses that belong to patents designated as *breakthrough* patents. As shown in these figures, while the larger clusters contain more subclasses that belong to *breakthrough* patents, such subclasses can be found in any size of cluster and their distribution in different clusters varies from year to year.

$$\text{Originality}_k = 1 - \sum_{j \in \cup_{i \in B(k)} S_i} \left( \frac{\mid \{j \in \cup_{i \in B(k)} S_i\} \mid}{\mid \cup_{i \in B(k)} S_i \mid} \right)^2$$

[6] The OECD patent quality indicator dataset provides two data tables: one at patent level; and the other at cohort level by year of filing and technology field. We use the patent-level data set.

[7] We note that the OECD dataset provides two metrics on *forward citations*: one counts citations within five years after patent's publication, and the other, within seven years. The publication date of a patent is usually within 18 months of the patent application filing date. Thus, patents with a more recent application date are expected to have fewer forward citations. In our study, we use the five-year post-publication forward citation numbers to utilise a longer time-series sample with a more accurate count of *forward citations*.

Table 3 provides the summary statistics of the variables listed above for PHARM and COMP. It allows a sectoral comparison of the variable values, showing that PHARM has a larger *patent scope* and *family size* on average than COMP. A smaller *patent scope* (smaller $S_k$) could be a potential contributor to a larger IRII, but we cannot conclude that this is the cause of the higher average IRII values in COMP as shown in Table 1.[8] PHARM patents also tend to cite more prior arts and receive more citations in five years after patent publication.[9] The variables show some significant variances. Despite a lower average value, COMP patents have a wider range of *patent scope* than PHARM. Indeed, the kurtosis of COMP *patent scope* is 24.0196, much higher than that of PHARM: 6.4531. Meanwhile, PHARM is wider in range than COMP in *family size*, *backward citations*, and *forward citations*. However, only with *forward citations* PHARM has higher skewness (PHARM: 37.1322, COMP: 22.8177) and kurtosis (PHARM: 2935.5600, COMP: 1067.0120).[10] In summary, the COMP sector, with lower mean values on these variables, has a more positively skewed and more leptokurtic distribution than PHARM in *family size* and *backward citations*, and the opposite is true for *patent scope* and *forward citations*. PHARM

[8]The minimum value of *patent scope* is reported as zero in both sectors. This may be taken as typo. Instead, in the original OECD data source there is one PHARM patent and four COMP patents codes with *patent scope* equal to zero. We have manually looked up these patents using the European Patent Office's patent search service, Espacenet, and found that each of them actually has one IPC subgroup/subclass. Therefore, the correct value of *patent scope* by definition should be one. After removing these patents from each sector's sample, the difference in the mean value of all the variables is at the 5th or 6th place after decimal point (See Supplement A in the Supplementary Datasheet attachment for the summary statistics excluding the zero-patent-scope observations.). We further manually computed the number of unique subclasses of each patent in the sample data and compared with the OECD dataset. Out of the 278,990 observations in PHARM, 4,454 show different values from *patent scope*, and the average difference is -0.0172. For COMP, 5,794 patents out of 282,506 have different *patent scope* values, with an average difference of -0.0232. So, we consider the impact due to this potential data inaccuracy to be minimal and continue to use the OECD dataset in the subsequent analysis.

[9]Both forward and backward citations include citations to and from patents within and outside of the sector.

[10]See Supplement B in the Supplementary Datasheet attachment for the statistics for all the variables.

patents also demonstrate higher mean values in the *originality* and *generality* indicators. Both sectors have negative skewness and positive kurtosis values in these two variables, with the absolute values larger in PHARM.

The summary statistics show that overall inventions in the two sectors carry different ex-ante and ex-post characteristics. PHARM patents tend to have a wider technological breadth and a larger set of patents filed in international patent jurisdictions that are related to the same priority filings. Both indicators have been used as the potential of the invention to generate higher commercial value for the patent owner [Lerner, 1994, Lanjouw et al., 1998, Harhoff et al., 2003]. In addition, having more backward and forward citations indicates that knowledge spillover among patents plays a bigger role in PHARM inventions than in COMP; while higher average values of *originality* and *generality* of PHARM patents point to the likelihood of inventions being more original [Trajtenberg et al., 1997] and more general-purpose [Hall and Trajtenberg, 2004], but less fundamentally novel [Lanjouw and Schankerman, 2001]. Among all the variables based on simple counts, except for *forward citations*, COMP summary statistics have distributions with higher peak and thinner tails.

Finally, we note that a reduced number of patents have data on *originality* and *generality*. In particular, data availability of *generality* varies significantly over time and across sector. We present the number of observations over the sample period split by sector in Figure 9. The figure clearly illustrates that the *generality* time series drop to rather low levels by 2018 for both PHARM and COMP. The lower number of observations is most likely linked to the increasingly shorter window over which forward citations can be observed. To check

that the limited availability of these two variables does not introduce a selection bias in our analysis, we will present computations both including and excluding these variables in the next subsection.

## 4.2 Patent-level Correlations

We begin our analysis with a discussion of the correlation matrices of the continuous variables that we introduced in the previous section for each sector under investigation: COMP and PHARM. In Tables 4 and 5 we present the pairwise correlation coefficients in two parts: sub-tables (a) in each table presents results based on the sample where data for all the variables except *originality* and *generality* are available, and sub-tables (b) include the pairwise correlations for the full list of variables. The sample size for the computations of sub-tables (a) is much larger than the ones that is used for sub-tables (b) due to the limited data availability for *generality* as shown in Figure 9 and lesser extent *originality* as revealed in Table 3.

Tables 4 and 5 both include the pairwise correlation coefficients for IRII and NTR computed using three different reference windows. Overall, the pairwise correlations between our technological novelty indices and the patent quality indicators mostly decrease or stay at the same level as the reference time window increases. There are two exceptions to this observation: the pairwise correlations between *family size* and IRII11 and IRII13 in PHARM, where we observe a slight increase as the length of the reference window increases. Based on this observation, we will not make a distinction between the same index computed with different reference windows in the discussion below.

Focusing on the pairwise correlations with IRII and NTR - our proposed new measures of ex-ante novelty - we can identify some clear patterns. The IRII, computed for different reference windows, is consistently negatively correlated with the other variables except for a weak positive correlation with *backward citations* in the PHARM sector. Among these negative correlations, the strongest in absolute value is with *Patent scope* and the second-highest is with NTR. The weakest correlation for PHARM is with *family size*, and for COMP with *backward citations* and *forward citations*. A comparison across sectors reveals a general tendency for the strength of pairwise correlation between IRII and the other patent quality indicators to be weaker in PHARM and stronger in COMP.

In the case of NTR, the two sectors are more distinct in the correlation with patent quality indicators. PHARM NTR is only positively correlated with *patent scope* and *generality*, and has weak negative correlations with the other variables. For COMP, NTR is positively correlated with all the conventional patent quality measures. The COMP NTR correlations are also stronger in absolute values compared to those in PHARM.

In contrast, by comparing parts (a) and (b) of each table, we do not detect substantial differences. We can point out that *originality* is only weakly correlated with IRII in PHARM, but the correlation is much stronger in COMP, while the correlation between *originality* and NTR doesn't differ as much. With respect to *generality*, IRII exhibits the stronger correlation in both sectors compared to NTR, and for both IRII and NTR the correlation coefficients are stronger in COMP.

Overall, the discussion on the pairwise correlations presented in Tables 4

and 5 suggest that IRII, our proposed new ex-ante technological recombination novelty index - has the expected signs of correlation coefficients with the established indicators of patent quality. As IRII is an *inverse* index, the negative correlation coefficients are in line with the expectation that a higher intensity of re-combination of technological components, i.e. lower IRII, is associated with higher patent quality as captured by one of the established indicators. For NTR, the picture is more obscure and it is hard to draw a summary of the different coefficient signs and weak correlations, especially in relation to *family size*, *backward* and *forward citations*, and *originality*. We could say that the novelty brought by new technologies in COMP tend to be more aligned with the establish indicators of ex-ante patent quality. Overall, for both sectors, the correlation coefficients between IRII and NTR, respectively, and the other patent quality indicators are low with the notable exception of the pairwise correlation between IRII and *patent quality* where at its highest - in Table 5-b - it can categorised moderately high, suggesting that IRII and NTR capture different information sets. We also note that the degree of correlation varies across sectors which, along with the discussion of earlier figures and tables, points to a need for a sector-specific empirical analysis. To render this analysis more accurate, we endeavour to include variations across socio-economic environments by including sector-country-level controls.

## 4.3 Sector-Country-level Data and Summary Statistics

We source country-sector-level data from the OECD MSTI database which covers a wide range of sector-country-level variables for the OECD member states

and seven non-member ones starting from 1981 onward. The MSTI database contains a wide range of variables from which we have chosen a selection of controls that fall into one of the three categories that are relevant to this study: (1) three sector-level variables: Business and Enterprise R&D expenditure (BERD), trade balance, and export market shares (defined in Table 6 as B_COMP, B_PHARM, TD_COMP, TD_PHARM, TD_XCOMP and TD_XPHARM); (2) country-level capital R&D expenditure variables such as the R&D expenditure in three major segments: Business and Enterprise, Government Intramural, and Higher Education; all measured both in current Purchasing Power Parity (PPP) $ (defined in Table 6 as B_PPP, GV_PPP, and H_PPP); and (3) a country-level human resources in research variable measured in full-time equivalent unit (FTE) (defined in Table 6 as TP_RS).

We use the patent filing date and applicants' residence information, both included in the OECD REGPAT data, to control for cohort effects and country-fixed effects. The information allows us to control for factors that are common for all patents but vary from cohort-to-cohort such as the state of the world economy and world-wide technological frontier; as well as account for differences in regulatory environment and policy at the national level, which are invariant over time in the period under investigation. Since our key independent variables are characteristics of individual patents, the sector-country-year variables obtained from the OECD MSTI database need to be transformed to patent level. We do so by defining patent-level MSTI variables as the weighted average of the sector-country level variable where the weights are the share of applicants residing in each unique country listed on the patent application. For example, consider a PHARM patent that was filed in 2000 and listed two authors located

in Germany and one author in Japan. For this patent, each of the MSTI variables mentioned in the first paragraph of this subsection will be computed using the MSTI variable of PHARM-Germany-2000, weighted by the authors' country share of 2/3, added to the MSTI variable of PHARM-Japan-2000, weighted by Japan's share among authors' residency of 1/3.

The datasets for both sectors cover the period from 1981 to 2018 for 25 countries [11]. Table 7 provides the descriptive statistics of the patent-level MSTI variables for both sectors during the period from 1981 to 2018. We note that PHARM and COMP have very similar values for both the mean and standard deviation statistics of all MSTI variables, thus, any sectoral differences in patent quality is unlikely to be driven by any of these factors.

In Table 8 we present country-level data on the total number of patent applications, number of breakthrough patents, and the percentage of breakthrough patents per sector for the OECD countries in the sample period.[12]. The data in Table 6 suggests an imperfect, at best, correlation, between patent volume and breakthrough rate. In both sectors the countries with the top breakthrough rates, those above 1%, are all placed in the bottom half of the table in terms of applications volume; in PHARM the countries with the most and least patent applications, USA (ISO code US) and Portugal (ISO code PT), have comparable breakthrough rates: 0.427% and 0.420%, respectively; and in COMP, Germany (ISO code DE), the country with the third highest volume, has the forth low-

[11] The dataset includes information from the following countries: Australia, Austria, Belgium, Canada, Switzerland, China, Germany, Denmark, Spain, Finland, France, the United Kingdom, Ireland, Israel, Italy, Japan, South Korea, Luxembourg, the Netherlands, Norway, Russia, Singapore, Sweden, Taiwan and the United States. Data is not available for certain countries in some years, as noted in Tables 10 to 15.

[12] Country codes are taken from the ISO 3166 alpha-2 standard definition issued by the International Organization for Standardization, which can be accessed at: https://www.iso.org/iso-3166-country-codes.html

est breakthrough rate of 0.09%. Based on the data in Table 6 in the following sections we will further explore country-specific influences on the ex-ante and ex-post novelty of patenting activities.

# 5 Relationship between IRII and NTR and Invention Quality

Guided by previous research studying the determinants of invention quality using large data, we use *forward citations* and *breakthrough* as ex-post measures for the ex-post technological quality of an invention due to these variables widely recognized significance in predicting a patent's influence in the future technological development. Our main objective is to study how our two ex-ante novelty measures, IRII and NTR, correlate with the ex-post patent quality beyond what established factors contribute. Among these factors we include other patent-level measures of quality and sector-country level variables that we have already discussed extensively in Section 4. To this end, we perform regression analysis for each sector with and without the country-specific controls to have a more thorough investigation of the role played by IRII and NTR. In addition to IRII and NTR, other patent quality indicators including *patent scope*, *family size* and *backward citations* are included as potentially relevant factors. The two variables *originality* and *generality* are not included in the regression models due to their limited data availability.

Formally, the patent-level regression models for patent $k$ of cohort $t$ are presented below. We start with the Poisson regression model with *forward*

$citations_k$ as the dependent variable:

$$\log(\textit{forward citations})_k = \alpha_0 + \alpha_1 \text{IRIIX}_k + \alpha_2 \text{NTRX}_k + \boldsymbol{\zeta_1}' \mathbf{QI}_k + \boldsymbol{\zeta_2}' \mathbf{MSTI}_k + \mu_k + \lambda_t \tag{3}$$

Next, we present the Probit regression model with $breakthrough_k$ as the dependent variable:

$$Pr[\textit{breakthrough}_k = 1] = \Phi\left(\beta_0 + \beta_1 \text{IRIIX}_k + \beta_2 \text{NTRX}_k + \right.$$
$$\left. + \boldsymbol{\phi_1}' \mathbf{QI}_k + \boldsymbol{\phi_2}' \mathbf{MSTI}_k + \nu_k + \theta_t\right) \tag{4}$$

We recall that $forward\ citations_k$ is the number of forward citations received by patent $k$ in 5 years following its publication and that $breakthrough_k$ is an indicator variable that equals 1 if patent $k$ is among the top 1% of patents filed in the same year $t$ by number of forward citations in the following 5 years. The explanatory variable $\text{IRIIX}_k$ is the Inverse Recombination Intensity Index of patent $k$ computed for one of three reference windows $X \in \{11, 31, 51\}$; Similarly, $\text{NTRX}_k$ is the New Technology Ratio of patent $k$ computed for one of three reference windows $X \in \{11, 31, 51\}$. The additional controls listed in the regression models (3) and (4) are the following: $\mathbf{QI}_k$ which is a vector of patent quality indicators defined in Section 4.1, namely, $patent\ scope_k$, $family\ size_k$, and $backward\ citations_k$; $\mathbf{MSTI}_k$ which is a vector of sector-country specific variables from the OECD MSTI database discussed in Section 4.1 and listed in Table 7; $\mu_k$ and $\nu_k$ are vectors of five country dummy variables that indicate if at least one of the applicants listed on patent $k$'s application resides in one of

the top five countries by volume of patents filed in the sector as presented in Table 6; and $\lambda_t$ and $\theta_t$ are patent $k$'s year of application, $t$, cohort fixed effect.

