Research Policy 54 (2025) 105295



Contents lists available at ScienceDirect

Research Policy

journal homepage: www.elsevier.com/locate/respol



Text vs. citations: A comparative analysis of breakthrough and disruption metrics in patent innovation

Alex J. Yang *

School of Information Management, Nanjing University, Nanjing, China
Center for Labor and a Just Economy, Harvard University, Cambridge, MA, USA

ARTICLE INFO

JEL classification:
I23
O31
O33
O38

Keywords:
Patent
Breakthrough
Disruption
Team
Natural language processing

ABSTRACT

This study examines two dynamic metrics for assessing technological innovation— the text-based breakthrough index (KI index) and the citation-based disruption index (CD index)—both of which integrate ex-ante (novelty) and ex-post (impact) information. The KI index identifies breakthrough inventions by measuring their novelty (low similarity to prior patents) and impact (high similarity to future patents), whereas the CD index quantifies technological disruption by analyzing shifts in citation patterns. Using a dataset of over six million patents filed with the USPTO between 1980 and 2017, this paper finds that KI and CD indices are highly correlated and both effectively capture technological breakthroughs. Patents with high KI or CD scores typically originate from original and narrowly focused knowledge bases. However, the two indices exhibit distinct patterns: (1) the KI index fluctuates with economic cycles, while the CD index has experienced a steady decline over time; (2) the KI index positively correlates with future patent citation impact, whereas the CD index follows a U-shaped relationship with patent citation impact; and (3) small and remote teams produce higher KI but lower CD scores, potentially because larger teams cite newer, widely recognized references. I discuss innovation concepts—breakthroughs, disruptions, and beyond—to contextualize these findings and explore their implications for understanding technological advancement. These results contribute to the discourse on measuring innovation and underscore the complementary strengths of text-based and citation-based approaches in assessing technological progress.

1. Introduction

Technological breakthroughs and disruptions are pivotal drivers of economic growth and industrial transformation, introducing novel ideas, reshaping knowledge landscapes, and influencing future innovations. Breakthroughs represent inventions that combine significant originality with widespread future impact, while disruptions fundamentally alter existing technological paradigms by displacing prior knowledge. Accurately measuring these phenomena is crucial for policymakers, industry leaders, and researchers aiming to foster and understand technological progress (Fortunato et al., 2018; Higham et al., 2021). However, dynamically assessing how a patent contributes to knowledge evolution remains challenging, requiring metrics that integrate ex-ante novelty (originality at filing) and ex-post impact (subsequent influence).

Conventional metrics for assessing innovation typically focus on either novelty or impact, each with inherent limitations. Ex-ante measures, such as atypical knowledge combinations, novel terminology, or interdisciplinarity based on backward citations, capture a patent's originality at the time of filing (Arts et al., 2025; Azoulay et al., 2011; Di Bona et al., 2025; Fleming, 2001; Foster et al., 2015). In contrast, ex-post metrics, like forward citation counts or economic value estimates, gauge a patent's influence after filing (Hall and Harhoff, 2012; Kogan et al., 2017). For instance, forward citations reflect how frequently a patent is referenced by later inventions, while economic value assessments quantify market significance. However, these approaches often fail to bridge novelty and impact, limiting their ability to identify true breakthroughs that combine originality with transformative influence (Higham et al., 2021). This gap highlights the need for metrics that dynamically connect a patent's origins and its future contributions.

The text-based breakthrough index (KI index), proposed by Kelly et al. (2021), addresses this need by measuring breakthrough inventions through textual analysis. The KI index evaluates a patent's novelty (low textual similarity to prior patents) and impact (high textual similarity to

* School of Information Management, Nanjing University, Nanjing 210023, China.
E-mail address: alexjieyang@outlook.com.

https://doi.org/10.1016/j.respol.2025.105295
Received 22 October 2024; Received in revised form 24 June 2025; Accepted 24 June 2025
0048-7333/© 2025 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

future patents) using natural language processing (NLP). By leveraging global patent data within specific time windows, it offers a citation-independent perspective, capturing breakthroughs that set new technological agendas. In contrast, the citation-based disruption index (CD index), introduced by Funk and Owen-Smith (2017), quantifies disruption by analyzing how a patent alters citation patterns, displacing prior technologies in subsequent citations. Relying on local citation networks, the CD index has revealed trends such as the higher disruptiveness of small, localized teams and a decline in disruptive innovation over time (Lin et al., 2023; Park et al., 2023; Wu et al., 2019). Both indices integrate ex-ante novelty and ex-post impact, linking a patent's predecessors and successors, though other related measures, such as textual factor-based (Cong et al., 2024) and knowledge entity-based (Wang et al., 2023) approaches, exist. This study focuses on comparing the KI and CD indices to elucidate their distinct roles in measuring breakthroughs and disruptions.

This paper provides a comprehensive comparison between the text-based KI index and the citation-based CD index, using a dataset of over six million patents filed between 1980 and 2017. The findings reveal that both indices effectively identify transformative technological breakthroughs, with high-scoring patents often emerging from original and narrowly focused knowledge bases. However, key differences emerge: (1) the KI index fluctuates with economic cycles, while the CD index shows a steady decline; (2) the KI index positively correlates with future patent citation impact, whereas the CD index exhibits a U-shaped relationship; and (3) small and remote teams tend to produce patents with higher KI scores but lower CD scores, possibly due to larger teams citing newer, more popular references. These findings offer valuable implications for research and policy. Researchers benefit from understanding the complementary strengths of text-based and citation-based metrics, enabling better tool selection for innovation studies. Policymakers gain insights into how economic conditions and team dynamics shape breakthroughs, guiding strategies to promote transformative technologies through tailored support for diverse team structures or during economic fluctuations.

The structure of this paper is similar to Higham et al. (2021), which compares different technological innovation measures at the patent level, but the content and implications are distinct. First, the two innovative metrics analyzed in this paper incorporate both ex-ante (novelty) and ex-post (impact) information. While their rationale and definitions are similar, they employ fundamentally different methodologies—one based on complex networks and the other on natural language processing (NLP) techniques. Second, the KI and CD indices are relatively new, and their patterns and implications remain underexplored. Third, this paper contributes to recent debates surrounding disruptiveness in science and technology (Leibel and Bornmann, 2024; Macher et al., 2024). The discussion of the CD index has sparked widespread interest, yet no comparable text-based metric has been available for comparison until now. By bridging this gap, this paper offers valuable insights into the multidimensional nature of innovation and patent impact.

The remainder of this paper is structured as follows. Section 2 provides theoretical background on citation-based and text-based innovation metrics and describes the dataset. Section 3 details the methodologies used to compute the KI and CD indices, along with the regression models employed. Section 4 presents the main findings, comparing the dynamic patterns of the two indices. Finally, Section 5 concludes with a discussion of the broader implications for understanding technological innovation.

2. Background

2.1. Technological innovation

The literature presents various definitions related to innovation—such as disruption, breakthroughs, creativity, and novelty (Azoulay et al., 2011; Fortunato et al., 2018; Foster et al., 2015; Hall and Jaffe, 2018)—though a comprehensive review of this extensive body of work is beyond the scope of this discussion. Instead, this paper adopts a pragmatic approach, focusing on empirical data derived primarily from patents (Arts et al., 2021; Higham et al., 2021).

Traditional methods for measuring technological innovation in the context of patents often operate from an ex-ante perspective, neglecting the focal patents' potential future impact. Such approaches are typically associated with the concept of novelty rather than disruption. Theoretical perspectives emphasize the recombination of existing knowledge as a primary driver of breakthrough innovation (Fleming and Sorenson, 2001; Fontana et al., 2020; Uzzi et al., 2013). Innovation frequently arises from the integration of ideas or knowledge across disparate fields, with cross-sectoral knowledge flows playing a critical role in this process (Gates et al., 2019). The innovation process has often been described as an exploration of combinatorial possibilities, from which new insights and breakthroughs emerge. Prior studies have underscored the importance of recombination in linking innovation with substantial scientific and technological advancements. Azoulay et al. (2011) characterize creativity as a cumulative and interactive process of recombining existing knowledge, while Uzzi et al. (2013) quantified the relative commonness of journal pairs cited in a given paper, using the lowest 10th percentile of commonness scores to represent novelty and median scores to reflect conventionality.

However, identifying novel patents through such methodologies does not necessarily indicate that they represent breakthrough inventions. For an invention to be genuinely transformative, it must be adopted and utilized within its field, analogous to how a biological mutation must reproduce to exert evolutionary influence (Mokyr, 1990). Not all combinations of distant technologies result in mainstream innovations, and there is frequently skepticism surrounding novelty. Arthur (2007) contended that truly radical technologies emerge from pre-existing building blocks, alongside a profound theoretical understanding of how to manipulate these technologies for specific applications. Accordingly, disruptive innovations must incorporate novel combinations of existing technologies in ways that are both relevant and applicable within their domain. Foster et al. (2015) observed that innovative publications tend to exert a greater impact compared to more conservative works, though they are also subject to a higher risk of failure. Wang et al. (2017) highlighted that highly novel articles are more likely to be published in journals with lower impact factors, suggesting that unconventional combinations often struggle to gain immediate acceptance within the academic mainstream. While novel ideas inherently involve risk, many fail to result in major breakthroughs, as new combinations alone may be insufficient for genuine disruption (Sun et al., 2021). Therefore, ex-post dimensions, such as the future impact of patents, should also be considered when attempting to quantify breakthrough inventions.

2.2. Text-based breakthrough measures

As outlined above, measuring transformative innovation requires an examination of both ex-ante novelty and ex-post impact of a given item. Consequently, quantifying breakthrough innovation in science and technology solely on the basis of textual data presents challenges, particularly due to the difficulty in assessing ex-post impact without relying on citation data. However, recent advancements in natural language processing (NLP) and large-scale full-text data analysis have led to the development of methods aimed at quantifying innovation at the textual level.

Hofstra et al. (2020) employ NLP techniques such as phrase extraction and structural topic modeling to identify terms representing substantive concepts across millions of documents in the ProQuest Dissertations database. By summing the number of novel conceptual co-occurrences in each thesis, they create a measure of conceptual novelty. They also track how frequently a thesis's new conceptual linkages are

incorporated into subsequent documents over time. Similarly, Sun et al. (2021) apply topic modeling to a corpus comprising the full text of patents within a specific technological field. The topics identified define a "technological space," from which the authors develop continuous measures of a patent's relevance—based on its similarity to the mainstream technological trajectory within this space over time. Their method enables the identification of potential breakthrough inventions that may influence the rate and direction of technological development in the focal field. Crucially, their framework calculates not only backward similarity (relevance to past work) but also forward dynamic similarity (relevance to future developments).

This paper aims to compare text-based breakthrough measures with citation-based disruption measures. One of the primary measures examined is the recent breakthrough metric (KI index) introduced by Kelly et al. (2021). Building on the work of Kogan et al. (2017), Kelly and colleagues construct a comprehensive full-text patent dataset, covering nearly all patents issued by the U.S. Patent and Trademark Office (USPTO). They identify breakthrough patents by analyzing the textual similarity of a given patent to both preceding and subsequent work, focusing on patents that are distinct from prior work yet related to subsequent innovations. Notably, this method does not rely on citation data and encompasses the global textual knowledge represented in the USPTO patent database.