We first introduce the estimation results without the sector-country MSTI controls, i.e. these are the estimations where the vector of parameters $\zeta_{\mathbf{2}} = \mathbf{0}$ in (3). Tables 9 and 13 present the parameter estimates using *forward citations* as the dependent variable for the two sectors: PHARM and COMP, respectively. We present two variants of the basic model: excluding country-specific dummies (the first three columns of each table); and, including the country of residence fixed effects $\mu_k$ for patents where an author resides in a top five country by ptent application volume (the last three columns of each table). All estimates include the cohort year fixed effects and the other control variables listed in Equation 3. We also estimate the model using IRII and NTR computed with the three different reference windows: 11, 31, and 51. The results show that with the reference window is one year, technological recombination has a positive and statistically significant impact on the number of 5-year forward citations received by a patent for both sectors. At its largest, the marginal effects of IRII suggest that a marginal decrease in IRII would increase the probability of a patent with average sample characteristics by 0.32 and 0.2 percentage points for PHARM and COMP, respectively. Yet, there are some notable differences: The estimated marginal effect of of IRII is the largest in absolute value and significance for IRII11 and then decreases in magnitude and significance for IRII31 and IRII51 in PHARM, with the lowest absolute value and significance using IRII31; whereas the opposite is true in COMP where the largest marginal effect in magnitude and significance is estimated for the model using IRII51.

From the summary statistics in Table 1, we know that the average values and standard deviations of the indices using the three window estimations are remarkably similar for both COMP and PHARM, however, COMP exhibits on average a lower level of re-combination across all three windows, and, as displayed in the time series plot in Figure 6, a lower degree of fluctuation. The tables with the pairwise correlation coefficients (Tables 3 and 4) confirm the same observation that the information content of IRII11, IRII31, IRII51 is much more overlapping in COMP and more distinct across the three indices in PHARM. This would imply that the network of IPC subclasses and its partition, constructed at the first stage of the IRII methodology, using the three different sets of patent filings (1-year window, 3-year window, and 5-year window) identify very similar groupings of IPC subclasses by utilization in COMP and that the groupings are more dependent on the cohort of patent filings in PHARM, a scenario similar to the examples in Figure 4 and 5. This could explain why in COMP we observe similar statistical significance across the three versions of IRII; while in PHARM the statistical significance varies.

Differences in the process of innovation are also suggested by the results on the NTR. The addition of relatively more new subgroups previously non-utilized by patents in the sector as measured by NTR shows a significant negative marginal effect on the number of forward citations for COMP, regardless of the reference window lengths, but decreases in significance for PHARM as the reference window increases from one year to five years. The absolute value of NTR's marginal effect is smaller with COMP than PHARM, indicating a larger value of NTR with COMP patents on average, as shown in Table 1. Notably, the estimated marginal effects are all negative. This suggests that the poten-

tial for a patent to be cited by future inventions in either of the two sector is not enhanced by incorporating new knowledge origins from other technological sectors but rather the novel use of core sector technological components that is captured by IRII.

Turning our attention to the established measures of patent quality, we observe that *patent scope*, *family size*, and *backward citations* contribute significantly to the number of future patents influenced which confirms finding found previously in the literature. All marginal effects are of similar magnitude across all variables, model estimations, and both technology sectors.

In the last three columns of each table, we present the estimation of the modules including country-specific fixed effect. The inclusion of the country dummies do not effect the estimated coefficients of the other variables, however, they point to statistically significant differences in a patent's potential for having more forward citations depending on the residence of its authors. In particular, a patent application with at least one applicant residing in the U.S. or Japan leads to a higher likelihood of an increase in its number of forward citations compared to an average application where no applicant is from one of the top 5 countries in both sectors though the magnitude of the effects is larger in COMP. In PHARM, having an applicant residing in France or the UK could lower the number of received forward citations compared to an application where none of the applicants reside in a top 5 country by the volume of applications. In COMP, the South Korea indicator variable has the largest statistically significant and positive coefficient, however, both having an applicant from France or Germany is estimated to lower the probability in a statistically significant sense

in this sector. These results suggest that even among the most prolific countries there are important national-level factors that influence the potential of their resident innovators to produce a breakthrough patent. Before we turn to the investigation of this aspect further in the following discussion, let's look at the results of Equation 4.

The estimates using *breakthrough* as independent variable are presented in Table 9 for PHARM and Table 13 for COMP. Compared to Table 8 and 12, we could summarize that the results of IRII and NTR show similar sectoral differences, as well as the pattern with different reference windows, but the magnitude and statistical significance of these two key indices across all the estimation models are lower. The likely explanation is that *breakthrough* is defined to label the patents with the top 1% by the number of forward citations their receive. Therefore, compared to *forward citations* which is a continuous variable, *breakthrough* is a less differentiating outcome. Aside from this, the country-specific fixed effects also show some differences compared to the estimates against *forward citations*: In the PHARM sector, having an applicant from Japan now has the largest and most statistically significant and positive coefficient, overtaking the U.S. The negative significance of the France indicator is at 5% level. In the COMP sector, the country indicator U.S. supersedes South Korea to be the dummy variable with the largest magnitude of effect, while having an applicant from Germany becomes a statistically significant and negative factor for the likelihood of a patent being a breakthrough.

Now let's bring the individualized country R&D input variables into the equations. Table 10 and 11 provide the results of PHARM, with *forward cita-*

*tions* and *breakthrough* as the independent variable, respectively, and including the individualized MSTI variables. And Table 14 and 15 are the COMP versions. In each table, we first present the estimation results using IRII11 and NTR11 in column 1-4, where each model involves different MSTI variables. Then, based on the model in column 4, we extend the analysis to different reference window lengths (column 4-6) and to include the country-specific dummy variables indicating the five most productive countries by number of patent filings in each sector (column 7-9). The results show that the inclusion of individualized MSTI variables does not affect the estimates of IRII and NTR for PHARM. However, in the COMP sectors, adding the sectoral MSTI control variables results in a decrease in the statistical significance in NTR, as show in column 3 and 4 in Table 14 and 15. We still see that with PHARM, IRII11 and NTR11 have the most statistically significant effect on the independent variables, and with COMP, IRII51 has the largest magnitude of the coefficient's absolute value and the highest significance level. With both sectors, IRII and NTR still have negative effects on either the probability of being a breakthrough patent or being cited by future patents, across all the models.

Finally, we come to the individualized MSTI variables. In the PHARM sector, with the independent variable being *forward citations*, the sectoral export market share is the only MSTI regressor that's weakly significant at 5% level when the country-specific dummies are excluded (column 3-6 in Table 10). Then Table 11 shows that the Government Intramural Expenditure on R&D has a statistically significant and negative effect on the probability that a patent is *breakthrough* as long as the country-specific fixed effects are excluded, and having higher R&D expenditure in higher education has a statistically significant

and positive effect on a PHARM patent to be the outstanding top 1% forward citation elites across all the models. For a COMP patent, as shown in Table 14, the inclusion of sectoral variables appears to cause changes to the model estimates: starting from column 3, GV_PPP turns to positive and statistically significant from negative, B_PPP becomes statistically significant and negative, and H_PPP loses its statistical significance. The BERD performed in COMP (B_COMP) also has a significant and positive relationship with the number of forward citations. But all these variables' statistical significance drop when the country-specific dummies are included in the last three columns. In Table 15, the inclusion of sectoral MSTI variables from column 3 also brings changes to the estimates. But with *breakthrough* as the dependent variable, GV_PPP and B_PPP remain statistically significant in the last 3 columns with the five country dummy variables.

Compared to IRII and NTR, the individualized MSTI variables show less of a persistent pattern and the regression estimates are less robust. The COMP results seem to show that the sectoral variables (sectoral business enterprise expenditure on R&D, trade balance, and export market share) could interact with the non-sectoral expenditure variables. This could be caused by the inherent relationship between some of the variables by definition. For example, a higher B_PPP is likely to be associated with higher sectoral BERD. However, as we are modeling the regression using weighted country-level data on individual patents, and the MSTI variables are included as controls, we will not do into depth in this field. It is still worth-mentioning that these estimates again show that the two sectors under analysis have their unique characteristics.

# 6 Discussion and Conclusion

Throughout all the analysis in this paper, from the descriptive statistics of IRII and NTR to the regression results, from the network visualization to the correlation coefficients between IRII and NTR and the conventional patent quality indicators, patterns of sectoral differences can be observed. This calls for further reflection and investigation. In PHARM a patent's potential to be breakthrough carries the largest correlation to the level of re-combination of knowledge origins relative to the patterns observed in the previous cohort of applications and in the COMP this is relative to the patterns derived from pooling the application data across the previous five cohorts. One may see this as evidence of the greater potential of an invention to influence new technological trends if it is destructive to the long-term pattern of utilization of subclasses in the COMP sector while in the PHARM sector a similar breakthrough potential is generated by a similar degree of destruction to the short-term pattern. Given that in PHARM compared to COMP we observe on average a higher degree of re-combination of IPC subclasses in each patent application, lower ratio of new subgroups (see Table 1), higher patent scope and lower rate of backward citations (see Table 2) and that the grouping of IPC subclasses by usage differs more depending on the length of the observation window, we may be detecting the effects of a different technological process of invention between the two sectors. We conjecture that in PHARM the quality of an invention is linked to destructive innovation at the frontier and patterns of combining knowledge become obsolete quicker; in COMP the sector-specific technological frontier is better captured by the patterns of utilization of IPC subclasses over a longer time period as inventions

tend to be more geared towards entering new technological fields rather than making existing sector technology obsolete; This would explain why the largest marginal effect is estimated for IRII51.

# 7 Conclusion

This research fills the gap in current literature by proposing an ex-ante perspective into the emergence and composition of disruptive innovation. With PHARM and COMP used for empirical analysis, the method we propose is generic and can be applied regardless of sector or scale. The relationships we have found between the new metrics and the conventional patent indicators confirm our proposed significance of technological recombination and the new knowledge in producing high-quality, influential inventions. The analysis involving country-level R&D resources further validates the value of IRII and NTR as they bring added information where the MSTI variables do not show significant predicting power. National and institutional decision makers may benefit from this study to monitor the status of disruptive innovation and obtain information to take actions at an earlier stage. Our research also reveals sectoral differences between Pharmaceuticals and Computer Technology, and divergence in technological strengths and specialisations among countries. The differences can be clearly seen from summary statistics of basic variables through to the regression analysis. This prompts us to remind the readers that comparison between the two sectors and interpretation of results must be done in the context sector-specific knowledge of the process new technology accumulation and creation. These issues are not fully explored in this paper but present promising

dimensions for future research.

# Figures

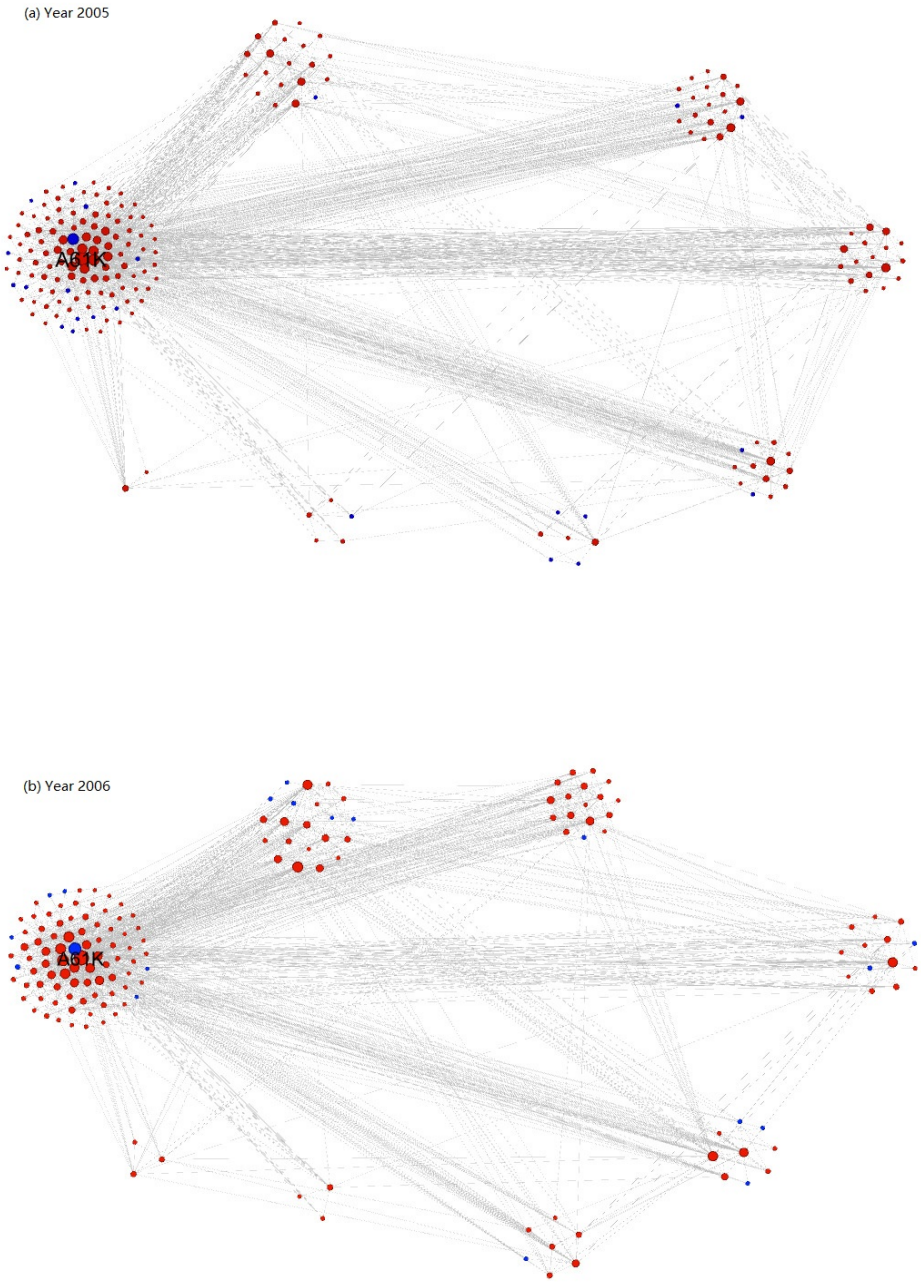


Figure 1: 8-cluster PHARM network partition highlighting subclasses containing new subgroups with 1-year reference window

Nodes in red color represent the subclasses containing new subgroups that are not found in patents of the PHARM sector in the previous year, and the the blue nodes do not contain new subgroups. Node size is proportional to the node degree in the network, i.e. the number of connections to other nodes. Labels are shown for nodes with degree above the median value. The edge lengths are not indicative of the connection strength.