The KI index measures a patent's breakthrough potential by assessing its textual novelty (low similarity to prior patents) and impact (high similarity to future patents). Using a point-in-time Term Frequency-Inverse Document Frequency (TF-IDF) approach, it computes pairwise textual similarities between a focal patent and other patents. The KI index is defined as the logarithmic ratio of summed textual similarities with forward patents to those with backward patents, where higher values indicate greater innovation.

This approach assumes that patent text accurately reflects technological content and that textual similarity captures meaningful relationships. However, a key limitation of the KI index is its reduced effectiveness over long-time windows (Kelly et al., 2021), where the ratio of signal to noise may decline due to the sparsity of relevant future patents, weakening the informativeness of textual similarity. Despite this, the KI index captures four core dimensions: novelty (textual divergence from prior work), impact (alignment with future innovations), global scope (considering all USPTO patents), and dynamic evolution (linking past and future knowledge). It provides a global view of technological trends, is independent of citation data, and demonstrates strong predictive power (Cong et al., 2024; Kelly et al., 2021).

2.3. Citation-based disruption measures

Technological disruption is often closely associated with the concept of technological breakthroughs, as both entail sudden shifts that alter the prevailing status quo (Bower and Christensen, 1995; Kuhn, 1962). The number of forward citations is frequently used to estimate a patent's future technological impact, as citation data is widely accessible, making citation-based disruption measures relatively straightforward to compute. However, one significant limitation of citation-based metrics is that citation information tends to be local rather than global. In many cases, knowledge is not accurately acknowledged by authors or patent examiners, and citations serve as an imperfect proxy for knowledge flows. This is particularly problematic due to the considerable heterogeneity among different teams, sections, and industries, which may introduce biases in citation analysis. Although some studies combine ex-ante novelty measures with ex-post citation counts to assess breakthroughs (Esposito, 2023), this paper does not focus on such hybrid approaches. Instead, the emphasis is placed on the citation-based disruption measure known as the CD index, which, despite its relatively recent introduction, has gained widespread usage, including in three research articles published in *Nature* (Lin et al., 2023; Park et al., 2023; Wu et al., 2019).

The CD index, introduced by Funk and Owen-Smith (2017), offers a structural analysis of the deep citation network, providing a systematic framework for assessing the disruptive potential of academic papers. This metric quantifies the extent to which a paper aligns with established knowledge while also evaluating its capacity to induce paradigm shifts, thereby offering a novel approach to assessing scholarly innovation. Funk and Owen-Smith (2017) argue that the CD index captures four key dimensions: (i) the degree to which future inventions build upon prior technologies, (ii) variations in how inventions modify the utilization of their predecessors over time, (iii) the range from large-scale transformations to incremental shifts in technological consolidation and destabilization, and (iv) the distinction between consolidating and destabilizing technologies that may produce similar levels of impact.

The CD index has been extensively validated and applied across diverse domains, demonstrating its flexibility and robustness to variations in citation practices. Wu et al. (2019) find that Nobel Prize-winning papers have an average CD index of approximately 0.1, ranking in the top 2 % of Web of Science papers. The CD index has been used to identify disruptive innovations in fields such as colorectal surgery, plastic and craniofacial surgery, radiology, pediatric medicine, and cancer research, often aligning with practitioner judgments of clinical significance (Becerra et al., 2025; Grunvald et al., 2021; Horen et al., 2021). It has also been applied to track disruption in case law (Lee et al., 2024), AI development (Jurowetzki et al., 2025), materials science, and art history (Shinichi and Matsui, 2024), reflecting its adaptability to disciplinary norms and temporal changes in citation behavior. These studies incorporate robustness checks, normalization strategies, and alternative specifications to address variations in citation practices across fields and over time (Lin et al., 2023; Park et al., 2023; Wu et al., 2019; Yang and Deng, 2024; Yang et al., 2025b).

Despite its strengths, the CD index has notable limitations. It assumes that a patent's knowledge foundation is fully captured by its cited references, yet foundational knowledge may be unacknowledged or inaccurately cited in practice. Additionally, citation behaviors vary across technological fields (Nielsen and Andersen, 2021), complicating the index's interpretation. Validity concerns, such as those raised by Bornmann et al. (2020), question the CD index's ability to identify expert-selected breakthroughs. Furthermore, measurement and truncation bias concerns, as highlighted by Macher et al. (2024), suggest that the index may be skewed by incomplete or time-limited citation data. To address these issues, researchers have proposed alternative disruption measures (Deng et al., 2025; Leydesdorff and Bornmann, 2021; Wang et al., 2023; Yang et al., 2023a; Yang et al., 2024a), though these also depend on citation data. Complementing citation-based measures like the CD index with non-citation-based approaches, such as the text-based KI index, can provide a more comprehensive understanding of technological innovation.

2.4. Patent data description

This study utilizes full-text and citation data provided by the United States Patent and Trademark Office (USPTO) and Patents View, a patent data-sharing initiative supported by the Office of the Chief Economist at the USPTO. The USPTO patent dataset comprises 7,566,869 patents and 117,308,691 citations between 1976 and 2022, representing the curated digital patent application data. Notably, each patent is associated with two key dates: the filing (or publication) date and the issue date. For the purposes of this study, text-based and citation-based innovation measures are calculated based on the filing date. The Inventor Protection Act of 1999 mandates that U.S. patent applications be published 18 months after the initial filing.

The USPTO categorizes patents into several types, including 'utility,' 'reissue,' 'design,' 'TVPP,' 'plant,' 'statutory invention registration,' and 'defensive publication.' Consistent with prior research (Park et al., 2023; Wu et al., 2019), this analysis focuses exclusively on utility patents due

to their higher novelty and more stringent issuance and publication standards. I do not differentiate between citations added by inventors and those added by patent examiners (Higham et al., 2021), although this will be a very promising direction for further analysis.

Text-based and citation-based innovation measures are calculated using a five-year window (Macher et al., 2024; Park et al., 2023), covering 6,384,413 patents filed between 1980 and 2017. However, the calculation of the CD index results in a reduced sample size, retaining only 3,036,683 patents. This reduction occurs because the CD index requires each patent to have at least one backward citation and one forward citation.

Consistent with existing research (Lin et al., 2023; Park et al., 2023; Wu et al., 2019) and given the observational nature of this study, no statistical methods were applied to predetermine the sample size. The regression dataset is structured as a panel by filing year, CPC section, and inventor, meaning patents appear multiple times if classified under multiple sections or if multiple inventors are listed. As a result, the regression panel dataset contains 25,037,733 observations.

3. Methods

3.1. Text-based innovation (the KI index)

The text-based innovation measure draws on the breakthrough measure proposed by Kelly et al. (2021). Specifically, they utilize large-scale full-text patent data from the USPTO dataset to construct pairwise textual similarity or distance between patents. These similarities are computed using a retrospective, or point-in-time, version of the TF-IDF (Term Frequency-Inverse Document Frequency) method.

As illustrated in Fig. 1, to calculate text-level disruption measures, Kelly et al. (2021) apply natural language processing (NLP) techniques to preprocess the textual data of patents. Using the Python NLTK package, they parse the abstract, claims, and description sections of each patent into individual terms. In this process, stop words and non-words that provide little semantic value are removed, and all capital letters are converted to lowercase. To enhance the signal-to-noise ratio in the textual analysis, the dictionary is further reduced by excluding terms that occur either extremely frequently or extremely infrequently; specifically, terms that appear in fewer than 20 documents are excluded. This process results in a dictionary of 1,685,416 terms. Using this comprehensive dictionary, a document-term matrix (DTM) is then constructed, denoted as C, where each matrix element c_{dw} represents the count of term w in patent d.

A point-in-time version of TF-IDF is then employed to calculate the textual similarity or distance between each pair of patents. The backward inverse document frequency of term w for filing year t, denoted as $BIDF_{wt}$, is defined as the logarithmic frequency of documents containing term w in patents filed prior to year t:

$$BIDF_{wt} = \log\left(\frac{N(patents)^{prior\ to\ t}}{N(documents)_w^{prior\ to\ t} + 1}\right).$$

For a patent pair $(i\,j)$, the point-in-time version of TF-IDF for each term w in patents i and j is constructed as follows:

$$\begin{cases} TFBIDF_{iwt} = TF_{iw}BIDF_{wt} \\ TFBIDF_{jwt} = TF_{jw}BIDF_{wt} \end{cases}, t = \min(t_i, t_j).$$

These values are arranged in a W-dimensional vector $TFBIDF_{it}$, where W represents the size of the union of terms for the pair $(i\,j)$. Each TFBIDF vector is then normalized to have unit length, i.e., $V_{it} = TFBIDF_{iwt}/\|TFBIDF_{iwt}\|$. The textual similarity between the two normalized vectors is then computed as:

$$Textual\ similarity_{ij} = V_{it}V_{jt}.$$

The underlying logic of the text-level innovation measure is that breakthrough patents are expected to exhibit significant textual distance from existing patents filed prior to the focal patent's filing year t, reflecting novelty and divergence from established technologies. Besides, breakthrough patents should also have relatively low distance from future patents filed after year t, as breakthrough innovations are

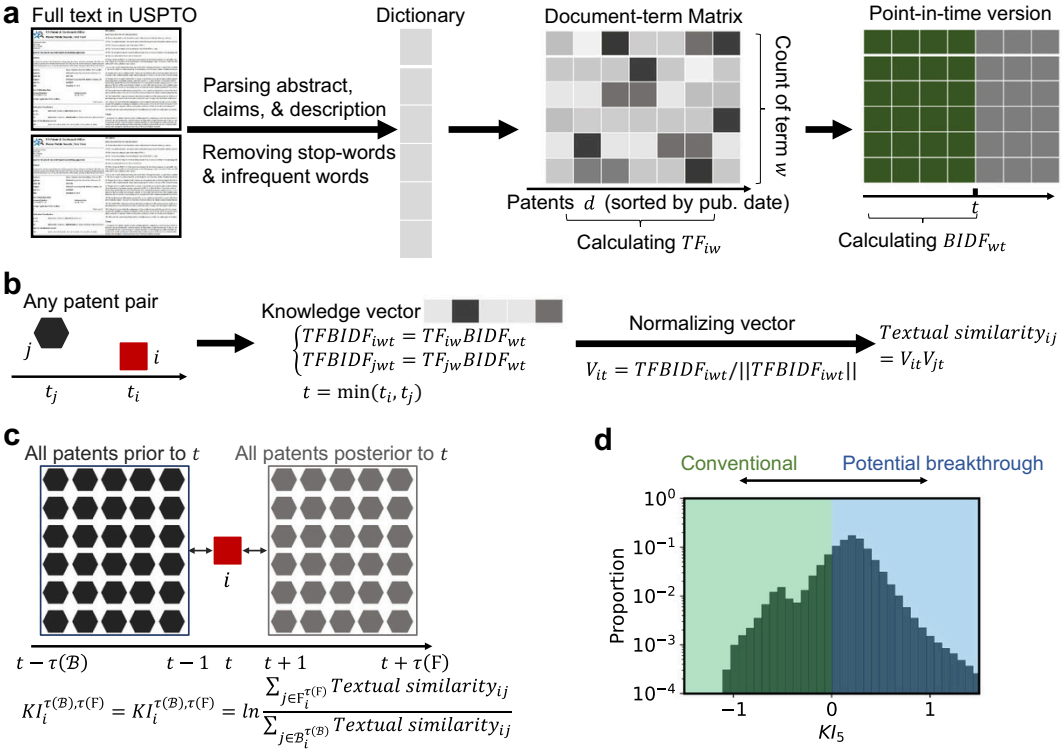


Fig. 1. Quantifying text-based innovation index (the KI index) in technological patents.

likely to shift paradigms and lead to new technological developments, thusly future technologies should be similar to the focal breakthrough.

It is important to note that a patent with a high backward distance signifies strong ex-ante novelty, though its impact on future innovations may not yet be visible. On the other hand, a patent exhibiting a low forward distance may reflect its central role in redefining the trajectory of future technological developments. Alternatively, it could signal its alignment with a larger shift in technology initiated by a predecessor. To measure innovation, both the backward distance (positively associated with innovation) and forward distance (negatively associated with innovation) are integrated.

The Kelly's Knowledge Innovation (KI) Index is thus defined as:

$$KI = ln\frac{\sum_{j\in F_i^{\tau(F)}} Textual\ similarity_{ij}}{\sum_{j\in \mathscr{B}_i^{\tau(\mathscr{B})}} Textual\ similarity_{ij}},$$

where the numerator sums the textual similarities between the focal patent and all future patents within the time window $[t-\tau(\mathscr{B}), t-1]$, while the denominator sums the textual similarities with all prior patents in $[t+1, t+\tau(F)]$. The ratio captures the balance between novelty and influence, providing a measure of innovation. To address potential bias due to fluctuations in annual patent filings, a robustness check is conducted by calculating the average textual similarity instead of the sum (detailed in the Supplementary Materials).

The primary measure, the KI_5 index (referred to as KI_5 hereafter for brevity), is based on five-year backward ($\tau(\mathscr{B}) = 5$) and forward ($\tau(F) = 5$) windows. Extending the method to longer time windows (e.g., 10 years) is not recommended, as the sparsity of relevant patents over extended periods may reduce the informativeness of textual similarity calculations, limiting meaningful insights.

3.2. Citation-based disruption (the CD index)

The citation-based disruption measure is relatively straightforward in its calculation, as it relies on human-curated citations linking patents, thereby eliminating the need to compute global pairwise information. Instead, the focus is on the local citation network of the focal patent—specifically, its references and citations, as well as the pairwise citations among them. To quantify technological disruption, I employ the widely-used CD index, a metric introduced by Funk and Owen-Smith (2017). This index, grounded in the structural characteristics of the local citation network, provides a systematic approach for assessing the disruptive potential of technological patents. By evaluating a patent's relationship to established knowledge and its capacity to initiate paradigm shifts, the CD index offers a valuable perspective for measuring innovation.

The underlying intuition is that a disruptive patent is less likely to have its predecessors cited by subsequent work that cites the focal patent; in other words, the foundational ideas of its predecessors become less relevant for future researchers. Conversely, a consolidating patent is more likely to have its predecessors cited by subsequent work, indicating that the knowledge on which the patent is built remains crucial for future research.

As illustrated in Fig. 2, the CD index is defined as:

$$CD_\tau = \frac{1}{n^\tau}\sum_{i=1}^{n}\frac{-2f_i^\tau b_i^\tau + f_i^\tau}{w_i^\tau}, \text{with}$$

$$f_i^\tau = \begin{cases} 1\ if\ i\ cites\ FP \\ 0\ otherwise \end{cases}, \text{and}$$

$$b_i^\tau = \begin{cases} 1\ if\ i\ cites\ any\ FP's\ predecessors \\ 0\ otherwise \end{cases}.$$

Here, τ denotes the forward citation window, and n^τ represents the sum of patents that cite: (1) only the focal patent (type f), (2) both the focal patent and parts of its predecessors (type b), or (3) only parts of its predecessors within τ years of the focal patent's publication. w_i^τ indexes a matrix of weights assigned to patents. A forward citation increases the consolidating (or disruptive) nature of the focal patent if the citing patent does (or does not) cite the predecessor patents that were also referenced by the focal patent. The CD index ranges from -1 (consolidating) to $+1$ (disruptive). A lower CD index suggests that a patent closely aligns with established knowledge, while a higher CD index signals a greater potential for the patent to disrupt existing technological paradigms. Although the CD index assumes that the cited references of a focal patent constitute its knowledge base—and can be influenced by the number of references—it remains widely adopted due to its relative simplicity in citation-based analyses.

Following Park et al. (2023), the forward citation window τ is set to five years, that is the CD_5 index (referred to as CD_5 hereafter for brevity). I calculate the CD index for patents that have at least one backward citation (i.e., a reference) and at least one forward citation, as the absence of either prevents the computation of a valid CD index. Leibel and Bornmann (2024) suggested using a higher threshold to obtain more meaningful results; therefore, in the Supplementary Materials, I also calculate the CD index for patents with at least five backward citations and five forward citations.

In addition to the CD index, Funk and Owen-Smith (2017) introduced the mCD index, which is the product of the CD index and forward

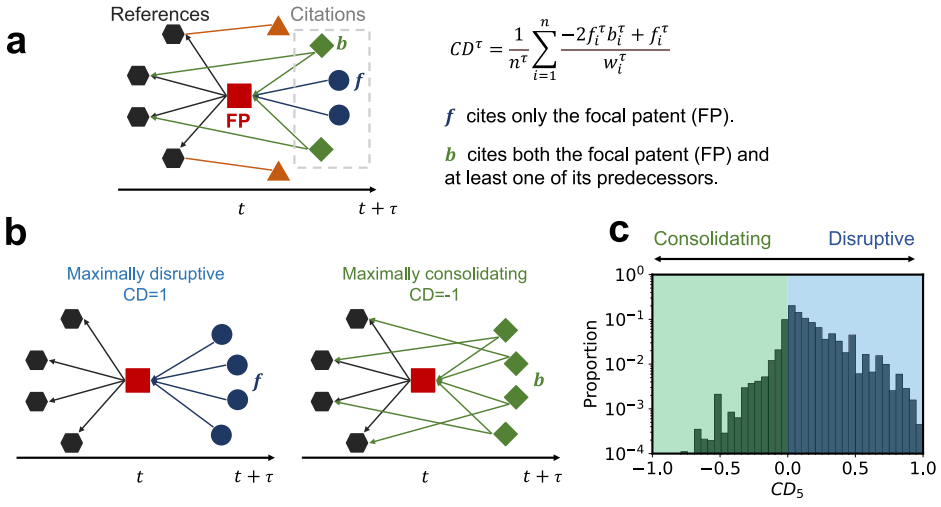


Fig. 2. Quantifying citation-based disruption index (the CD index) in technological patents.

citations. Although the mCD index is less commonly used, I analyze its patterns in Appendix C. The findings indicate that the mCD index might occupy a conceptual space that bridges the KI and CD indices. It shares the KI index's sensitivity to economic fluctuations, while also retaining the CD index's non-linear relationship with citation impact.

3.3. Related measures

To examine the implications of text-based and citation-based technological innovation metrics, I construct a set of related measures. Invention, as a fundamentally social activity (Jones et al., 2008; Wuchty et al., 2007), is influenced by the structure of the team that produces new technology, which may in turn affect the technology's significance. Building on existing research (Lin et al., 2023; Wu et al., 2019), I measure both the size of the inventor team and the collaboration distance for each patent. Team size is defined as the number of inventors associated with a given patent. In addition, following the methodology outlined by Lin et al. (2023), I calculate the collaboration distance among team members. The Patents View database provides name-disambiguated inventors and their respective geographical locations, including latitude and longitude coordinates associated with their reported addresses. Collaboration distance is thus measured as the average or median geographical distance between team members' locations.

To assess the technological impact of patents, I use the number of forward citations as a primary indicator. It is important to note, however, that while the CD index is a citation-based metric, it captures the proportion of disruptive citations rather than their absolute count. As a result, the CD index may not exhibit a positive correlation with the total number of forward citations. I utilize five-year and ten-year windows to quantify forward citations (Cit_5 and Cit_{10}). In addition, I calculate forward citations for each patent through the year 2022 to assess their long-term technological impact. To mitigate potential biases arising from variations in citation patterns across different years and technology sections, I construct an additional measure: the identification of "hits" or "home runs" (Mukherjee et al., 2017; Yang et al., 2024b). These are patents that rank in the top 10 % of citations (as of 2022) within their filing year and Cooperative Patent Classification (CPC) section. This approach allows for a more standardized assessment of patent impact, accounting for temporal and sectoral differences in citation behavior.

To assess the consistency of text-based and citation-based innovation metrics, I employ several commonly used originality measures (Foster et al., 2015; Leahey et al., 2023; Strumsky and Lobo, 2015; Wang et al., 2017) based on the CPC group, which is a fine-grained subclassification under the CPC system. Initially, I determine the first occurrence of each CPC group and CPC group pair (two-tuple). Subsequently, I calculate the average age of the CPC group. A patent that introduces a new CPC group or a novel combination of groups is considered more original compared to those that rely on established classifications. Conversely, if all CPC groups or combinations within a patent have been previously used (or old in terms of age), the patent may represent a refinement of existing technologies (Strumsky and Lobo, 2015). Additionally, I compute the Herfindahl-Hirschman Index (HHI) for the three-digit CPC subclass of each patent. A higher HHI suggests that the patent draws from a narrower knowledge base, while a lower HHI indicates a broader scope of knowledge.

To explain the distinctions between text-based and citation-based innovation metrics, I calculate reference age and reference popularity. Reference age is defined as the average time gap between the filing years of the patents cited (backward citations) and the focal patent. Reference popularity measures the average number of five-year forward citations received by the referenced patents.

3.4. Fixed effect regressions

I construct a regression dataset that includes more observations than unique patents. This dataset is structured as a panel by filing year and CPC section, with patents appearing multiple times if classified under multiple sections. It also incorporates inventor-level information using disambiguated author data from Patents View, duplicating patents when multiple inventors are listed. This structure allows for the inclusion of filing year, CPC section, and inventor fixed effects. By merging these dimensions, the dataset accounts for variations in patent quality across years and sections while mitigating potential selection biases—such as the Matthew effect, where highly innovative inventors are more likely to produce high-impact patents.

Given that both the KI index and CD index display non-standard distributions, as illustrated in Fig. 1 and Fig. 2, I follow Wu et al. (2019) to convert these indices into percentile data for analysis. When multiple patents share the same index value—a more common scenario with the CD index than the KI index—I assign them the median percentile value.

To examine the relationship between an innovation index (IND) and a patent's future technological impact or economic value, I employ ordinary least squares (OLS) regression models, formulated as:

$$ln(Forward\ citation_i + 1) = \alpha + \beta_1(IND_i) + Y_t(+S_s)(+A_a) + \epsilon_i,$$

where i represents the article, Y_t denotes year fixed effects, S_s captures section fixed effects, A_a captures author fixed effects, ϵ_i denotes the error term and α denotes the constant term. This approach rigorously assesses the relationship between innovation measures and patent impact.