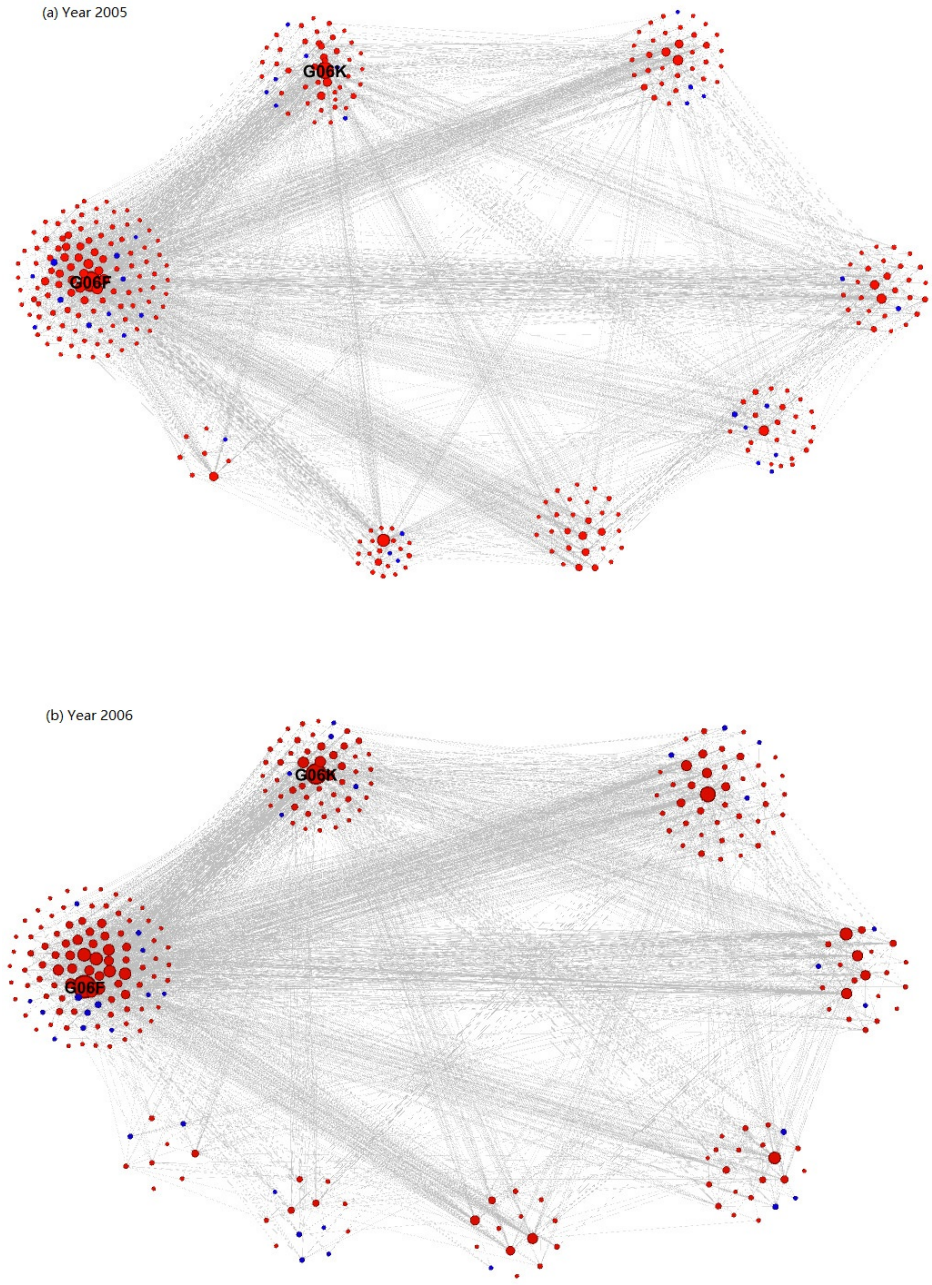


Figure 2: 8-cluster COMP network partition highlighting subclasses containing new subgroups with 1-year reference window

Nodes in red color represent the subclasses containing new subgroups that are not found in patents of the COMP sector in the previous year, and the the blue nodes do not contain new subgroups. Node size is proportional to the node degree in the network, i.e. the number of connections to other nodes. Labels are shown for nodes with degree above the median value. The edge lengths are not indicative of the connection strength.

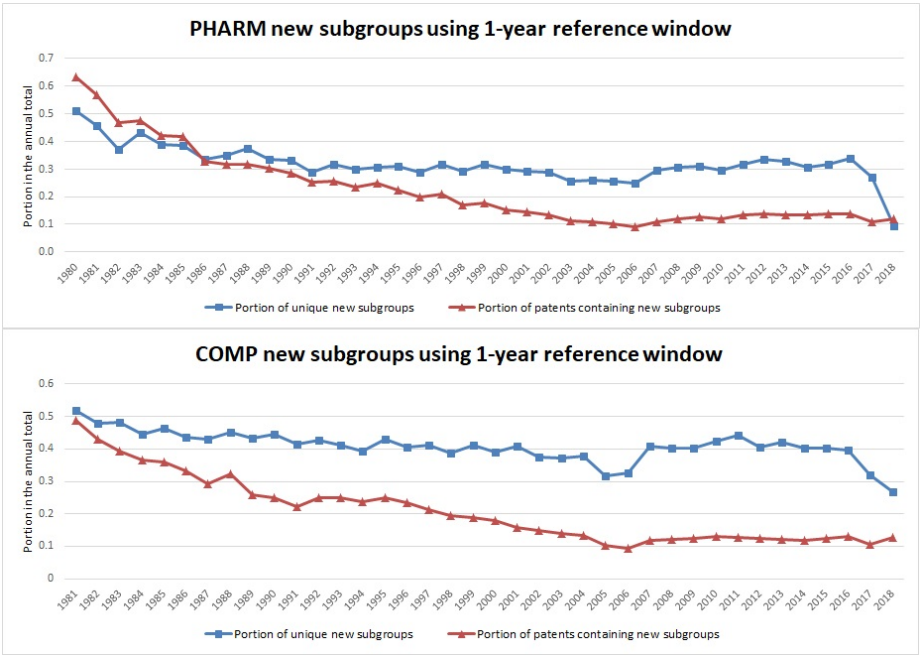


Figure 3: Subgroups portion in annual total quantities with one-year reference window

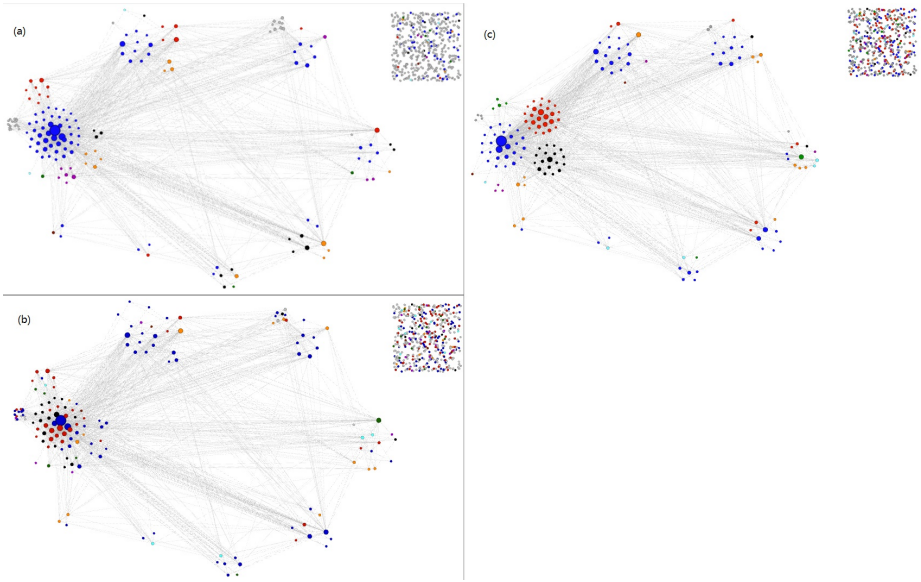


Figure 4: 8-cluster PHARM network partition of 2006 showing recombination in reference to the previous 1-year and 5-year windows

Each panel of the figures shows the partition with eight clusters generated from the network constructed using he cohort of PHARM patents filed in 2006, plus a square cluster of nodes at the upper-right corner that represents unconnected nodes and nodes (subclasses) not used in the patent cohort. In Panel (a), each cluster is further divided into eight sub-clusters, each with a distinct color representing the network partition generated using the cohot of PHARM patents filed in 2005. Panel (b) is in the same layout as Panel (a), but the color palette represents the network partition of the patent cohort in the 5-year reference window, 2001-2005. Nodes in Panel (c) have the same color representation as Panel (b), but with the sub-clusters visually grouped.

The color palette follows the cluster sizes in each partitioning: light grey for unconnected or unused nodes, blue for nodes in the largest cluster, red for nodes in the second largest cluster, black the third, and yellow, purple, green, light blue, and brown. Node size is proportional to the node degree in the network, i.e. the number of connections to other nodes. Node labels are omitted for visual clearance. The edge lengths are not indicative of the connection strength.

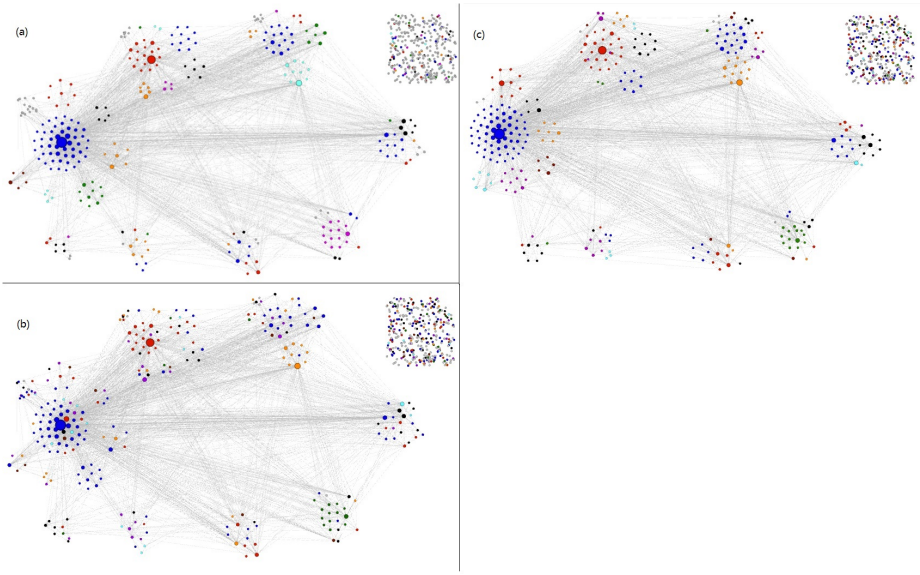


Figure 5: 8-cluster COMP network partition of 2006 showing recombination in reference to the previous 1-year and 5-year windows

Each panel of the figures shows the partition with eight clusters generated from the network constructed using he cohort of COMP patents filed in 2006, plus a square cluster of nodes at the upper-right corner that represents unconnected nodes and nodes (subclasses) not used in the patent cohort. In Panel (a), each cluster is further divided into eight sub-clusters, each with a distinct color representing the network partition generated using the cohort of COMP patents filed in 2005. Panel (b) is in the same layout as Panel (a), but the color palette represents the network partition of the patent cohort in the 5-year reference window, 2001-2005. Nodes in Panel (c) have the same color representation as Panel (b), but with the sub-clusters visually grouped.

The color palette follows the cluster sizes in each partitioning: light grey for unconnected or unused nodes, blue for nodes in the largest cluster, red for nodes in the second largest cluster, black the third, and yellow, purple, green, light blue, and brown. Node size is proportional to the node degree in the network, i.e. the number of connections to other nodes. Node labels are omitted for visual clearance. The edge lengths are not indicative of the connection strength.