To analyze the relationship between team size or collaboration distance and the innovation indices, I use OLS regression models, specified as:

$$IND_i = \alpha + \beta_1(Team\ size_i) + Y_t(+S_s)(+\mathrm{A}_a) + \epsilon_i.$$

Additionally, to explore the mechanisms driving the different patterns in the KI and CD indices, I incorporate interactions between reference age and reference popularity as follows:

$$IND_i = \alpha + \beta_1(Team\ size_i) + \beta_2(Ref\ age_i) + \beta_3(Team\ size_i \times Ref\ age_i) + Y_t(+S_s)(+A_a) + \epsilon_i.$$

4. Results

4.1. Consistency and effectiveness

The distributions of KI_5 and CD_5 are highly similar, both centered around 0, as shown in Fig. 1 and Fig. 2. KI_5 has an average value of 0.16 with a standard deviation of 0.28 (see Appendix A, Table A1 for the full descriptive statistics of all measures in the sample). A KI_5 value >0 indicates that a patent shares greater textual similarity with its descendants than with its antecedents, suggesting a higher degree of innovation. CD_5 ranges from −1 to 1, with an average value of 0.17 and a standard deviation of 0.21. A CD_5 value above 0 suggests that a patent has a higher tendency to disrupt rather than consolidate existing knowledge. Although the Pearson correlation between KI_5 and CD_5 at the patent level is low (0.01; see Appendix A, Fig. A1 for the full Pearson correlation matrix), patents with high CD_5 values often exhibit high KI_5 values, while those with low CD_5 values tend to have low KI_5 values (Fig. 3d). Moreover, at the year-section aggregated level, the average KI_5 and CD_5 are strongly positively correlated (Fig. 3e).

To assess whether KI_5 and CD_5 are effective in identifying technological breakthroughs, I present results based on a set of expert-selected technological milestones (Kelly et al., 2021), detailed in Appendix D, Table D1. The findings indicate that both the text-based and citation-based innovation indices successfully capture these breakthroughs. Specifically, all corresponding KI_5 values are above 0, and the majority of CD_5 values also exceed 0. Furthermore, the distributions of KI_5 and CD_5 for breakthrough patents are significantly higher than those for other patents (Fig. 3a–b). To provide a clearer comparison, Fig. 3c reports the average percentile ranks of KI_5 and CD_5 within these

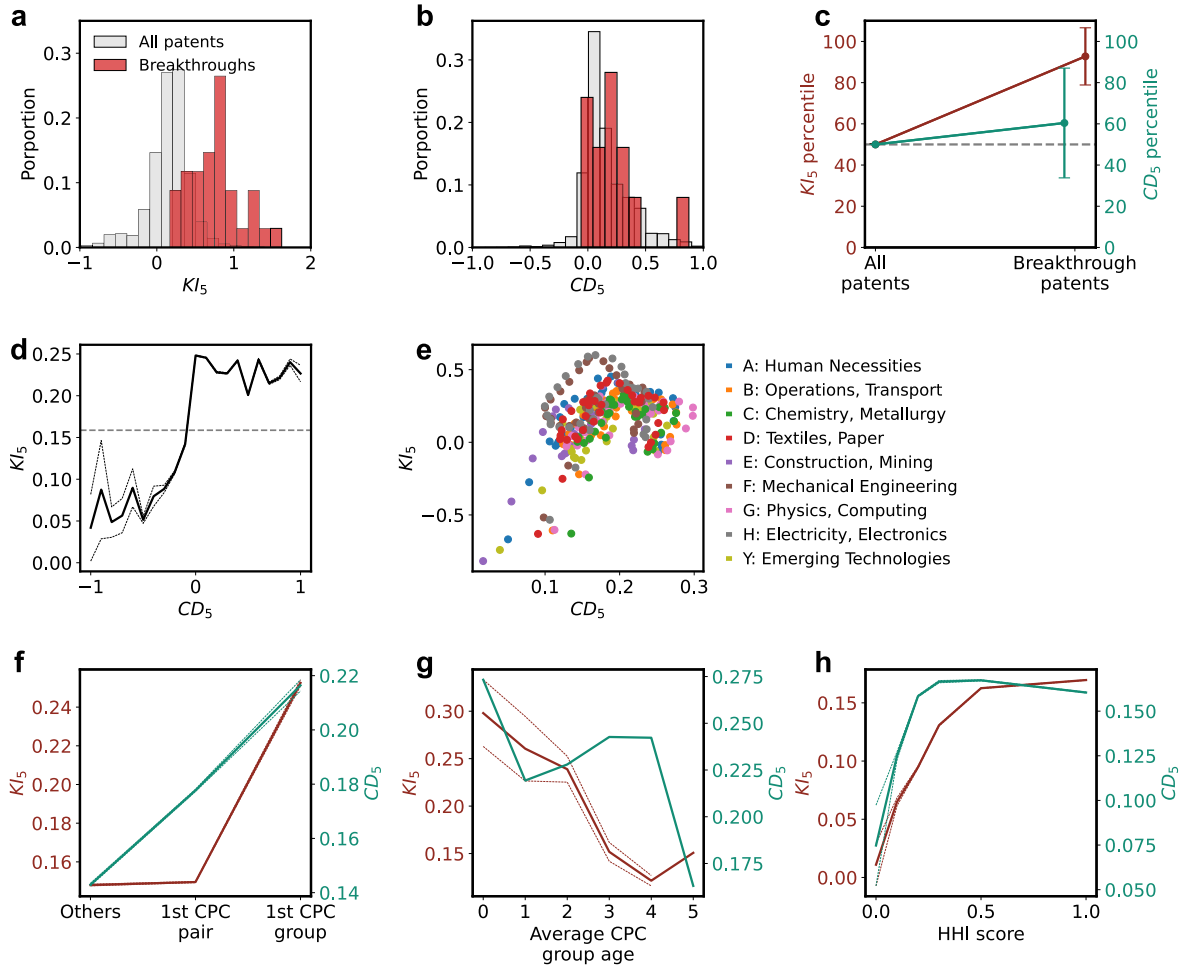


Fig. 3. Effectiveness and Consistency of the Text-Based KI Index and Citation-Based CD Index. (a–b) Distributions of KI_5 and CD_5 for all patents (grey) and expert-selected technological milestones/breakthroughs (red), illustrating the concentration of breakthrough patents in the upper tails of both distributions. (c) Average percentiles of KI_5 and CD_5 for all patents and expert-selected technological milestones, with standard deviations shown as error bars. (d) Binned line plot (bin width = 0.1) depicting the patent-level relationship between KI_5 and CD_5, indicating a positive association between high CD_5 and high KI_5. (e) Aggregated relationship between KI_5 and CD_5 at the section-year level, where each point represents the average KI_5 and CD_5 for a specific section-year, with colors denoting different CPC sections. (f) Comparison of average KI_5 and CD_5 across patents that introduce a new CPC group, a new CPC group pair, and all other patents, demonstrating that patents incorporating novel classification groups exhibit significantly higher KI_5 and CD_5 scores. (g) Relationship between CPC group age and KI_5/CD_5, showing that patents associated with older technological classifications tend to have lower KI_5 and CD_5 scores. (h) Relationship between KI_5, CD_5, and the Herfindahl-Hirschman Index (HHI) of the three-digit CPC subclass, indicating that both KI_5 and CD_5 are associated with patents that originate from more specialized or narrowly focused knowledge domains. Dotted lines indicate 95 % bootstrapped confidence intervals. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

breakthroughs. Breakthrough patents have an average KI_5 percentile of 92.8 ± 18.9 (std) and a CD_5 percentile of 60.4 ± 26.7 (std). These results indicate that both indices effectively identify technological breakthroughs. At least in this context, KI_5 appears to be a more reliable indicator of breakthroughs than CD_5.

Using CPC group-based ex-ante originality measures, I find that more original patents tend to have higher KI_5 and CD_5 scores. Specifically, patents that introduce a new (first) CPC group exhibit KI_5 scores approximately 0.1 points higher and CD_5 scores about 0.07 points higher than other patents (Fig. 3f). Additionally, as the average CPC group age increases—indicating an aging technological domain—both KI_5 and CD_5 decline significantly (Fig. 3g).

Furthermore, KI_5 and CD_5 are strongly correlated with the Herfindahl-Hirschman Index (HHI) of the three-digit CPC subclass for each patent (Fig. 3h). A higher HHI suggests that a patent is based on a more specialized knowledge base, whereas a lower HHI indicates broader knowledge integration. This relationship suggests that both KI_5 and CD_5 are closely linked to patents that draw from specialized rather than interdisciplinary knowledge.

Overall, KI_5 and CD_5 exhibit strong consistency across these features, indicating that they capture similar aspects of patent evaluation—specifically, technological breakthroughs that emerge from original and narrowly focused knowledge.

4.2. Temporal trends

Fig. 4 presents the temporal evolution of KI_5 and CD_5, revealing distinct trends over time. Contrary to expectations, these two metrics do not follow similar patterns. The average CD_5 has shown a clear decline since 1980, a finding consistent with Park et al. (2023).[1] Interestingly,

[1] In Appendix C and Supplementary Materials, I analyze the temporal trend of the mCD index (Funk and Owen-Smith, 2017) and examine the CD index with a higher threshold (requiring at least five backward and forward citations) (Leibel & Bornmann, 2024). These alternative measures do not exhibit a declining trend. Instead, they follow a pattern similar to that of KI_5, decreasing during economic expansions and increasing during recessions.

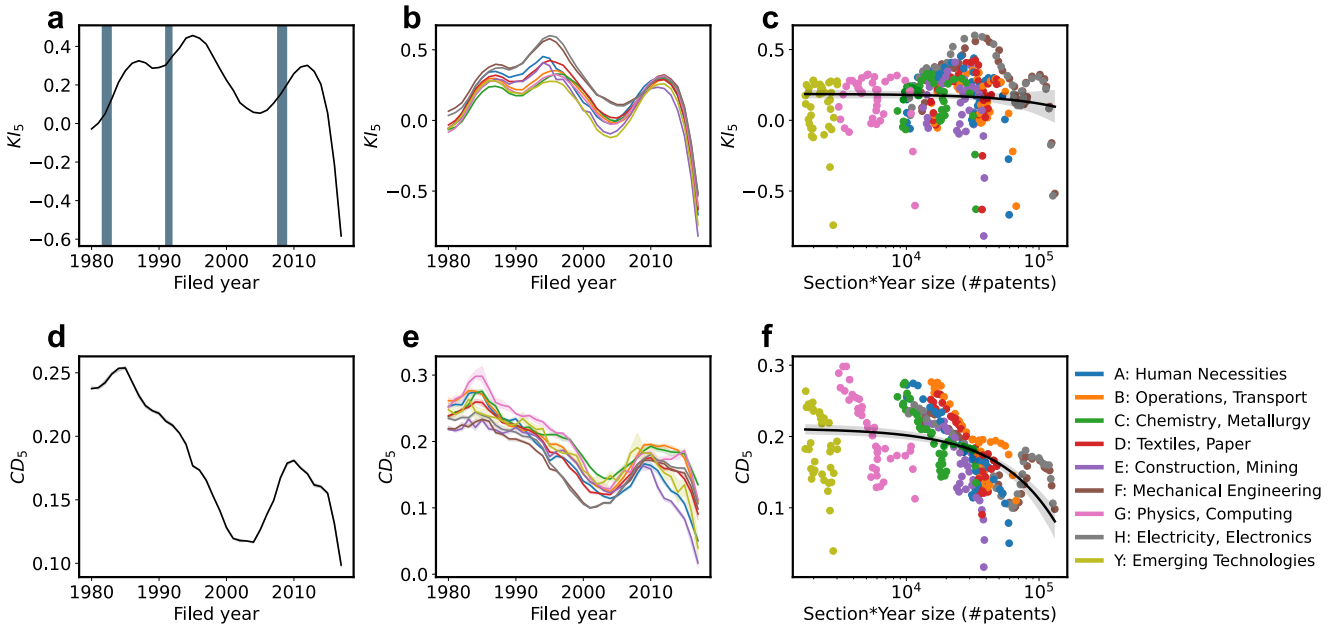


Fig. 4. Temporal Evolution of the Text-Based KI Index and Citation-Based CD Index. (a) Annual average KI_5 from 1980 to 2017 ($n = 6{,}384{,}413$), showing fluctuations that generally align with global economic cycles. Shaded blue regions indicate periods of economic recession as identified by the NBER. (b) Temporal trends in KI_5 remain consistent across different CPC sections. (c) KI_5 does not exhibit a significant correlation with section-year size. Each point represents the average KI_5 for a specific section-year, with colors denoting different CPC sections. (d) Annual average CD_5 from 1980 to 2017 ($n = 3{,}036{,}683$), displaying a long-term decline, with a notable upward deviation in the 2000s, potentially linked to changes in patent legislation. (e) Temporal trends in CD_5 remain stable across different CPC sections. (f) CD_5 shows a negative correlation with section-year size, suggesting that larger technological fields tend to have lower average CD_5 values. Shaded areas represent 95 % confidence intervals. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

CD_5 exhibits an upward trend in the 2000s, potentially linked to changes in U.S. patent law—specifically, the Inventor Protection Act of 1999, which required patent applications to be published 18 months after filing.

In contrast, the average KI_5 does not display a downward trend. Instead, it appears to fluctuate with the global economic cycle.[2] Periods of economic expansion are associated with a lower proportion of breakthrough technologies, while recessions see an increase in such innovations. One possible explanation aligns with Manso et al. (2023), who suggest that firms shift toward exploration during economic downturns and toward exploitation during periods of growth. As a result, during economic expansions, firms focus more on refining known technologies, leading to a decline in the average KI_5. Conversely, during recessions, firms pursue more explorative innovations, driving an increase in KI_5 (Manso et al., 2023). It is important to note that these trends reflect average values rather than a universal pattern, as Park et al. (2023) also find that the number of highly disruptive inventions has not declined over time.

Notably, the temporal trends of KI_5 and CD_5 remain consistent across different technological sections, as shown in Fig. 4b and e. In a recent study, Chu and Evans (2021) observed that in scientific fields, as the annual volume of publications increases, citations tend to concentrate disproportionately on already well-cited papers, reducing the likelihood that newly published works will disrupt existing research. This raises the question of whether larger section-years, as measured by the number of patents, tend to generate fewer disruptive innovations. To explore this, I analyze the relationship between section-year size and the average KI_5 and CD_5. The results, presented in Fig. 4c and f, reveal a clear divergence: KI_5 shows no significant correlation with section-year size, whereas CD_5 exhibits a negative correlation. This discrepancy may be due to the fundamental differences in how these indices are constructed. The text-based KI index does not rely on citation data, whereas the citation-based CD index is influenced by variations in citation practices across different technological fields.

4.3. Correlation with technological impact

I further examine the relationship between text-based and citation-based innovation measures and a patent's technological impact, as measured by the number of forward citations—a widely recognized indicator of patent significance. Fig. 5a–c reveals a strong, nearly linear positive relationship between KI_5 and forward citations. This suggests that patents with higher KI_5—indicating greater textual similarity to future patents than to antecedents—are significantly more likely to be highly cited and become "hits." Patents with extremely high KI_5 values ($KI_5 > 1$) receive over five times as many citations within five years and more than ten times as many citations within ten years compared to patents with $KI_5 < 0$. The fixed effects regression analysis (Table 1, Models 1 and 3) further supports these findings. After controlling for field, year, section, and author fixed effects, a one-unit increase in KI_5 is associated with a 17.5 % increase in five-year forward citations and a 52.6 % increase in ten-year citations.

In contrast, the relationship between the citation-based disruption measure and forward citations, as shown in Fig. 5d–f, follows a non-linear pattern[3]. Patents with either extremely low or extremely high CD_5 values are the most likely to receive high citation counts and become hits. Specifically, patents with extreme CD_5 values (close to −1 or 1) have a >50 % chance of becoming hits, whereas those with moderate CD_5 values (around 0) receive the fewest citations. The fixed effects regression analysis (Table 1, Models 2 and 4) confirms this U-shaped relationship, as evidenced by the significantly positive

[2] See https://www.nber.org/research/business-cycle-dating for the details of Business Cycle Dating.

[3] When selecting patents for the CD index analysis (which requires at least one backward and one forward citation), the patents in this subset generally have higher citation counts and hit rates compared to the full sample.

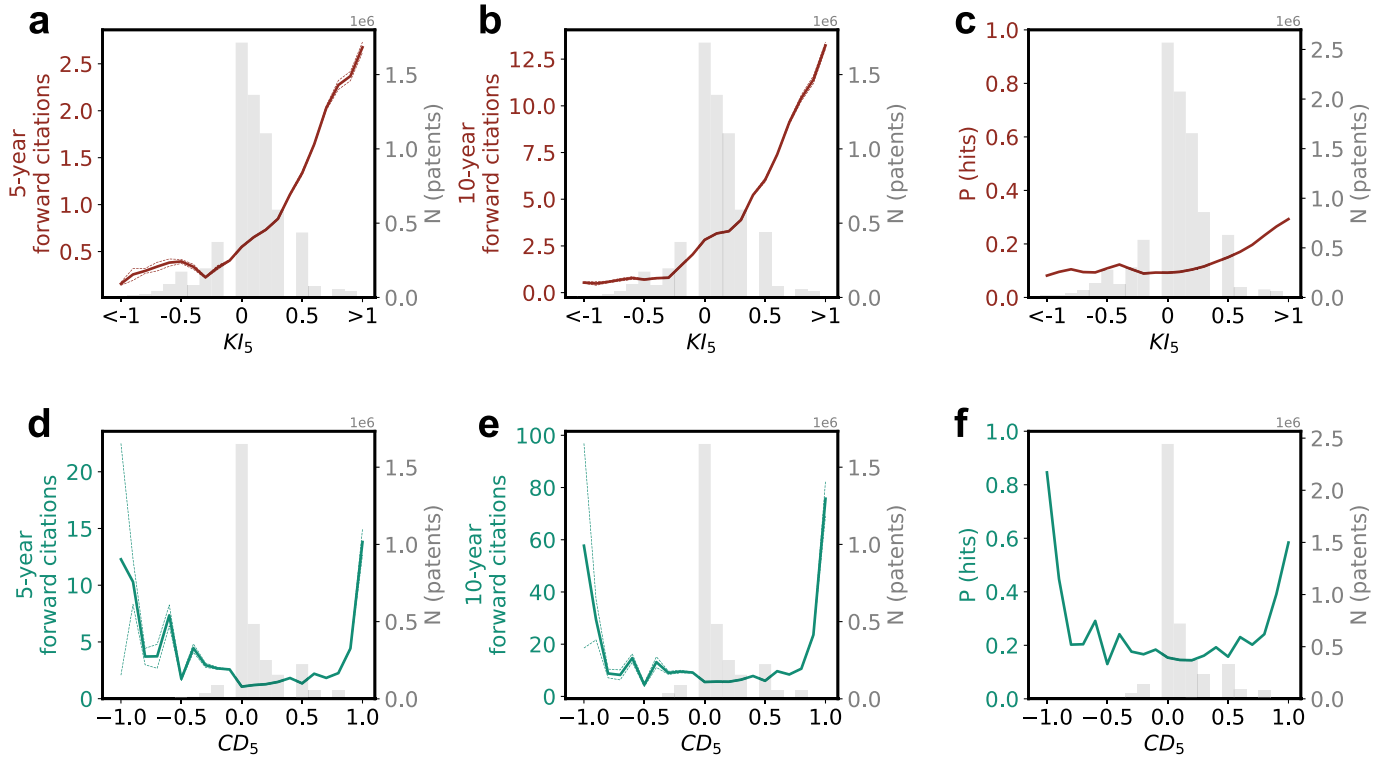


Fig. 5. Technological Innovation and Patent Impact. (a) The colored line represents the average number of five-year forward citations for patents grouped by KI_5 score (bin width = 0.1), while the grey histogram depicts the distribution of patents across KI_5 bins. The results show that higher KI_5 scores are associated with increased forward citations. (b) A similar positive relationship is observed between KI_5 and ten-year forward citations. (c-d) The probability of a patent being classified as a "hit" (i.e., ranking within the top 10 % of highly cited patents in its year and CPC section, measured up to 2022) increases with KI_5, reinforcing its predictive power for patent impact. (d-f) In contrast, forward citations and the probability of "hit" classification do not follow a linear pattern with CD_5. Instead, a U-shaped relationship emerges, suggesting that patents with either very low or very high CD_5 scores tend to receive more citations. Only patents with both backward and forward citations are included in the CD_5 analysis. Dotted lines indicate 95 % bootstrapped confidence intervals across all panels.

Table 1
Relationship between KI_5, CD_5 and forward citations of patents.

Models	(1)	(2)	(3)	(4)
	$\ln(Cit_5 + 1)$		$\ln(Cit_{10} + 1)$	
KI_5	0.1754***		0.5261***	
	(0.0286)		(0.0600)	
CD_5		0.0612		0.4305***
		(0.0699)		(0.0949)
CD_5 square		0.7385***		0.7112***
		(0.1167)		(0.1461)
Year FE	Yes	Yes	Yes	Yes
Section FE	Yes	Yes	Yes	Yes
Author FE	Yes	Yes	Yes	Yes
Obs.	25,036,484	11,461,290	25,036,484	11,461,290
R^2	0.3491	0.3994	0.4699	0.5360

Note: All models are estimated using OLS regression. Standard errors, clustered at the filed year level, are reported in parentheses. Statistical significance is indicated as follows: $^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. The sample size exceeds the number of patents, as patents are duplicated when associated with multiple sections or authors.

coefficient for the squared CD_5 term.

4.4. Correlation with team features

Wu et al. (2019) demonstrated that smaller teams tend to drive disruption in science and technology by generating novel ideas and identifying new opportunities, whereas larger teams are more likely to develop existing knowledge. In a recent study, Lin et al. (2023) found that researchers in remote teams are less likely to make breakthrough discoveries compared to their on-site counterparts, as indicated by a lower probability of producing disruptive papers or patents (CD index >0).