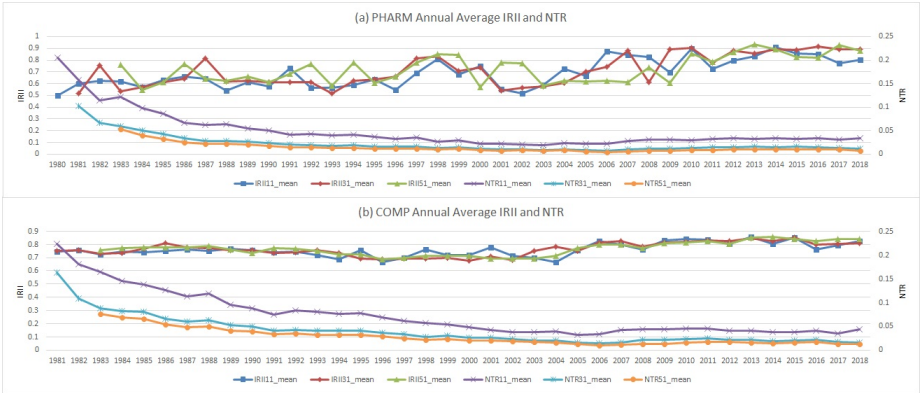


Figure 6: Annual sectoral average IRII and NTR with different reference windows

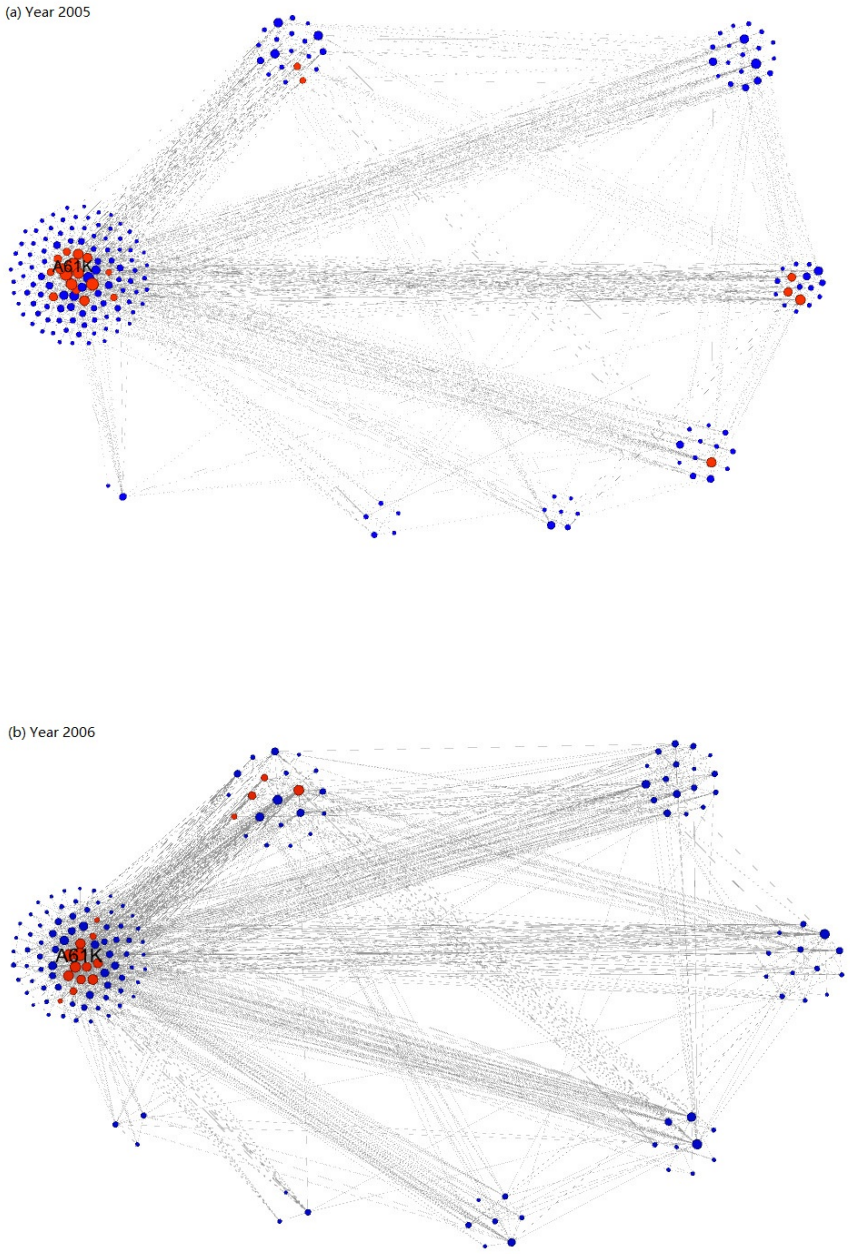


Figure 7: 8-cluster PHARM network partition highlighting subclasses of Breakthrough patents

Nodes in red color represent the subclasses assigned to Breakthrough PHARM patents of the years, and the blue nodes are not assigned to Breakthrough patents. Node size is proportional to the node degree in the network, i.e. the number of connections to other nodes. Labels are shown for nodes with degree above the median value. The edge lengths are not indicative of the connection strength.

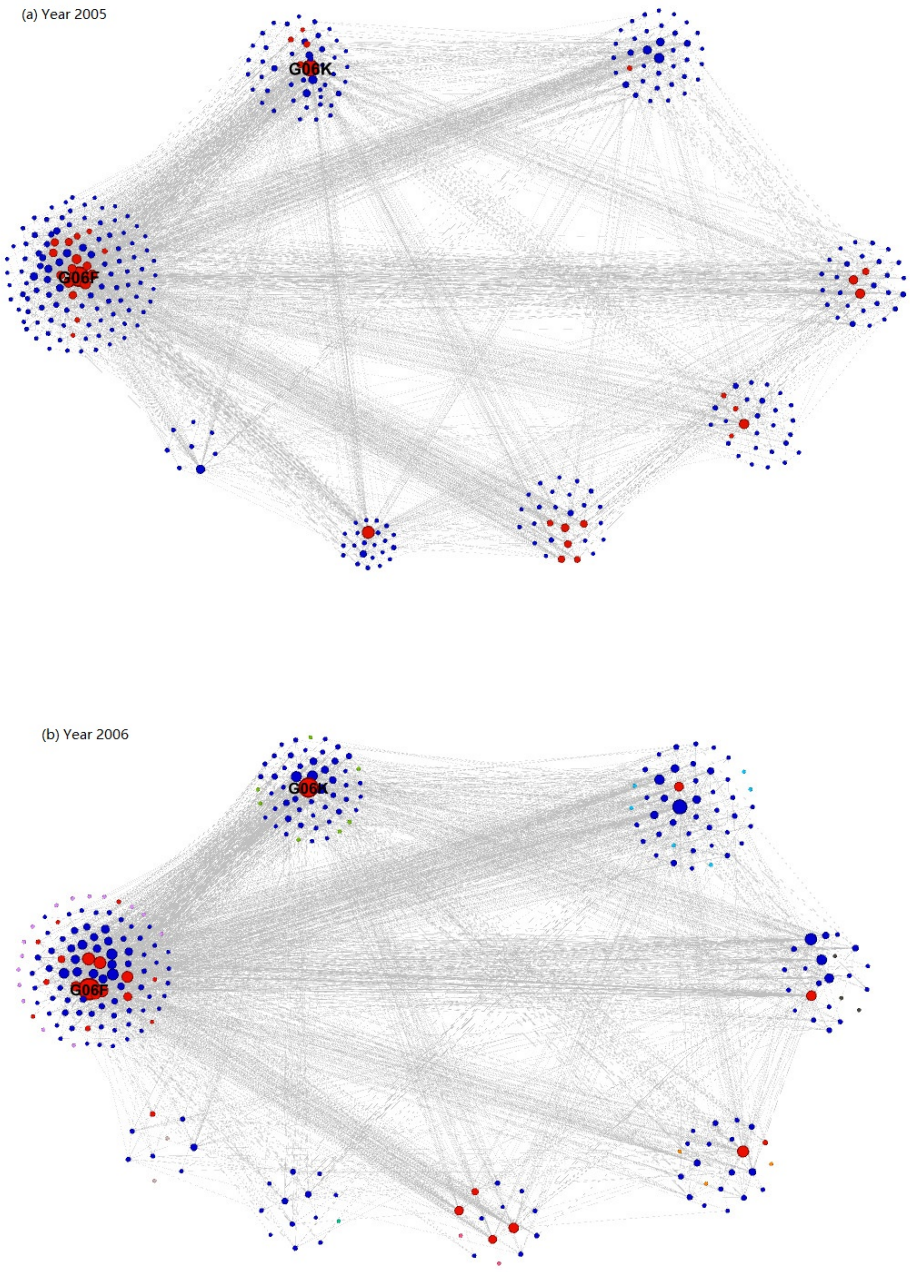


Figure 8: 8-cluster COMP network partition highlighting subclasses of Breakthrough patents

Nodes in red color represent the subclasses assigned to Breakthrough COMP patents of the years, and the blue nodes are not assigned to Breakthrough patents. Node size is proportional to the node degree in the network, i.e. the number of connections to other nodes. Labels are shown for nodes with degree above the median value. The edge lengths are not indicative of the connection strength.

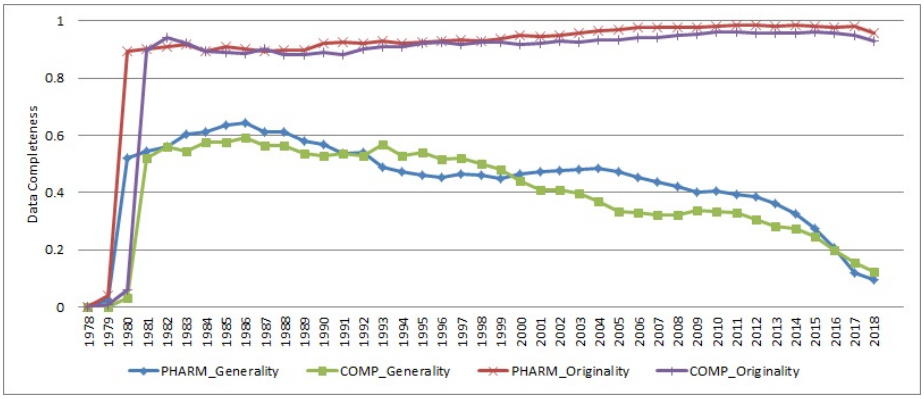


Figure 9: Originality and Generality data completeness of each sector

# Tables

Table 1: Sectoral IRII and NTR summary statistics

| PHARM (1980-2018) | | | | | |
|---|---|---|---|---|---|
| Variable | Sample size | Mean | Std. dev. | Min | Max |
| IRII11 | 278,990 | 0.7171 | 0.2314 | 0.0278 | 1 |
| IRII31 | 277,922 | 0.7295 | 0.2298 | 0.0556 | 1 |
| IRII51 | 275,003 | 0.7272 | 0.2341 | 0.0625 | 1 |
| NTR11 | 278,990 | 0.0347 | 0.1000 | 0 | 1 |
| NTR31 | 277,922 | 0.0156 | 0.0653 | 0 | 1 |
| NTR51 | 275,003 | 0.0105 | 0.0532 | 0 | 1 |

| COMP (1981-2018) | | | | | |
|---|---|---|---|---|---|
| Variable | Sample size | Mean | Std. dev. | Min | Max |
| IRII11 | 282,506 | 0.7726 | 0.2540 | 0 | 1 |
| IRII31 | 282,506 | 0.7769 | 0.2490 | 0.0278 | 1 |
| IRII51 | 280,401 | 0.7763 | 0.2475 | 0.0400 | 1 |
| NTR11 | 282,506 | 0.0502 | 0.1393 | 0 | 1 |
| NTR31 | 282,506 | 0.0263 | 0.0983 | 0 | 1 |
| NTR51 | 280,401 | 0.0195 | 0.0834 | 0 | 1 |

# References


Gautam Ahuja and Curba Morris Lampert. Entrepreneurship in the large corporation: A longitudinal study of how established firms create breakthrough inventions. *Strategic management journal*, 22(6-7):521–543, 2001.

Sam Arts and Reinhilde Veugelers. Technology familiarity, recombinant novelty, and breakthrough invention. *Industrial and Corporate Change*, 24(6):1215–1246, 2015.

Joseph L Bower and Clayton M Christensen. Disruptive technologies: catching the wave. 1995.

Kristie Briggs. Co-owner relationships conducive to high quality joint patents. *Research Policy*, 44:1566–1573, 2015.

Paola Criscuolo and Bart Verspagen. Does it matter where patent citations come from? inventor vs. examiner citations in european patents. *Research policy*, 37(10):1892–1908, 2008.

Kristina B Dahlin and Dean M Behrens. When is an invention really radical? defining and measuring technological radicalnes. *Research Policy*, 34:717–737, 2005.

Robert E Evenson. Patents, r&d, and invention potential: international evidence. *The American Economic Review*, 83(2):463–468, 1993.

Carsten Fink, Mosahid Khan, and Hao Zhou. Exploring the worldwide patent surge. *Economics of Innovation and New Technology*, 25(2):114–142, 2016.

Lee Fleming. Recombinant uncertainty in technological search. *Management science*, 47(1):117–132, 2001.

Alberto Galasso. Cep discussion paper no 1072 august 2011 trading and enforcing patent rights alberto galasso, mark schankerman and carlos j. serrano. 2011.

Yuan Gao and Emiliya Lazarova. A new empirical index to track the technological novelty of inventions: A sector level analysis. Technical report, School of Economics, University of East Anglia, Norwich, UK., 2022.

Paul Gompers, Josh Lerner, and David Scharfstein. Entrepreneurial spawning: Public corporations and the genesis of new ventures, 1986 to 1999. *The journal of Finance*, 60(2):577–614, 2005.

Zvi Griliches. Patents: Recent trends and puzzles, 1989.

Zvi Griliches. Patent statistics as economic indicators: a survey. In *R&D and productivity: the econometric evidence*, pages 287–343. University of Chicago Press, 1998.

Zvi Griliches, Ariel Pakes, and Bronwyn H Hall. The value of patents as indicators of inventive activity, 1986.

Bronwyn H Hall and Manuel Trajtenberg. Uncovering gpts with patent data. Technical report, National Bureau of Economic Research, 2004.

Bronwyn H Hall, Adam Jaffe, and Manuel Trajtenberg. Market value and patent citations. *RAND Journal of economics*, pages 16–38, 2005.

Dietmar Harhoff and Stefan Wagner. The duration of patent examination at the european patent office. *Management Science*, 55(12):1969–1984, 2009.

Dietmar Harhoff, Frederic M Scherer, and Katrin Vopel. Citations, family size,

opposition and the value of patent rights. *Research policy*, 32(8):1343–1363, 2003.

Rebecca Henderson, Adam B Jaffe, and Manuel Trajtenberg. Universities as a source of commercial technology: a detailed analysis of university patenting, 1965–1988. *Review of Economics and statistics*, 80(1):119–127, 1998.

Zoltan J. Acs and David B Audretsch. Patents as a measure of innovative activity. *Kyklos*, 42(2):171–180, 1989.

Adam B Jaffe and Manuel Trajtenberg. *Patents, citations, and innovations: A window on the knowledge economy*. MIT press, 2002.

Sarah Kaplan and Keyvan Vakili. The double-edged sword of recombination in breakthrough innovation. *Strategic Management Journal*, 36:1435–1457, 2015. URL DOI:10.1002/smj.2294.

PH Kaya and A Joseph. Schumpeter's perspective on innovation international'. *Journal of Economics, Commerce and Management*, 3(8):25–37, 2015.

William R Kerr. Breakthrough inventions and migrating clusters of innovation. *Journal of urban Economics*, 67(1):46–60, 2010.

Jinyoung Kim and Gerald Marschke. Accounting for the recent surge in us patenting: changes in r&d expenditures, patent yields, and the high tech sector. *Economics of Innovation and New Technology*, 13(6):543–558, 2004.

Jean O Lanjouw and Mark Schankerman. Characteristics of patent litigation: a window on competition. *RAND journal of economics*, pages 129–151, 2001.

Jean O Lanjouw and Mark Schankerman. Patent quality and research productivity: Measuring innovation with multiple indicators. *The Economic Journal*, 114(495):441–465, 2004.

Jean O Lanjouw, Ariel Pakes, and Jonathan Putnam. How to count patents and value intellectual property: The uses of patent renewal and application data. *The journal of industrial economics*, 46(4):405–432, 1998.

Anne Layne-Farrar and Josh Lerner. To join or not to join: Examining patent pool participation and rent sharing rules. *International Journal of Industrial Organization*, 29(2):294–303, 2011.

Joshua Lerner. The importance of patent scope: an empirical analysis. *The RAND Journal of Economics*, pages 319–333, 1994.

Mikhail Makarov. The eighth edition of the ipc. *World Patent Information*, 28 (2):122–126, 2006.

Stéphane Maraut, Hélène Dernis, Colin Webb, Vincenzo Spiezia, and Dominique Guellec. The oecd regpat database: a presentation. *OECD Science, Technology and Industry Working Papers*, 2008(2):0_1, 2008.

Carlo Piccardi. Finding and testing network communities by lumped markov chains. *PloS one*, 6(11):e27028, 2011.

David Popp, Nidhi Santen, Karen Fisher-Vanden, and Mort Webster. Technology variation vs. r&d uncertainty: What matters most for energy patent success? *Resource and Energy Economics*, 35(4):505–533, 2013.

Pierre Régibeau and Katharine Rockett. Innovation cycles and learning at the

patent office: does the early patent get the delay? *The Journal of Industrial Economics*, 58(2):222–246, 2010.

Lori Rosenkorf and Atul Nerkar. Beyond local search: Boundary-spanning, exploration, and impact in the optical disc industry. *Strategic Management Journal*, 22:287–306, 2001.

Chandra Nath Saha and Sanjib Bhattacharya. Intellectual property rights: An overview and implications in pharmaceutical industry. *Journal of Advanced Pharmaceutical Technology & Research*, 2:88–93, 2011. doi: 10.4103/2231-4040.82952.

Joseph Schumpeter. Creative destruction. *Capitalism, socialism and democracy*, 825:82–85, 1942.

Joseph A Schumpeter. *Theory of economic development*. Routledge, 2017.

William G Shepherd and Joanna M Shepherd. *The economics of industrial organization*. Waveland Press, 2003.

Daniela Silvestri, Massimo Riccaboni, and Antonio Della Malva. Sailing in all winds: Technological search over the business cycle. *Research Policy*, 47(10): 1933–1944, 2018.

Jasjit Singh. Distributed r&d, cross-regional knowledge integration and quality of innovative output. *Research Policy*, 37(1):77–96, 2008.

Mariagrazia Squicciarini, Hélène Dernis, and Chiara Criscuolo. Measuring patent quality. 2013.

Manish K Srivastava and Devi R Gnyawali. When do relational resources matter? leveraging portfolio technological resources for breakthrough innovation. *Academy of Management Journal*, 54(4):797–810, 2011.

Jessica C Stahl. Mergers and sequential innovation: evidence from patent citations. 2010.

Deborah Strumsky and Jose Lobo. Identifying the sources of technological novelty in the process of invention. *Research Policy*, 44:1445–1461, 2015.

Manuel Trajtenberg. A penny for your quotes: patent citations and the value of innovations. *The Rand journal of economics*, pages 172–187, 1990.

Manuel Trajtenberg, Rebecca Henderson, and Adam Jaffe. University versus corporate patents: A window on the basicness of invention. *Economics of Innovation and new technology*, 5(1):19–50, 1997.

Dennis Verhoeven, Jurriën Bakker, and Reinhilde Veugelers. Measuring technological novelty with patent-based indicators. *Research Policy*, 45(3):707–723, 2016.

WIPO. World intellectual property indicators. Technical report, 2022. DOI:10.34667/tind.47082.

Table 2: Conventional patent quality indicators list

| Indicator | Definition |
|---|---|
| Patent scope | The number of distinct 4-digit IPC subclasses assigned to the patent |
| Family size | The number of patent offices operating in different jurisdictions at which a given invention has been protected |
| Backward citation | The number of citations of prior art listed on a patent applications as a source of knowledge in the development of the invention |
| Forward citation | The number of citations a patent receives within five years after the publication date |
| Breakthrough | A binary variable which equals 1 for patents in the the top 1% by the number of forward citations among those filed in the same year within the next 5 years; and 0 otherwise. |
| Originality | A measure of knowledge diversification in the development of a patent based on the range of subclasses included in the backward citations of the patent application |
| Generality | The ex-post counterpart of Originality, by using forward citations to capture the scope and degree of general-purpose technology that a patent enables |

Table 3: Sectoral patent quality indicators summary statistics

| PHARM (1980-2018) | | | | | |
|---|---|---|---|---|---|
| Variable | Sample size | Mean | Std. dev. | Min | Max |
| patent scope | 278,990 | 3.0289 | 1.4706 | 0 | 21 |
| family size | 278,990 | 9.8403 | 7.6757 | 1 | 57 |
| backward citations | 278,990 | 7.8111 | 20.9826 | 0 | 1013 |
| forward citations | 278,990 | 1.3189 | 4.9999 | 0 | 672 |
| originality | 266,616 | 0.7955 | 0.1626 | 0 | 0.9863 |
| generality | 122,563 | 0.5056 | 0.2247 | 0 | 0.9388 |

| COMP (1981-2018) | | | | | |
|---|---|---|---|---|---|
| Variable | Sample size | Mean | Std. dev. | Min | Max |
| patent scope | 282,506 | 2.0965 | 1.2761 | 0 | 30 |
| family size | 282,506 | 4.7957 | 2.8728 | 1 | 45 |
| backward citations | 282,506 | 4.5430 | 6.5689 | 0 | 498 |
| forward citations | 282,506 | 0.9442 | 3.1078 | 0 | 270 |
| originality | 264,878 | 0.6798 | 0.2272 | 0 | 0.9823 |
| generality | 103,344 | 0.3529 | 0.2803 | 0 | 0.9378 |

Table 4: PHARM patent-level IRII and NTR correlations with quality indicators

Table 4-a (1980-2018)

| N=275,003 | IRII11 | IRII31 | IRII51 | NTR11 | NTR31 | NTR51 | patent scope | family size | backward citations | forward citations |
|---|---|---|---|---|---|---|---|---|---|---|
| IRII11 | 1.0000 | | | | | | | | | |
| IRII31 | 0.7032 | 1.0000 | | | | | | | | |
| IRII51 | 0.5136 | 0.5449 | 1.0000 | | | | | | | |
| NTR11 | -0.1992 | -0.1731 | -0.1637 | 1.0000 | | | | | | |
| NTR31 | -0.1841 | -0.1590 | -0.1475 | 0.7702 | 1.0000 | | | | | |
| NTR51 | -0.1767 | -0.1524 | -0.1387 | 0.6866 | 0.8958 | 1.0000 | | | | |
| patent scope | *-0.4982* | *-0.4475* | *-0.3606* | *0.1861* | *0.1630* | *0.1514* | 1.0000 | | | |
| family size | *-0.0197* | *-0.0505* | *-0.0247* | *-0.0392* | *-0.0336* | *-0.0297* | 0.0731 | 1.0000 | | |
| backward citations | *0.0424* | *0.0540* | *0.0143* | *-0.0178* | *-0.0129* | *-0.0132* | -0.0256 | 0.0588 | 1.0000 | |
| forward citations | *-0.0619* | *-0.0630* | *-0.0443* | *-0.0064* | *-0.0051* | *-0.0013* | 0.0977 | 0.1530 | 0.1044 | 1.0000 |

Table 4-b (1980-2014)

| | IRII11 | IRII31 | IRII51 | NTR11 | NTR31 | NTR51 | patent scope | family size | backward citations | forward citations | originality | generality |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| IRII11 | 1.0000 | | | | | | | | | | | |
| IRII31 | 0.7072 | 1.0000 | | | | | | | | | | |
| IRII51 | 0.4936 | 0.5401 | 1.0000 | | | | | | | | | |
| NTR11 | -0.2077 | -0.1750 | -0.1653 | 1.0000 | | | | | | | | |
| NTR31 | -0.1870 | -0.1603 | -0.1478 | 0.7780 | 1.0000 | | | | | | | |
| NTR51 | -0.1765 | -0.1493 | -0.1386 | 0.6902 | 0.8981 | 1.0000 | | | | | | |
| patent scope | *-0.4986* | *-0.4485* | *-0.3492* | *0.1809* | *0.1586* | *0.1496* | 1.0000 | | | | | |
| family size | *-0.0135* | *-0.0390* | *-0.0107* | *-0.0357* | *-0.0331* | *-0.0290* | 0.0671 | 1.0000 | | | | |
| backward citations | *0.0478* | *0.0494* | *0.0154* | *-0.0196* | *-0.0143* | *-0.0129* | -0.0244 | 0.0621 | 1.0000 | | | |
| forward citations | *-0.0550* | *-0.0519* | *-0.0342* | *-0.0061* | *-0.0050* | *-0.0015* | 0.0943 | 0.1508 | 0.1128 | 1.0000 | | |
| originality | *-0.0529* | *-0.0711* | *-0.0722* | *-0.0213* | *-0.0226* | *-0.0191* | 0.1407 | 0.0722 | 0.1461 | 0.0468 | 1.0000 | |
| generality | *-0.1939* | *-0.1795* | *-0.1514* | *0.0272* | *0.0107* | *0.0060* | 0.3294 | 0.0732 | 0.0257 | 0.1414 | 0.1405 | 1.0000 |

Table 5: COMP patent-level IRII and NTR correlations with quality indicators

Table 5-a (1981-2018)

| N=280,401 | IRII11 | IRII31 | IRII51 | NTR11 | NTR31 | NTR51 | patent scope | family size | backward citations | forward citations |
|---|---|---|---|---|---|---|---|---|---|---|
| IRII11 | 1.0000 | | | | | | | | | |
| IRII31 | 0.8003 | 1.0000 | | | | | | | | |
| IRII51 | 0.7903 | 0.8516 | 1.0000 | | | | | | | |
| NTR11 | -0.2857 | -0.2543 | -0.2400 | 1.0000 | | | | | | |
| NTR31 | -0.2463 | -0.2235 | -0.2089 | 0.7852 | 1.0000 | | | | | |
| NTR51 | -0.2319 | -0.2103 | -0.1978 | 0.7107 | 0.9071 | 1.0000 | | | | |
| patent scope | *-0.6721* | *-0.6511* | *-0.6564* | *0.2977* | *0.2606* | *0.2467* | 1.0000 | | | |
| family size | *-0.1219* | *-0.1189* | *-0.1260* | *0.0749* | *0.0701* | *0.0704* | 0.1835 | 1.0000 | | |
| backward citations | *-0.0624* | *-0.0545* | *-0.0573* | *0.0399* | *0.0335* | *0.0307* | 0.0821 | 0.0330 | 1.0000 | |
| forward citations | *-0.0931* | *-0.0957* | *-0.0982* | *0.0483* | *0.0443* | *0.0467* | 0.1590 | 0.1805 | 0.0589 | 1.0000 |

Table 5-b (1981-2014)

| | IRII11 | IRII31 | IRII51 | NTR11 | NTR31 | NTR51 | patent scope | family size | backward citations | forward citations | originality | generality |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| IRII11 | 1.0000 | | | | | | | | | | | |
| IRII31 | 0.7984 | 1.0000 | | | | | | | | | | |
| IRII51 | 0.7968 | 0.8510 | 1.0000 | | | | | | | | | |
| NTR11 | -0.2877 | -0.2534 | -0.2361 | 1.0000 | | | | | | | | |
| NTR31 | -0.2485 | -0.2237 | -0.2047 | 0.7902 | 1.0000 | | | | | | | |
| NTR51 | -0.2319 | -0.2091 | -0.1932 | 0.7124 | 0.9077 | 1.0000 | | | | | | |
| patent scope | *-0.6735* | *-0.6527* | *-0.6621* | *0.2963* | *0.2605* | *0.2473* | 1.0000 | | | | | |
| family size | *-0.1265* | *-0.1214* | *-0.1261* | *0.0753* | *0.0710* | *0.0727* | 0.1988 | 1.0000 | | | | |
| backward citations | *-0.1005* | *-0.0907* | *-0.0952* | *0.0624* | *0.0541* | *0.0500* | 0.1264 | 0.0665 | 1.0000 | | | |
| forward citations | *-0.0951* | *-0.0959* | *-0.0964* | *0.0489* | *0.0446* | *0.0482* | 0.1689 | 0.1824 | 0.0548 | 1.0000 | | |
| originality | *-0.2182* | *-0.2123* | *-0.2241* | *0.0692* | *0.0612* | *0.0572* | 0.2577 | 0.0509 | 0.2824 | 0.0374 | 1.0000 | |
| generality | *-0.2744* | *-0.2791* | *-0.2736* | *0.1142* | *0.0943* | *0.0864* | 0.3241 | 0.0836 | 0.0575 | 0.1835 | 0.2202 | 1.0000 |

Table 6: OECD MSTI Variable List

| MSTI variable | definition | unit |
|---|---|---|
| B_COMP | BERD performed in the computer, electronic and optical industry (current PPP $) | USD $MM |
| B_PHARM | BERD performed in the pharmaceutical industry (current PPP $) | USD $MM |
| TD_BCOMP | Trade Balance: Computer, electronic and optical industry (current prices) | USD $MM |
| TD_BPHARM | Trade Balance: Pharmaceutical industry (current prices) | USD $MM |
| TD_XCOMP | Export market share: Computer, electronic and optical industry | % |
| TD_XPHARM | Export market share: Pharmaceutical industry | % |
| B_PPP | Business Enterprise Expenditure on R&D (BERD) at current PPP $ | USD $MM |
| GV_PPP | Government Intramural Expenditure on R&D (GOVERD) at current PPP $ | USD $MM |
| H_PPP | Higher Education Expenditure on R&D (HERD) at current PPP $ | USD $MM |
| TP_RS | Total researchers (FTE) | FTE |