Building on these findings, I investigate the relationship between team size, collaboration distance, and both text-based and citation-based innovation metrics. Collaboration distance is measured as the average or median geographical distance between team members, using disambiguated data from Patents View and focusing exclusively on patents produced by teams with more than one member. I employ fixed effects regression models, using single-member teams (team size = 0) or fully on-site teams (distance = 0) as baselines, to examine the effects of team size and collaboration distance on KI_5 and CD_5 while controlling for field, year, section, and author fixed effects. The results are presented in Fig. 6a–c, with regression tables available in the Supplementary Materials. To provide a clearer perspective on effect sizes, I also report KI_5 and CD_5 percentiles in Fig. 6d–f.

The findings on team size and CD_5 align with Wu et al. (2019), showing that average CD_5 decreases as team size increases. However, I also find that KI_5 increases with team size. The regression results indicate that, compared to single-member teams, teams of two members experience an 11.6 % increase in KI_5 but a 17.6 % decrease in CD_5, while teams of five members see a 44.7 % increase in KI_5 but a 97.2 % decrease in CD_5. These effects remain consistently significant regardless of how large or small teams are defined, and the effect sizes are substantial.

Similarly, the relationship between collaboration distance and CD_5 corroborates Lin et al. (2023). CD_5 declines as collaboration distance increases, with the sharpest drop occurring between on-site and remote teams. In contrast, my analysis reveals that KI_5 increases with greater collaboration distance. The regression results suggest that compared to

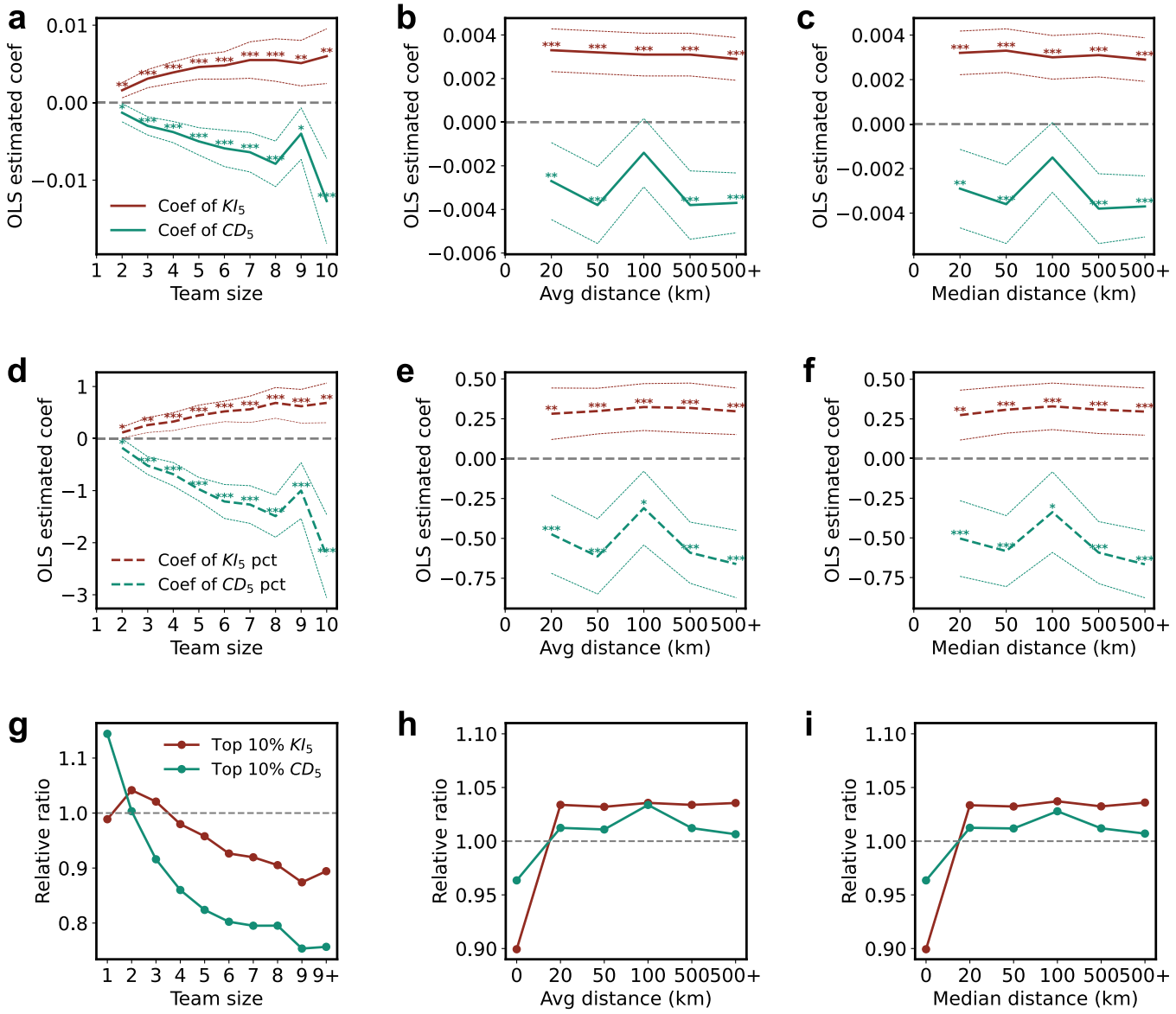


Fig. 6. Team Size, Collaboration Distance, and Technological Innovation. (a) Estimated coefficients of KI_5 and CD_5 with team size using OLS fixed-effects regressions, with single-member teams (team size = 0) as the baseline (grey horizontal line). 95 % confidence intervals are based on standard errors clustered at the filing year level. Statistical significance is indicated as follows: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. (b) Estimated coefficients of KI_5 and CD_5 with average collaboration distance, using fully on-site teams (distance = 0) as the benchmark. Distance categories include 0–20 km, 20–50 km, 50–100 km, 100–500 km, and >500 km. (c) Estimated coefficients of KI_5 and CD_5 with median collaboration distance. (d–f) Same as (a–c), but reporting KI_5 and CD_5 percentile values instead of raw scores. These results indicate that high KI_5 is associated with larger and more geographically dispersed teams, whereas high CD_5 is linked to smaller, on-site teams. (g–h) The relative ratio of each team structure in producing potential breakthroughs ranked in the top 10 % of KI_5 and CD_5.

on-site teams, teams with an average distance of 0–20 km experience a 28.1 % increase in KI_5 but a 47.5 % decrease in CD_5, while teams with an average distance >500 km see a 29.7 % increase in KI_5 but a 66.1 % decrease in CD_5.

These findings suggest that small and remote teams tend to produce patents with lower CD_5 values but higher KI_5 values. This indicates that while they are less likely to generate disruptive innovations within citation networks, they may still contribute significantly to technological advancement through influential innovations. Despite their distinct relationships with team characteristics, KI_5 and CD_5 exhibit similar patterns in identifying which teams are most likely to produce breakthrough patents. Here, I define potentially breakthrough patents as those in the top 10 % of KI_5 and CD_5 and analyze their relative ratio across different team structures. The results, presented in Fig. 6g–i, show that small teams are significantly more likely to generate patents with KI_5 and CD_5 above the 90th percentile compared to large teams. Similarly, remote teams have a much higher likelihood of producing top-tier KI_5 or CD_5 patents than on-site teams.

4.5. Alternative explanations

To further explore why large teams tend to produce patents with lower CD_5 but higher KI_5, I examine the role of references (i.e., backward citations). Leibel and Bornmann (2024) suggest that the CD index may be influenced by the citations received by a focal patent's references. Building on this idea, I consider reference age and reference popularity as potential explanatory factors. My analysis reveals that KI_5 is positively associated with reference age and negatively associated with reference popularity, whereas CD_5 exhibits the opposite pattern—negatively correlated with reference age and positively correlated

with reference popularity. This suggests that patents citing older or less popular references tend to have higher KI_5 but lower CD_5. This relationship is intuitive: KI_5 captures originality and frontier knowledge, making it more strongly linked to new and highly cited references. Conversely, CD_5 is affected by the citation patterns of these references (Leibel and Bornmann, 2024), influencing its denominator and driving its relationship with reference characteristics in the opposite direction.

To test this mechanism, I incorporate reference age and reference popularity as mediating variables. The results, presented in Table 2 and Fig. 7, confirm this mediation effect. Larger teams tend to cite newer and more widely cited references (lower reference age, higher reference popularity), which leads to an increase in KI_5 and a decrease in CD_5. Interaction effects between team size and reference characteristics—such as reference age, popularity, and count—further support this mediation pathway. However, this is just one possible explanation. These findings suggest that the differing trends in KI_5 and CD_5 may be partially driven by their distinct relationships with backward citations. Future research should further investigate these dynamics to determine which team structures are most conducive to fostering innovation.

5. Discussion

5.1. Comparing the KI and CD indices

This paper advances the study of innovation metrics by comparing two distinct measures of patent innovation dynamics: the text-based KI index, which measures breakthrough innovation, and the citation-based CD index, which measures technological disruption. Table 3 summarizes the strengths, weaknesses, similarities, and differences between these indices. Both aim to identify transformative patents but employ different methodologies—the KI index analyzes the full text of patents to assess breakthrough potential, while the CD index examines citation structures to evaluate disruption.

The KI index captures global breakthrough innovation by evaluating a patent's textual novelty (low similarity to prior patents) and impact (high similarity to future patents) within a specified time window. This approach tracks broader technological evolution independently of citation behaviors, making it robust to biases in citation practices (Yang, 2025; Yang et al., 2023b; Yang et al., 2025a). In contrast, the CD index is computationally simpler, relying on citation data to assess how a patent disrupts its knowledge domain by altering citation patterns between its antecedents and descendants. The CD index remains effective over long-time windows, providing reliable insights into sustained disruptive effects.

Each metric has limitations. The KI index is computationally intensive, requiring advanced natural language processing (NLP) techniques to extract meaningful patterns from patent text. It is also less effective over long-time windows due to the sparsity of relevant patents, which reduces the informativeness of textual similarity calculations. Furthermore, the KI index does not explicitly capture relationships between a patent's predecessors and successors. Conversely, the CD index is limited by its localzed scope, reflecting only first-order citation networks and missing broader knowledge flows. It is highly sensitive to the number and citation impact of a patent's references and requires longer observation periods for reliable results. Despite these differences, both indices exhibit strong correlations at an aggregated level and align with ex-ante originality measures based on CPC codes. Notably, the KI index excels at identifying expert-selected technological milestones, while both indices effectively highlight patents within narrow knowledge domains, indicating that breakthrough innovations often arise in specialized technological areas.

A key distinction between these indices lies in their temporal trends. The KI index fluctuates with economic cycles, reflecting broader shifts in innovation activity, while the CD index has shown a long-term decline, likely due to changes in patenting and citation behaviors. Their relationships with patent impact also differ—the KI index has a nearly linear positive relationship with future citations, whereas the CD index follows a U-shaped pattern, indicating that both highly disruptive and highly consolidating patents tend to receive more citations. The indices also diverge in their association with team structures. Larger teams tend to produce patents with higher KI indices but lower CD indices, whereas smaller teams are more likely to generate disruptive patents. Similarly, collaboration distance plays a role—patents from remote teams are associated with higher KI indices but lower CD indices, whereas patents from on-site teams are more likely to be disruptive.

Overall, the KI and CD indices provide complementary insights into technological breakthroughs. The KI index offers a broad, text-based measure of innovation, while the CD index identifies localized disruptions in citation networks. The KI index serves as a stable global indicator of novelty, whereas the CD index remains valuable for detecting shifts within specific knowledge domains. Given their distinct strengths and limitations, future research could integrate both metrics to develop a more comprehensive approach to measuring technological disruption and breakthrough innovation.

5.2. Rethinking innovation: breakthroughs, disruptions, and beyond

The comparative analysis of the KI and CD indices reveals both synergies and divergences that may shed light on the multifaceted concepts of “breakthrough,” “disruption,” and related terms like “novelty” and “impact” in the innovation literature (Hall and Jaffe, 2018; Higham et al., 2021). While my empirical findings highlight the indices' similarities (e.g., identifying patents from original, narrowly focused knowledge bases) and differences (e.g., the KI index's sensitivity to

Table 2
Alternative explanations.

Models	(1)	(2)	(3)	(4)	(5)	(6)
	KI_5 percentile			CD_5 percentile		
ln(Team size)	0.3422*** (0.0894)	0.9812** (0.2831)	1.441*** (0.1415)	−0.0081 (0.1422)	2.985*** (0.2801)	−1.911*** (0.2183)
ln(#Ref)	−0.1724** (0.0574)			−10.02*** (0.1871)		
ln(Team size) × ln(#Ref)	0.0789*** (0.0209)			0.1599*** (0.0357)		
ln(Ref age)		−1.231*** (0.2156)			6.363*** (0.4978)	
ln(Team size) × ln(Ref age)		−0.2089 (0.1192)			−1.923*** (0.1243)	
ln(Ref pop)			3.427*** (0.3620)			−19.21*** (0.4326)
ln(Team size) × ln(Ref pop)			−0.5455*** (0.0654)			0.8766*** (0.1105)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Section FE	Yes	Yes	Yes	Yes	Yes	Yes
Author FE	Yes	Yes	Yes	Yes	Yes	Yes
Obs.	23,144,064	22,594,285	22,814,933	11,461,192	11,394,577	11,460,754
R^2	0.8139	0.8140	0.8169	0.4735	0.4176	0.4829

Note: All models are estimated using OLS regression. Standard errors, clustered at the filed year level, are reported in parentheses. Statistical significance is indicated as follows: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. The sample size exceeds the number of patents, as patents are duplicated when associated with multiple sections or authors.

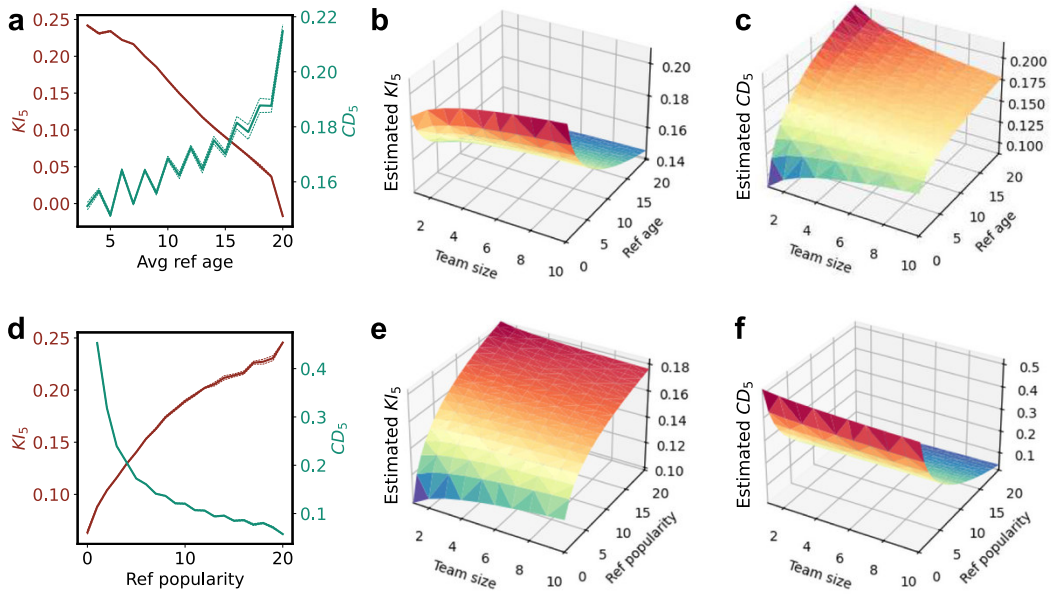


Fig. 7. Alternative Explanations for the Effect of Team Size on KI_5 and CD_5. (a) Relationship between KI_5, CD_5, and the average reference age (the mean time gap between the filing years of cited patents and the focal patent). KI_5 is negatively associated with reference age, whereas CD_5 shows a positive association. (b–c) Marginal predictions of KI_5 and CD_5 under different combinations of team size and reference age, based on fixed-effects OLS regression models with interaction terms. (d) Relationship between KI_5, CD_5, and reference popularity (the average five-year forward citations of a patent's references). KI_5 is positively associated with reference popularity, whereas CD_5 is negatively associated. (e–f) Marginal predictions of KI_5 and CD_5 under different combinations of team size and reference popularity, based on fixed-effects OLS regression models with interaction terms. These findings suggest that larger teams tend to cite newer and more popular references, influencing the observed differences between KI_5 and CD_5.

economic cycles versus the CD index's long-term decline, or their contrasting relationships with team size and citation impact), I offer the following conjectures on how these patterns might relate to the varied interpretations of innovation-related terms, without claiming definitive resolutions.

First, the KI index's linear correlation with forward citations and its fluctuation with economic cycles suggest it may align closely with a concept of "breakthrough" that emphasizes sustained influence and contextual adaptability. Breakthroughs, in this sense, could be innovations that not only introduce novel ideas but also resonate widely with future technological trajectories, particularly during periods of economic exploration (e.g., recessions). The KI index's reliance on global textual similarity might capture this by reflecting a patent's ability to set a new technological agenda (Cong et al., 2024; Kelly et al., 2021). I speculate that the KI index's strength in identifying such innovations could make it a valuable tool for studying breakthroughs as globally influential, context-sensitive advancements, though its computational intensity may limit its applicability to smaller-scale studies.

In contrast, the CD index's U-shaped relationship with citations and its focus on local citation networks suggest it may better reflect "disruption" as a polarized phenomenon—encompassing both radical displacements of prior technologies (high CD scores) and significant consolidations of existing knowledge (low CD scores). This aligns with the view of disruption as a shift in citation patterns (Funk and Owen-Smith, 2017), where a patent either destabilizes or reinforces its knowledge domain. The CD index's decline over time and sensitivity to reference characteristics might indicate that disruption, as measured within citation networks, is increasingly constrained by citation practices or patenting norms. I hypothesize that the CD index may be particularly suited to studying disruptions within specific technological niches, where citation networks are dense, but less effective for capturing broader, textually driven shifts in innovation.

Finally, the interplay between novelty and impact, central to both indices, raises questions about how these dimensions relate to "creativity," "originality," or "transformative innovation" in the literature (Hall and Jaffe, 2018; Higham et al., 2021). The KI index's emphasis on textual divergence from the past and convergence with the future might align with creativity as a process of recombining distant ideas into influential outcomes (Uzzi et al., 2013). The CD index, by contrast, might capture transformative innovation as a structural shift in knowledge flows, where the novelty lies in altering citation dependencies (Funk and Owen-Smith, 2017). I speculate that integrating the KI index's global textual perspective with the CD index's local citation focus could help disentangle these concepts, potentially revealing whether creativity precedes disruption or if transformative innovations require both novelty and structural impact.

These conjectures are not definitive but offer a framework for future research to explore how text-based and citation-based metrics can refine our understanding of innovation. By acknowledging the complementary strengths of the KI and CD indices, I hope researchers might better navigate the conceptual ambiguities surrounding "breakthrough" and "disruption," tailoring their choice of metrics to specific innovation contexts or combining them for a more holistic view of technological advancement.

5.3. Limitations

Despite the contributions of this study, several limitations must be acknowledged. First, while this paper aims to provide a comprehensive observational analysis, it does not seek to establish causal relationships. As such, the observed trends in technological innovation indices and their implications should be interpreted with caution. Both large and small teams offer distinct advantages, and this study does not intend to critique small and remote teams but rather to offer an alternative perspective on their disruptive potential. Second, although the text-based breakthrough measure attempts to capture global information by incorporating all patent pairs, it is not without limitations. One key challenge lies in distinguishing meaningful patterns from noise, particularly due to variations in language style and expression (Kelly et al., 2021). Addressing these nuances remains a task for future research. Both the text-based and citation-based disruption measures are dynamic, and further studies could explore their evolving patterns over time, moving

Table 3
Strengths, weaknesses, similarities and differences of the KI and CD Indices.

Aspect	Text-Based KI Index	Citation-based CD index
Strengths	1. Captures global innovation by analyzing all patents before and after the focal patent within a specific time window. 2. Independent of citation data, avoiding biases from citation practices. 3. Effective with short observation windows.	1. Measures disruption within local citation networks, reflecting how a focal patent alters relationships between its antecedents and descendants. 2. Relatively straightforward to compute. 3. Remains effective over long-time windows.
Weaknesses	1. Computationally intensive, requiring NLP techniques and complex processing. 2. Requires methods to enhance the signal-to-noise ratio in text analysis. 3. Does not explicitly capture relationships between a focal patent's antecedents and descendants. 4. Less effective over long-time windows due to sparsity of relevant patents.	1. Only reflects local citation relationships within the first-order citation network, missing broader knowledge flows. 2. Sensitive to the number and citation impact of a patent's references. 3. Requires a longer time window to yield reliable results. 4. Limited validity in identifying expert-selected breakthroughs. 5. Susceptible to measurement and truncation biases from incomplete or time-limited citation data.
Similarities	1. Patents with high KI index scores often have high CD index scores, showing strong correlations at an aggregated level. 2. Both align with ex-ante originality measures based on CPC codes. 3. Both reflect patents associated with narrow knowledge domains.	
Differences	1. Purpose: the KI index measures breakthrough innovation through textual novelty and impact, while the CD index measures disruption through shifts in citation patterns. 2. Temporal Trends: the KI index fluctuates with economic cycles, while the CD index exhibits a long-term decline. 3. Forward Citations: the KI index has a nearly linear positive relationship with future citations, whereas the CD index follows a U-shaped pattern. 4. Team Size: Larger teams tend to produce patents with higher KI but lower CD scores. 5. Collaboration Distance: Remote teams tend to generate patents with higher KI but lower CD scores.	

beyond binary perspectives on innovation. Lastly, this study focuses exclusively on patent-based innovations, which may not fully capture the disruptive potential of non-patentable innovations, particularly in industries where open-source practices or informal knowledge exchange are prevalent. Future research should investigate the applicability of text-based breakthrough measures to a broader spectrum of innovation outputs, including scientific publications, software development, and data-driven industries.

CRediT authorship contribution statement

Alex J. Yang: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The author declares no competing interest relevant to this article.