Table 7: Summary statistics of weighted MSTI variables at patent level for each sector

| PHARM (1981-2018) | | | | | |
|---|---|---|---|---|---|
| Variable | Sample size | Mean | Std. dev. | Min | Max |
| GV_PPP | 259,786 | 19362.5200 | 18181.0000 | 16.5214 | 84124.8000 |
| B_PPP | 256,906 | 114276.6000 | 110427.3000 | 10.6821 | 429134.4000 |
| H_PPP | 259,428 | 21850.8700 | 20196.3200 | 0.9419 | 74722.0000 |
| TP_RS | 246,689 | 616141.3000 | 452805.4000 | 708.2000 | 1866109.0000 |
| B_PHARM | 211,376 | 16235.6500 | 19105.7300 | 0.2389 | 66202.0000 |
| TD_BPHARM | 270,283 | -3210.1860 | 14102.5800 | -67899.7300 | 47548.0700 |
| TD_XPHARM | 270,239 | 8.2620 | 4.5362 | 0.0010 | 19.6955 |

| COMP (1981-2018) | | | | | |
|---|---|---|---|---|---|
| Variable | Sample size | Mean | Std. dev. | Min | Max |
| GV_PPP | 275,046 | 21432.7100 | 19334.5700 | 16.5214 | 84124.8000 |
| B_PPP | 274,203 | 129885.9000 | 115365.9000 | 24.0575 | 429134.4000 |
| H_PPP | 274,913 | 24064.1600 | 20798.5800 | 0.9419 | 74722.0000 |
| TP_RS | 266,437 | 681733.3000 | 449777.6000 | 815.1000 | 1866109.0000 |
| B_COMP | 207,172 | 33336.0400 | 24746.7300 | 0.5354 | 78575.0000 |
| TD_BCOMP | 278,922 | -24781.9600 | 77526.7600 | -212256.7000 | 186830.1000 |
| TD_XCOMP | 278,847 | 9.4136 | 7.3226 | 0.0001 | 31.2690 |

Table 8: Total number of patent filings and Breakthrough patents of OECD countries, ranked by number of patent filings

| PHARM | | | | COMP | | | |
|---|---|---|---|---|---|---|---|
| | Patent No. | Breakthrough No. | Breakthrough% | | Patent No. | Breakthrough No. | Breakthrough% |
| US | 111,930 | 478 | 0.427% | US | 109,172 | 702 | 0.643% |
| DE | 32,552 | 91 | 0.280% | JP | 56,127 | 181 | 0.322% |
| JP | 26,420 | 108 | 0.409% | DE | 24,368 | 22 | 0.090% |
| FR | 20,067 | 38 | 0.189% | FR | 16,414 | 32 | 0.195% |
| GB | 17,373 | 44 | 0.253% | KR | 13,773 | 53 | 0.385% |
| CH | 14,518 | 46 | 0.317% | NL | 11,630 | 24 | 0.206% |
| NL | 7,189 | 34 | 0.473% | GB | 8,657 | 20 | 0.231% |
| IT | 6,774 | 24 | 0.354% | CN | 7,771 | 4 | 0.051% |
| CA | 5,531 | 27 | 0.488% | CA | 5,563 | 14 | 0.252% |
| SE | 5,143 | 8 | 0.156% | SE | 4,764 | 10 | 0.210% |
| BE | 4,040 | 16 | 0.396% | FI | 4,544 | 48 | 1.056% |
| DK | 3,673 | 28 | 0.762% | CH | 4,338 | 11 | 0.254% |
| IL | 3,669 | 3 | 0.082% | TW | 2,842 | 3 | 0.106% |
| KR | 3,465 | 2 | 0.058% | IT | 2,681 | 1 | 0.037% |
| ES | 3,182 | 5 | 0.157% | IL | 2,466 | 20 | 0.811% |
| AU | 3,112 | 11 | 0.353% | AU | 1,510 | 3 | 0.199% |
| CN | 3,038 | 1 | 0.033% | BE | 1,363 | 6 | 0.440% |
| IN | 2,717 | 3 | 0.110% | IE | 1,043 | 0 | 0.000% |
| AT | 2,671 | 11 | 0.412% | AT | 915 | 1 | 0.109% |
| IE | 1,550 | 2 | 0.129% | SG | 812 | 0 | 0.000% |
| NO | 1,048 | 3 | 0.286% | IN | 751 | 0 | 0.000% |
| TW | 922 | 0 | 0.000% | DK | 750 | 3 | 0.400% |
| FI | 896 | 0 | 0.000% | ES | 693 | 7 | 1.010% |
| LU | 770 | 11 | 1.429% | NO | 520 | 1 | 0.192% |
| HU | 744 | 0 | 0.000% | RU | 311 | 0 | 0.000% |
| RU | 582 | 6 | 1.031% | LU | 294 | 0 | 0.000% |
| SG | 547 | 0 | 0.000% | TR | 259 | 0 | 0.000% |
| TR | 538 | 0 | 0.000% | HU | 93 | 0 | 0.000% |
| SI | 451 | 0 | 0.000% | PT | 78 | 0 | 0.000% |
| CZ | 311 | 0 | 0.000% | CZ | 53 | 0 | 0.000% |
| PT | 238 | 1 | 0.420% | SI | 40 | 0 | 0.000% |

Notes: ISO 3166 alpha-2 country codes are used in the table. The full definition can be accessed at: https://www.iso.org/iso-3166-country-codes.html.

Table 9: PHARM Poisson regression with forward citation number as dependent variable

| | Dependent Variable: Number of forward citations in 5 years | | | | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| N | 278,990 | 277,922 | 275,003 | 278,990 | 277,922 | 275,003 |
| chi2 | 187475.3677 | 186656.9167 | 184983.6661 | 194220.4737 | 193484.2034 | 191697.0658 |
| IRII11 | -0.2266*** (0.0100) | | | -0.2158*** (0.0100) | | |
| NTR11 | -0.7798*** (0.0196) | | | -0.7437*** (0.0196) | | |
| IRII31 | | -0.0215* (0.0101) | | | -0.0109 (0.0102) | |
| NTR31 | | -0.9671*** (0.0301) | | | -0.9284*** (0.0300) | |
| IRII51 | | | 0.0159 (0.0093) | | | 0.0175 (0.0093) |
| NTR51 | | | -0.8608*** (0.0359) | | | -0.8315*** (0.0358) |
| patent scope | 0.1547*** (0.0011) | 0.1631*** (0.0011) | 0.1633*** (0.0011) | 0.1468*** (0.0011) | 0.1552*** (0.0011) | 0.1552*** (0.0011) |
| family size | 0.0485*** (0.0002) | 0.0484*** (0.0002) | 0.0486*** (0.0002) | 0.0491*** (0.0002) | 0.0491*** (0.0002) | 0.0492*** (0.0002) |
| backward citations | 0.0044*** (0.0000) | 0.0044*** (0.0000) | 0.0044*** (0.0000) | 0.0043*** (0.0000) | 0.0044*** (0.0000) | 0.0044*** (0.0000) |
| APPC_US | | | | 0.2489*** (0.0041) | 0.2500*** (0.0042) | 0.2476*** (0.0042) |
| APPC_DE | | | | 0.1089*** (0.0060) | 0.1061*** (0.0060) | 0.1038*** (0.0060) |
| APPC_JP | | | | 0.2711*** (0.0063) | 0.2704*** (0.0064) | 0.2690*** (0.0064) |
| APPC_FR | | | | -0.0943*** (0.0077) | -0.0967*** (0.0078) | -0.0949*** (0.0078) |
| APPC_GB | | | | -0.1254*** (0.0078) | -0.1297*** (0.0078) | -0.1391*** (0.0079) |
| cons | -0.6145*** (0.0279) | -0.6033*** (0.0224) | -0.4334*** (0.0189) | -0.7251*** (0.0281) | -0.7148*** (0.0226) | -0.5418*** (0.0191) |

Notes:
Sample of patents filed from 1980 to 2018. Year fixed effects are included in all the estimations.
Standard errors in parentheses. * p<0.05, ** p<0.01, *** p<0.001
APPC_US is a dummy variable defined to be 1 when at least one of the applicants of the patent is registered with address in the United States. DE: Germany, JP: Japan, FR: France, GB: United Kingdom.

Table 10: PHARM Probit regression with Breakthrough probability as dependent variable

| | Dependent Variable: Breakthrough | | | | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| N | 278,671 | 277,603 | 274,684 | 278,671 | 277,603 | 274,684 |
| chi2 | 1007.1574 | 998.8442 | 994.3050 | 1053.9714 | 1044.6836 | 1038.4280 |
| IRII11 | -0.2202** | | | -0.2163** | | |
| | (0.0692) | | | (0.0695) | | |
| NTR11 | -0.1625 | | | -0.1509 | | |
| | (0.1232) | | | (0.1233) | | |
| IRII31 | | -0.0415 | | | -0.0341 | |
| | | (0.0693) | | | (0.0697) | |
| NTR31 | | -0.0111 | | | 0.0036 | |
| | | (0.1751) | | | (0.1751) | |
| IRII51 | | | -0.0050 | | | -0.0015 |
| | | | (0.0635) | | | (0.0639) |
| NTR51 | | | 0.1829 | | | 0.1932 |
| | | | (0.1963) | | | (0.1962) |
| patent scope | 0.0837*** | 0.0903*** | 0.0910*** | 0.0807*** | 0.0875*** | 0.0881*** |
| | (0.0074) | (0.0072) | (0.0072) | (0.0075) | (0.0073) | (0.0072) |
| family size | 0.0261*** | 0.0261*** | 0.0261*** | 0.0267*** | 0.0266*** | 0.0267*** |
| | (0.0011) | (0.0011) | (0.0011) | (0.0011) | (0.0011) | (0.0011) |
| backward citations | 0.0028*** | 0.0028*** | 0.0028*** | 0.0027*** | 0.0027*** | 0.0027*** |
| | (0.0002) | (0.0002) | (0.0002) | (0.0002) | (0.0002) | (0.0002) |
| APPC_US | | | | 0.0801** | 0.0784** | 0.0746** |
| | | | | (0.0280) | (0.0281) | (0.0281) |
| APPC_DE | | | | -0.0056 | -0.0059 | -0.0142 |
| | | | | (0.0423) | (0.0423) | (0.0427) |
| APPC_JP | | | | 0.1790*** | 0.1762*** | 0.1745*** |
| | | | | (0.0409) | (0.0410) | (0.0412) |
| APPC_FR | | | | -0.1434* | -0.1448* | -0.1428* |
| | | | | (0.0582) | (0.0583) | (0.0583) |
| APPC_GB | | | | -0.1050 | -0.1051 | -0.1024 |
| | | | | (0.0556) | (0.0556) | (0.0557) |
| cons | -3.2768*** | -3.3682*** | -3.1660*** | -3.3149*** | -3.3999*** | -3.2036*** |
| | (0.1993) | (0.1670) | (0.1250) | (0.2024) | (0.1685) | (0.1273) |

Notes:
Sample of patents filed from 1980 to 2018. Year fixed effects are included in all the estimations. Standard errors in parentheses. * p<0.05, ** p<0.01, *** p<0.001
APPC_US is a dummy variable defined to be 1 when at least one of the applicants of the patent is registered with an address in the United States. DE: Germany, JP: Japan, FR: France, GB: United Kingdom.

Table 11: PHARM Poisson regression with forward citation number as dependent variable, MSTI variables included in estimation

| | Dependent Variable: Number of forward citations in 5 years | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| N | 255,645 | 246,077 | 201,350 | 198,111 | 198,111 | 198,111 | 198,111 | 198,111 | 198,111 |
| chi2 | 180567.4021 | 176845.7183 | 156651.2926 | 156023.7005 | 155508.8998 | 155193.0922 | 157881.0363 | 157389.2689 | 157089.0649 |
| IRII11 | -0.1654*** (0.0104) | -0.1820*** (0.0106) | -0.2098*** (0.0118) | -0.2253*** (0.0119) | | | -0.2258*** (0.0119) | | |
| NTR11 | -0.7851*** (0.0207) | -0.7909*** (0.0213) | -0.8439*** (0.0249) | -0.8664*** (0.0252) | | | -0.8553*** (0.0251) | | |
| IRII31 | | | | | -0.0524*** (0.0121) | | | -0.0555*** (0.0121) | |
| NTR31 | | | | | -1.1548*** (0.0395) | | | -1.1497*** (0.0395) | |
| IRII51 | | | | | | 0.0807*** (0.0111) | | | 0.0897*** (0.0111) |
| NTR51 | | | | | | -1.0197*** (0.0458) | | | -1.0111*** (0.0458) |
| patent scope | 0.1569*** (0.0011) | 0.1552*** (0.0012) | 0.1563*** (0.0013) | 0.1562*** (0.0013) | 0.1626*** (0.0013) | 0.1658*** (0.0012) | 0.1551*** (0.0013) | 0.1614*** (0.0013) | 0.1651*** (0.0012) |
| family size | 0.0485*** (0.0002) | 0.0488*** (0.0002) | 0.0506*** (0.0002) | 0.0511*** (0.0002) | 0.0509*** (0.0002) | 0.0509*** (0.0002) | 0.0510*** (0.0002) | 0.0508*** (0.0002) | 0.0508*** (0.0002) |
| backward citations | 0.0043*** (0.0000) | 0.0043*** (0.0000) | 0.0042*** (0.0000) | 0.0042*** (0.0000) | 0.0042*** (0.0000) | 0.0042*** (0.0000) | 0.0042*** (0.0000) | 0.0042*** (0.0000) | 0.0042*** (0.0000) |
| GV_PPP | -0.0051 (0.0069) | -0.0557*** (0.0073) | 0.0460*** (0.0093) | 0.0138 (0.0094) | 0.0119 (0.0094) | 0.0122 (0.0094) | 0.3111*** (0.0157) | 0.3101*** (0.0156) | 0.3105*** (0.0156) |
| B_PPP | 0.0038** (0.0015) | -0.0103*** (0.0018) | -0.0135*** (0.0017) | -0.0612*** (0.0022) | -0.0602*** (0.0022) | -0.0599*** (0.0022) | -0.0801*** (0.0025) | -0.0794*** (0.0025) | -0.0793*** (0.0025) |
| H_PPP | 0.0468*** (0.0064) | 0.1000*** (0.0070) | -0.0206* (0.0099) | -0.0992*** (0.0101) | -0.0954*** (0.0101) | -0.0961*** (0.0101) | -0.1881*** (0.0116) | -0.1856*** (0.0115) | -0.1865*** (0.0115) |
| TP_RS | | 0.0030*** (0.0002) | | 0.0106*** (0.0003) | 0.0105*** (0.0003) | 0.0104*** (0.0003) | 0.0019*** (0.0005) | 0.0017** (0.0005) | 0.0017** (0.0005) |
| B_PHARM | | | 0.1717*** (0.0059) | 0.2884*** (0.0069) | 0.2862*** (0.0069) | 0.2869*** (0.0069) | 0.2717*** (0.0078) | 0.2705*** (0.0078) | 0.2715*** (0.0078) |
| TD_BPHARM | | | 0.0616*** (0.0049) | 0.0236*** (0.0052) | 0.0249*** (0.0052) | 0.0249*** (0.0052) | -0.0028 (0.0052) | -0.0021 (0.0052) | -0.0023 (0.0052) |
| TD_XPHARM | | | -0.0111*** (0.0008) | 0.0032*** (0.0009) | 0.0029** (0.0009) | 0.0027** (0.0009) | 0.0046*** (0.0014) | 0.0047*** (0.0014) | 0.0046*** (0.0014) |
| APPC_US | | | | | | | 0.5510*** (0.0173) | 0.5551*** (0.0173) | 0.5577*** (0.0173) |
| APPC_DE | | | | | | | 0.1344*** (0.0132) | 0.1305*** (0.0132) | 0.1291*** (0.0132) |
| APPC_JP | | | | | | | 0.5666*** (0.0184) | 0.5703*** (0.0184) | 0.5718*** (0.0184) |
| APPC_FR | | | | | | | -0.2007*** (0.0176) | -0.2026*** (0.0176) | -0.2053*** (0.0176) |
| APPC_GB | | | | | | | -0.2127*** (0.0269) | -0.2132*** (0.0269) | -0.2147*** (0.0269) |
| cons | -0.4845*** (0.0231) | -0.4723*** (0.0240) | -0.1390*** (0.0213) | -0.4757*** (0.0240) | -0.6100*** (0.0247) | -0.7187*** (0.0238) | -0.7478*** (0.0264) | -0.8847*** (0.0271) | -1.0047*** (0.0262) |

Notes:
Sample of patents filed from 1981 to 2018 are available for estimation 1 and 2. Patents filed from 1987 to 2018 are available for estimation 3 and 4. Year fixed effects are included in all the estimations. All the MSTI variables except TD_XPHARM have been divided by 10,000 from their original values.
Standard errors in parentheses. * p<0.05, ** p<0.01, *** p<0.001
APPC_US is a dummy variable defined to be 1 when at least one of the applicants of the patent is registered with address in the United States. DE: Germany, JP: Japan, FR: France, GB: United Kingdom.

Table 12: PHARM Probit regression with Breakthrough probability as dependent variable, MSTI variables included in estimation

| | Dependent Variable: Breakthrough | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| N | 255,372 | 245,857 | 201,301 | 198,062 | 198,062 | 198,062 | 198,062 | 198,062 | 198,062 |
| chi2 | 1006.0039 | 983.2936 | 873.9383 | 863.3163 | 859.6833 | 861.6608 | 892.2427 | 888.5972 | 890.7000 |
| IRII11 | -0.1358 (0.0730) | -0.1350 (0.0738) | -0.1493 (0.0826) | -0.1575 (0.0832) | | | -0.1583 (0.0834) | | |
| NTR11 | -0.1330 (0.1298) | -0.1653 (0.1344) | -0.0481 (0.1481) | -0.0980 (0.1529) | | | -0.1017 (0.1533) | | |
| IRII31 | | | | | -0.0177 (0.0833) | | | -0.0224 (0.0835) | |
| NTR31 | | | | | 0.0083 (0.2202) | | | -0.0014 (0.2210) | |
| IRII51 | | | | | | 0.1018 (0.0771) | | | 0.1076 (0.0772) |
| NTR51 | | | | | | 0.1719 (0.2474) | | | 0.1625 (0.2482) |
| patent scope | 0.0909*** (0.0077) | 0.0869*** (0.0079) | 0.0873*** (0.0088) | 0.0830*** (0.0090) | 0.0884*** (0.0088) | 0.0923*** (0.0086) | 0.0829*** (0.0090) | 0.0882*** (0.0088) | 0.0925*** (0.0086) |
| family size | 0.0261*** (0.0012) | 0.0263*** (0.0012) | 0.0268*** (0.0013) | 0.0270*** (0.0013) | 0.0268*** (0.0013) | 0.0268*** (0.0013) | 0.0272*** (0.0014) | 0.0270*** (0.0014) | 0.0270*** (0.0013) |
| backward citations | 0.0027*** (0.0002) | 0.0027*** (0.0002) | 0.0026*** (0.0002) | 0.0026*** (0.0002) | 0.0026*** (0.0002) | 0.0026*** (0.0002) | 0.0027*** (0.0002) | 0.0027*** (0.0002) | 0.0027*** (0.0002) |
| GV_PPP | -0.1485** (0.0497) | -0.1902*** (0.0524) | -0.1200 (0.0746) | -0.1573* (0.0769) | -0.1566* (0.0767) | -0.1539* (0.0766) | 0.0743 (0.1256) | 0.0733 (0.1254) | 0.0754 (0.1253) |
| B_PPP | 0.0081 (0.0102) | 0.0012 (0.0121) | -0.0051 (0.0118) | -0.0397* (0.0158) | -0.0393* (0.0158) | -0.0397* (0.0158) | -0.0439* (0.0177) | -0.0435* (0.0177) | -0.0440* (0.0177) |
| H_PPP | 0.1337** (0.0513) | 0.1739** (0.0561) | 0.0728 (0.0814) | 0.0234 (0.0845) | 0.0251 (0.0844) | 0.0231 (0.0843) | 0.0164 (0.0935) | 0.0187 (0.0934) | 0.0173 (0.0933) |
| TP_RS | | 0.0015 (0.0014) | | 0.0078** (0.0024) | 0.0077** (0.0024) | 0.0077** (0.0024) | 0.0002 (0.0042) | 0.0001 (0.0042) | 0.0001 (0.0042) |
| B_PHARM | | | 0.1335** (0.0415) | 0.2252*** (0.0508) | 0.2238*** (0.0508) | 0.2260*** (0.0509) | 0.1602** (0.0579) | 0.1590** (0.0579) | 0.1613** (0.0580) |
| TD_BPHARM | | | 0.0000 (0.0342) | -0.0262 (0.0359) | -0.0258 (0.0359) | -0.0263 (0.0359) | -0.0369 (0.0352) | -0.0366 (0.0352) | -0.0371 (0.0352) |
| TD_XPHARM | | | -0.0077 (0.0054) | 0.0041 (0.0065) | 0.0039 (0.0065) | 0.0038 (0.0065) | 0.0025 (0.0098) | 0.0025 (0.0098) | 0.0023 (0.0098) |
| APPC_US | | | | | | | 0.2114 (0.1261) | 0.2129 (0.1259) | 0.2140 (0.1259) |
| APPC_DE | | | | | | | 0.1253 (0.0944) | 0.1237 (0.0943) | 0.1236 (0.0943) |
| APPC_JP | | | | | | | 0.3626** (0.1357) | 0.3633** (0.1356) | 0.3635** (0.1357) |
| APPC_FR | | | | | | | -0.5317* (0.2238) | -0.5318* (0.2237) | -0.5328* (0.2235) |
| APPC_GB | | | | | | | -0.3991 (0.2859) | -0.4022 (0.2862) | -0.4056 (0.2866) |
| cons | -3.2728*** (0.1709) | -3.2317*** (0.1739) | -2.9729*** (0.1537) | -3.2034*** (0.1729) | -3.3062*** (0.1777) | -3.3993*** (0.1718) | -3.2960*** (0.1851) | -3.3966*** (0.1895) | -3.4985*** (0.1839) |

Notes:
Sample of patents filed from 1981 to 2018 are available for estimation 1 and 2. Patents filed from 1987 to 2018 are available for estimation 3 and 4. Year fixed effects are included in all the estimations. All the MSTI variables except TD_XPHARM have been divided by 10,000 from their original values.
Standard errors in parentheses. * p<0.05, ** p<0.01, *** p<0.001
APPC_US is a dummy variable defined to be 1 when at least one of the applicants of the patent is registered with address in the United States. DE: Germany, JP: Japan, FR: France, GB: United Kingdom.

Table 13: COMP Poisson regression with forward citation number as dependent variable

| | Dependent Variable: Number of forward citations in 5 years | | | | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| N | 282,506 | 282,506 | 280,401 | 282,506 | 282,506 | 280,401 |
| chi2 | 132464.9563 | 132671.9469 | 132530.1564 | 139922.1502 | 140139.5781 | 139914.0625 |
| IRII11 | -0.2576*** (0.0090) | | | -0.2426*** (0.0091) | | |
| NTR11 | -0.2156*** (0.0133) | | | -0.1720*** (0.0133) | | |
| IRII31 | | -0.2887*** (0.0089) | | | -0.2750*** (0.0090) | |
| NTR31 | | -0.2445*** (0.0177) | | | -0.1999*** (0.0177) | |
| IRII51 | | | -0.3140*** (0.0090) | | | -0.2956*** (0.0091) |
| NTR51 | | | -0.1734*** (0.0199) | | | -0.1299*** (0.0199) |
| patent scope | 0.1400*** (0.0010) | 0.1390*** (0.0010) | 0.1363*** (0.0010) | 0.1419*** (0.0010) | 0.1410*** (0.0010) | 0.1386*** (0.0010) |
| family size | 0.0848*** (0.0004) | 0.0849*** (0.0004) | 0.0849*** (0.0004) | 0.0846*** (0.0004) | 0.0847*** (0.0004) | 0.0847*** (0.0004) |
| backward citations | 0.0085*** (0.0001) | 0.0085*** (0.0001) | 0.0085*** (0.0001) | 0.0081*** (0.0001) | 0.0081*** (0.0001) | 0.0081*** (0.0001) |
| APPC_US | | | | 0.3254*** (0.0054) | 0.3265*** (0.0054) | 0.3273*** (0.0054) |
| APPC_JP | | | | 0.2375*** (0.0063) | 0.2368*** (0.0063) | 0.2371*** (0.0063) |
| APPC_DE | | | | -0.0877*** (0.0092) | -0.0854*** (0.0092) | -0.0830*** (0.0092) |
| APPC_FR | | | | -0.1406*** (0.0106) | -0.1381*** (0.0106) | -0.1345*** (0.0106) |
| APPC_KR | | | | 0.4469*** (0.0100) | 0.4508*** (0.0100) | 0.4525*** (0.0100) |
| cons | -0.4062*** (0.0308) | -0.3890*** (0.0307) | -0.2348*** (0.0249) | -0.6244*** (0.0312) | -0.6043*** (0.0311) | -0.4716*** (0.0255) |

Notes:
Sample of patents filed from 1981 to 2018. Year fixed effects are included in all the estimations.
Standard errors in parentheses. * p<0.05, ** p<0.01, *** p<0.001
APPC_US is a dummy variable defined to be 1 when at least one of the applicants of the patent is registered with address in the United States. JP: Japan, DE: Germany, FR: France, KR: South Korea.

Table 14: COMP Probit regression with Breakthrough probability as dependent variable

| | Dependent Variable: Breakthrough | | | | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| N | 282,506 | 282,506 | 280,401 | 282,506 | 282,506 | 280,401 |
| chi2 | 1618.3394 | 1625.9499 | 1606.7991 | 1811.4562 | 1819.0067 | 1803.0888 |
| IRII11 | -0.1885*** (0.0508) | | | -0.1969*** (0.0515) | | |
| NTR11 | -0.0418 (0.0690) | | | -0.0226 (0.0702) | | |
| IRII31 | | -0.2300*** (0.0502) | | | -0.2372*** (0.0509) | |
| NTR31 | | -0.0116 (0.0889) | | | 0.0052 (0.0905) | |
| IRII51 | | | -0.2872*** (0.0515) | | | -0.2932*** (0.0522) |
| NTR51 | | | -0.0250 (0.1031) | | | -0.0092 (0.1049) |
| patent scope | 0.1050*** (0.0061) | 0.1032*** (0.0059) | 0.1002*** (0.0060) | 0.1054*** (0.0061) | 0.1039*** (0.0060) | 0.1008*** (0.0061) |
| family size | 0.0574*** (0.0022) | 0.0575*** (0.0022) | 0.0578*** (0.0023) | 0.0558*** (0.0023) | 0.0559*** (0.0023) | 0.0562*** (0.0023) |
| backward citations | 0.0036*** (0.0007) | 0.0036*** (0.0007) | 0.0036*** (0.0007) | 0.0031*** (0.0007) | 0.0032*** (0.0007) | 0.0031*** (0.0007) |
| APPC_US | | | | 0.2525*** (0.0298) | 0.2529*** (0.0299) | 0.2385*** (0.0314) |
| APPC_JP | | | | 0.0669 (0.0369) | 0.0658 (0.0369) | 0.0443 (0.0384) |
| APPC_DE | | | | -0.3299*** (0.0714) | -0.3278*** (0.0714) | -0.3557*** (0.0747) |
| APPC_FR | | | | -0.1521* (0.0672) | -0.1502* (0.0672) | -0.1673* (0.0688) |
| APPC_KR | | | | 0.2117*** (0.0547) | 0.2156*** (0.0547) | 0.2014*** (0.0554) |
| cons | -2.6174*** (0.1180) | -2.5919*** (0.1172) | -2.8923*** (0.1381) | -2.7131*** (0.1223) | -2.6870*** (0.1213) | -2.9883*** (0.1418) |

Notes:
Sample of patents filed from 1981 to 2018. Year fixed effects are included in all the estimations.
Standard errors in parentheses. * p<0.05, ** p<0.01, *** p<0.001
APPC_US is a dummy variable defined to be 1 when at least one of the applicants of the patent is registered with address in the United States. JP: Japan, DE: Germany, FR: France, KR: South Korea.

Table 15: COMP Poisson regression with forward citation number as dependent variable, MSTI variables included in estimation

| | Dependent Variable: Number of forward citations in 5 years | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| N | 273,644 | 266,242 | 203,710 | 199,354 | 199,354 | 199,354 | 199,354 | 199,354 | 199,354 |
| chi2 | 132265.4454 | 129887.0765 | 108154.1924 | 106825.7658 | 106837.6564 | 106914.1630 | 107489.6526 | 107497.1431 | 107574.9565 |
| IRII11 | -0.2426*** (0.0092) | -0.2367*** (0.0093) | -0.2181*** (0.0110) | -0.2188*** (0.0112) | | | -0.2139*** (0.0112) | | |
| NTR11 | -0.1921*** (0.0135) | -0.1764*** (0.0138) | -0.0726*** (0.0163) | -0.0694*** (0.0166) | | | -0.0525** (0.0166) | | |
| IRII31 | | | | | -0.2181*** (0.0111) | | | -0.2112*** (0.0111) | |
| NTR31 | | | | | -0.0222 (0.0219) | | | -0.0018 (0.0219) | |
| IRII51 | | | | | | -0.2337*** (0.0112) | | | -0.2261*** (0.0112) |
| NTR51 | | | | | | 0.0709** (0.0242) | | | 0.0924*** (0.0242) |
| patent scope | 0.1441*** (0.0010) | 0.1427*** (0.0010) | 0.1543*** (0.0012) | 0.1548*** (0.0012) | 0.1551*** (0.0012) | 0.1529*** (0.0012) | 0.1543*** (0.0012) | 0.1548*** (0.0012) | 0.1526*** (0.0012) |
| family size | 0.0847*** (0.0004) | 0.0865*** (0.0004) | 0.0858*** (0.0005) | 0.0863*** (0.0005) | 0.0863*** (0.0005) | 0.0861*** (0.0005) | 0.0856*** (0.0005) | 0.0856*** (0.0005) | 0.0854*** (0.0005) |
| backward citations | 0.0079*** (0.0001) | 0.0080*** (0.0001) | 0.0080*** (0.0001) | 0.0080*** (0.0001) | 0.0080*** (0.0001) | 0.0080*** (0.0001) | 0.0079*** (0.0001) | 0.0080*** (0.0001) | 0.0079*** (0.0001) |
| GV_PPP | -0.0566*** (0.0061) | -0.0881*** (0.0061) | 0.3357*** (0.0118) | 0.2195*** (0.0133) | 0.2195*** (0.0133) | 0.2187*** (0.0133) | 0.1480*** (0.0195) | 0.1471*** (0.0195) | 0.1477*** (0.0195) |
| B_PPP | 0.0019 (0.0014) | -0.0312*** (0.0018) | -0.1109*** (0.0028) | -0.0999*** (0.0033) | -0.1002*** (0.0033) | -0.0999*** (0.0033) | -0.0692*** (0.0040) | -0.0697*** (0.0040) | -0.0695*** (0.0040) |
| H_PPP | 0.1043*** (0.0053) | 0.1787*** (0.0057) | -0.1780*** (0.0156) | 0.0333 (0.0196) | 0.0341 (0.0196) | 0.0327 (0.0196) | 0.0886*** (0.0251) | 0.0876*** (0.0251) | 0.0846*** (0.0251) |
| TP_RS | | 0.0063*** (0.0002) | | -0.0065*** (0.0006) | -0.0064*** (0.0006) | -0.0065*** (0.0006) | -0.0052*** (0.0008) | -0.0050*** (0.0008) | -0.0050*** (0.0008) |
| B_PHARM | | | 0.3574*** (0.0068) | 0.3503*** (0.0071) | 0.3518*** (0.0071) | 0.3520*** (0.0071) | 0.1602*** (0.0106) | 0.1627*** (0.0106) | 0.1628*** (0.0106) |
| TD_BPHARM | | | -0.0396*** (0.0017) | -0.0236*** (0.0021) | -0.0232*** (0.0021) | -0.0234*** (0.0021) | -0.0206*** (0.0031) | -0.0205*** (0.0031) | -0.0209*** (0.0031) |
| TD_XPHARM | | | 0.0033* (0.0013) | 0.0275*** (0.0019) | 0.0272*** (0.0019) | 0.0274*** (0.0019) | 0.0272*** (0.0028) | 0.0267*** (0.0028) | 0.0269*** (0.0028) |
| APPC_US | | | | | | | 0.2107*** (0.0343) | 0.2091*** (0.0343) | 0.2098*** (0.0343) |
| APPC_JP | | | | | | | 0.0515* (0.0243) | 0.0486* (0.0243) | 0.0503* (0.0243) |
| APPC_DE | | | | | | | -0.1095*** (0.0155) | -0.1070*** (0.0155) | -0.1056*** (0.0155) |
| APPC_FR | | | | | | | -0.1443*** (0.0189) | -0.1439*** (0.0189) | -0.1442*** (0.0189) |
| APPC_KR | | | | | | | 0.3368*** (0.0170) | 0.3365*** (0.0170) | 0.3382*** (0.0170) |
| cons | -0.4442*** (0.0310) | -0.5767*** (0.0315) | -1.0423*** (0.0364) | -1.1752*** (0.0384) | -1.1742*** (0.0383) | -1.1663*** (0.0383) | -1.1382*** (0.0421) | -1.1392*** (0.0421) | -1.1329*** (0.0420) |

Notes:
Sample of patents filed from 1981 to 2018 are available for estimation 1 and 2. Patents filed from 1987 to 2018 are available for estimation 3 and 4. Year fixed effects are included in all the estimations. All the MSTI variables except TD_XCOMP have been divided by 10,000 from their original values.
Standard errors in parentheses. * p<0.05, ** p<0.01, *** p<0.001
APPC_US is a dummy variable defined to be 1 when at least one of the applicants of the patent is registered with address in the United States. JP: Japan, DE: Germany, FR: France, KR: South Korea.

Table 16: COMP Probit regression with Breakthrough probability as dependent variable, MSTI variables included in estimation

| | Dependent Variable: Breakthrough | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| N | 273,644 | 266,242 | 203,710 | 199,354 | 199,354 | 199,354 | 199,354 | 199,354 | 199,354 |
| chi2 | 1664.1605 | 1673.3111 | 1551.1983 | 1545.3356 | 1548.4264 | 1546.7180 | 1586.6443 | 1589.2683 | 1587.6486 |
| IRII11 | -0.1987*** | -0.2217*** | -0.2544*** | -0.2630*** | | | -0.2624*** | | |
| | (0.0521) | (0.0535) | (0.0617) | (0.0634) | | | (0.0637) | | |
| NTR11 | -0.0285 | -0.0313 | 0.0127 | 0.0345 | | | 0.0536 | | |
| | (0.0705) | (0.0729) | (0.0840) | (0.0859) | | | (0.0863) | | |
| IRII31 | | | | | -0.2778*** | | | -0.2744*** | |
| | | | | | (0.0624) | | | (0.0627) | |
| NTR31 | | | | | 0.0873 | | | 0.1099 | |
| | | | | | (0.1108) | | | (0.1114) | |
| IRII51 | | | | | | -0.2714*** | | | -0.2689*** |
| | | | | | | (0.0634) | | | (0.0637) |
| NTR51 | | | | | | 0.1148 | | | 0.1373 |
| | | | | | | (0.1233) | | | (0.1240) |
| patent scope | 0.1038*** | 0.1032*** | 0.1140*** | 0.1148*** | 0.1152*** | 0.1153*** | 0.1134*** | 0.1141*** | 0.1143*** |
| | (0.0063) | (0.0064) | (0.0074) | (0.0076) | (0.0074) | (0.0075) | (0.0077) | (0.0075) | (0.0075) |
| family size | 0.0559*** | 0.0573*** | 0.0582*** | 0.0591*** | 0.0593*** | 0.0592*** | 0.0583*** | 0.0586*** | 0.0584*** |
| | (0.0023) | (0.0024) | (0.0027) | (0.0028) | (0.0028) | (0.0028) | (0.0028) | (0.0028) | (0.0028) |
| backward citations | 0.0030*** | 0.0030*** | 0.0028*** | 0.0027*** | 0.0028*** | 0.0027*** | 0.0027** | 0.0028*** | 0.0027*** |
| | (0.0007) | (0.0007) | (0.0008) | (0.0008) | (0.0008) | (0.0008) | (0.0008) | (0.0008) | (0.0008) |
| GV_PPP | 0.0549 | 0.0203 | 0.4137*** | 0.3825*** | 0.3833*** | 0.3809*** | 0.2416* | 0.2404* | 0.2399* |
| | (0.0332) | (0.0344) | (0.0668) | (0.0745) | (0.0745) | (0.0745) | (0.1063) | (0.1063) | (0.1063) |
| B_PPP | -0.0157* | -0.0278** | -0.0896*** | -0.1206*** | -0.1215*** | -0.1208*** | -0.0771*** | -0.0781*** | -0.0777*** |
| | (0.0079) | (0.0100) | (0.0148) | (0.0182) | (0.0182) | (0.0182) | (0.0212) | (0.0212) | (0.0212) |
| H_PPP | 0.0984*** | 0.1326*** | -0.2291** | -0.2465* | -0.2444* | -0.2449* | -0.0416 | -0.0423 | -0.0433 |
| | (0.0281) | (0.0315) | (0.0883) | (0.1170) | (0.1168) | (0.1168) | (0.1468) | (0.1465) | (0.1466) |
| TP_RS | | 0.0035** | | 0.0069* | 0.0070* | 0.0069* | 0.0062 | 0.0064 | 0.0064 |
| | | (0.0012) | | (0.0031) | (0.0031) | (0.0031) | (0.0042) | (0.0041) | (0.0041) |
| B_DRUG | | | 0.3096*** | 0.3566*** | 0.3577*** | 0.3575*** | 0.0926 | 0.0949 | 0.0948 |
| | | | (0.0408) | (0.0461) | (0.0462) | (0.0462) | (0.0637) | (0.0637) | (0.0637) |
| TD_BDRUG | | | -0.0301** | -0.0389** | -0.0386** | -0.0385** | -0.0124 | -0.0123 | -0.0123 |
| | | | (0.0099) | (0.0126) | (0.0126) | (0.0126) | (0.0181) | (0.0181) | (0.0181) |
| TD_XDRUG | | | -0.0266*** | -0.0258* | -0.0260* | -0.0257* | -0.0279 | -0.0283 | -0.0282 |
| | | | (0.0070) | (0.0112) | (0.0112) | (0.0112) | (0.0164) | (0.0164) | (0.0164) |
| APPC_US | | | | | | | 0.3581 | 0.3575 | 0.3603 |
| | | | | | | | (0.1909) | (0.1912) | (0.1910) |
| APPC_JP | | | | | | | 0.0062 | 0.0013 | 0.0041 |
| | | | | | | | (0.1329) | (0.1328) | (0.1326) |
| APPC_DE | | | | | | | -0.4076*** | -0.4022*** | -0.4016*** |
| | | | | | | | (0.1212) | (0.1212) | (0.1214) |
| APPC_FR | | | | | | | -0.2136 | -0.2137 | -0.2161 |
| | | | | | | | (0.1217) | (0.1215) | (0.1219) |
| APPC_KR | | | | | | | 0.3106** | 0.3101** | 0.3107** |
| | | | | | | | (0.0989) | (0.0989) | (0.0987) |
| cons | -2.6391*** | -2.6945*** | -3.1309*** | -3.4032*** | -3.3869*** | -3.3959*** | -3.2521*** | -3.2398*** | -3.2482*** |
| | (0.1189) | (0.1232) | (0.1997) | (0.2149) | (0.2145) | (0.2144) | (0.2373) | (0.2368) | (0.2366) |

Notes:
Sample of patents filed from 1981 to 2018 are available for estimation 1 and 2. Patents filed from 1987 to 2018 are available for estimation 3 and 4. Year fixed effects are included in all the estimations. All the MSTI variables except TD_XCOMP have been divided by 10,000 from their original values.
Standard errors in parentheses. * p<0.05, ** p<0.01, *** p<0.001
APPC_US is a dummy variable defined to be 1 when at least one of the applicants of the patent is registered with address in the United States. JP: Japan, DE: Germany, FR: France, KR: South Korea.

**Supplement A:**
**OECD patent quality indicator variables summary statistics excluding observations with zero value *Patent Scope***

***PHARM***

| Variable | Obs | Mean | Std. dev. | Min | Max |
|---|---|---|---|---|---|
| patent sco | 278,989 | 3.0289 | 1.4706 | 1 | 21 |
| family size | 278,989 | 9.8403 | 7.6757 | 1 | 57 |
| backward | 278,989 | 7.8111 | 20.9827 | 0 | 1013 |
| forward ci | 278,989 | 1.3189 | 4.9999 | 0 | 672 |
| originality | 266,615 | 0.7955 | 0.1626 | 0 | 0.9863 |
| generality | 122,562 | 0.5056 | 0.2247 | 0 | 0.9388 |

***COMP***

| Variable | Obs | Mean | Std. dev. | Min | Max |
|---|---|---|---|---|---|
| patent sco | 282,502 | 2.096527 | 1.276035 | 1 | 30 |
| family size | 282,502 | 4.795729 | 2.87282 | 1 | 45 |
| backward | 282,502 | 4.542987 | 6.568957 | 0 | 498 |
| forward ci | 282,502 | 0.944185 | 3.107796 | 0 | 270 |
| originality | 264,874 | 0.679839 | 0.227161 | 0 | 0.9823 |
| generality | 103,342 | 0.352898 | 0.280268 | 0 | 0.9378 |

**Supplement B:**
**Variance, Skewness and Kurtosis of OECD patent quality indicator variables**

***PHARM***

| Variable | Obs | Variance | Skewness | Kurtosis |
|---|---|---|---|---|
| patent scor | 278,990 | 2.1625 | 1.4206 | 6.4531 |
| family size | 278,990 | 58.9164 | 1.6361 | 5.8959 |
| backward c | 278,990 | 440.2706 | 19.0553 | 637.4957 |
| forward cit | 278,990 | 24.9991 | 37.1322 | 2935.5600 |
| originality | 266,616 | 0.0264 | -2.4886 | 11.1026 |
| generality | 122,563 | 0.0505 | -0.9662 | 3.1497 |

***COMP***

| Variable | Obs | Variance | Skewness | Kurtosis |
|---|---|---|---|---|
| patent scor | 282,506 | 1.6283 | 2.5068 | 24.0196 |
| family size | 282,506 | 8.2530 | 2.6833 | 15.8604 |
| backward c | 282,506 | 43.1506 | 46.9561 | 3175.0810 |
| forward cit | 282,506 | 9.6583 | 22.8177 | 1067.0120 |
| originality | 264,878 | 0.0516 | -1.5783 | 5.1308 |
| generality | 103,344 | 0.0786 | -0.1152 | 1.5451 |