Acknowledgments

This research was supported by the National Natural Science Foundation of China Young Student (Doctoral) Project (#724B2015). The author thanks Professor Dimitris Papanikolaou for providing data on the text-based KI metric and Professor Richard B. Freeman at Harvard University for insightful discussions. The author also appreciates the valuable feedback from the anonymous reviewers.

Appendix A. Descriptive analysis

Table A1
Descriptive statistics.

Variables	Obs.	Mean	STD	Min	10 %	50 %	90 %	Max
KI_5	6,384,413	0.16	0.28	−2.62	−0.15	0.19	0.44	4.23
CD_5	3,036,683	0.17	0.21	−0.99	0.00	0.11	0.50	1.00
Cit_5	6,384,427	0.79	3.30	0.00	0.00	0.00	2.00	560.00
Cit_{10}	6,384,427	3.61	13.89	0.00	0.00	0.00	9.00	2942.00
Team size	6,383,667	2.54	1.83	1.00	1.00	2.00	5.00	133.00
Avg team distance	6,384,427	456.3	1619.6	0.0	0.0	0.0	803.3	19,733.3
Median team distance	6,384,427	498.8	1831.0	0.0	0.0	0.0	740.1	19,733.3
Nominal economic value	2,185,903	23.8	64.4	0.0	0.1	7.0	55.5	4588.5
Real economic value	2,185,903	11.5	31.4	0.0	0.1	3.6	26.5	2573.8
1st CPC group	6,384,245	0.02	0.13	0.00	0.00	0.00	0.00	1.00
1st CPC group pair	5,559,063	0.55	0.50	0.00	0.00	1.00	1.00	1.00
Avg CPC group age	5,559,063	11.41	7.91	0.00	1.80	10.33	22.27	46.00
HHI score	6,384,245	0.78	0.27	0.11	0.50	1.00	1.00	1.00
#Reference	6,384,305	15.5	62.3	0.0	1.0	6.0	27.0	5859.0
Avg reference age	5,811,469	9.95	5.80	0.00	3.67	8.83	17.92	46.00
Reference popularity	5,857,843	8.83	12.50	0.00	1.75	5.63	18.64	1432.00

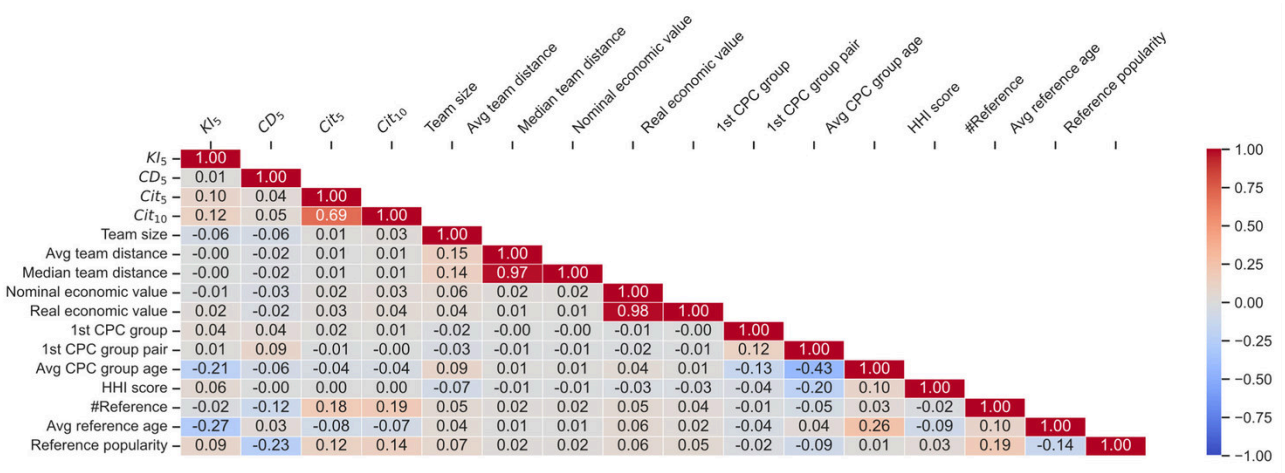


Fig. A1. Pearson correlation matrix of variables.

Appendix B. Correlation with economic values

I also assess the economic value of patents to evaluate their market or economic significance. The economic value of a patent can be quantified by analyzing stock market reactions to patent grants, as proposed by Kogan et al. (2017). This approach estimates the monetary value of innovations by observing fluctuations in a firm's stock price within a three-day window following the issuance of a patent. Both nominal and real values are considered, with the real values adjusted to 1982 dollars using the Consumer Price Index. The methodology for calculating the economic value of patents, as outlined by Kogan et al. (2017), emphasizes a firm's idiosyncratic return, R, which is defined as the firm's stock return minus the return on the market portfolio. The idiosyncratic return, R, can be decomposed into two components: the value associated with the patent j (ξ_j) as:

$$\xi_j = (1-\overline{\pi})\frac{M_j}{N_j}E\left[V_j|R_j\right],$$

where N_j is typically 1, but when multiple patents are granted to the same firm on the same day, N_j represents the average fraction. M_j denotes the firm's market capitalization prior to the patent issuance, and $\overline{\pi}$ represents the market's ex ante probability of patent approval, which is assumed to be 2.27 (Kogan et al., 2017).

Assuming the distribution of the value V follows a normal distribution truncated at zero, the expected value of V_j conditional on R_j, $E\left[V_j|R_j\right]$, can be expressed as:

$$E\left[V_j|R_j\right] = \delta_{ft}R_j + \sqrt{\delta_{ft}}\sigma\frac{\phi\left(\frac{-R_j\sqrt{\delta_{ft}}}{\sigma}\right)}{1-\Phi\left(\frac{-R_j\sqrt{\delta_{ft}}}{\sigma}\right)},$$

where ϕ and Φ represent the probability density function (PDF) and cumulative distribution function (CDF) of the standard normal distribution, respectively. δ refers to the signal-to-noise ratio, which can be estimated by observing the increase in volatility of firm returns around patent announcement days. σ is estimated using the regression:

$$\log\left(R_{fd}\right)^2 = \gamma I_{fd} + cZ_{fd} + \mu_{fd},$$

where R_{fd} refers to the 3-day idiosyncratic return of firm f, starting on day d. This regression captures the relationship between the firm's return volatility and patent announcement events. In the sample of this paper, δ is estimated to be 0.0142. While our text-based and citation-based measures are based on the filing date, the economic value of patents is more closely tied to the issue date. This measure estimates the monetary value of innovations based on changes in a firm's stock price within a three-day window following the patent's issuance. Both nominal and real values are considered, with the real value adjusted to 1982 dollars using the Consumer Price Index.

The results show a positive relationship between the text-based breakthrough measure (KI_5) and the economic value of patents when KI_5 is >0. However, patents with very low KI_5 can also exhibit high economic value. Interestingly, CD_5 demonstrates a significantly negative relationship with economic value, suggesting that patents contributing to technological consolidation tend to have higher economic value compared to disruptive ones—except in cases of the most groundbreaking innovations.

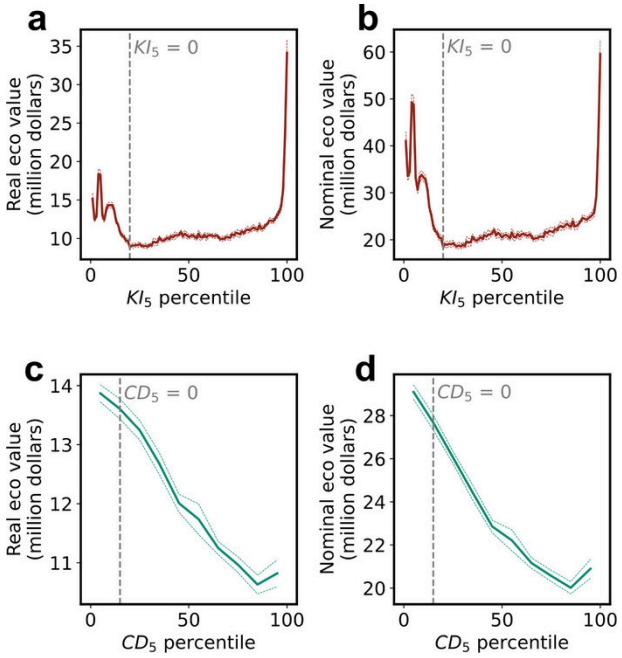


Fig. B1. Technological Innovation and the Economic Value of Patents. (a) For 2,185,903 patents filed between 1980 and 2017, the real economic value (adjusted to 1982 dollars using the Consumer Price Index) generally increases with KI_5 when KI_5 is above 0. However, patents with low KI_5 also exhibit very high economic values. The economic value is estimated based on the stock price change of the firm within a three-day window following the patent's issuance. The sample size is smaller because this measure is only available for patents where the assignee can be matched with firm PERMNO codes (Unique identifiers assigned to companies in the Center for Research in Security Prices (CRSP) database). (b) Similarly, the nominal economic value increases with KI_5 when KI_5 is above 0, though patents with low KI_5 also show high economic values. (c–d) In contrast, both real and nominal economic values decrease as CD_5 percentiles increase. Dotted lines in all panels represent 95 % bootstrapped confidence intervals.

Appendix C. Patterns of the mCD index

In addition to the CD index, Funk and Owen-Smith (2017) introduced the mCD index, which is calculated as the product of the CD index and forward citations. While the CD index primarily measures local disruption within citation networks, the mCD index incorporates citation impact, potentially offering a broader perspective on a patent's technological influence. Despite this advantage, the mCD index is less frequently used in the literature compared to the CD index. I analyze the temporal and structural patterns of the mCD index using a 5-year window (mCD_5) in Fig. C1.

The analysis reveals that mCD_5 exhibits temporal dynamics closely aligned with those of the KI index, displaying a clear responsiveness to global economic cycles. Unlike the CD index, which generally shows a downward trend over time, mCD_5 fluctuates in tandem with economic expansions and contractions. This suggests that mCD_5 may be more attuned to macroeconomic conditions shaping the innovation landscape, much like the KI index. Furthermore, the temporal stability of mCD_5 across sections indicates that this metric is robust to differences in technological fields, offering a consistent signal across the innovation ecosystem. Another distinguishing feature of mCD_5 is its apparent insensitivity to section-year size. Whereas the CD index is negatively correlated with the volume of patenting activity within specific technological domains, mCD_5 remains largely uncorrelated with such size effects. This insensitivity suggests that mCD_5, akin to the KI index, is less biased by the structural composition of the patent landscape and may better reflect the intrinsic properties of inventive activity.

Collaboration patterns also appear to shape mCD_5 values. Patents produced by larger teams and those involving remote or distributed collaborations tend to have higher mCD_5 scores. This stands in contrast to the CD index, where smaller, more localized teams tend to generate more disruptive patents. The alignment of mCD_5 with broader and more distributed team configurations further underscores its similarity to the KI index and suggests that it may be more reflective of the collaborative processes underpinning contemporary innovation. Finally, while mCD_5 shares many structural and temporal characteristics with the KI index, it diverges in its relationship with forward citations. The index follows a U-shaped curve with respect to forward citations, indicating that patents with either very low or very high mCD_5 values tend to receive more citations. This non-monotonic relationship mirrors the pattern observed with the CD index and stands in contrast to the KI index, which exhibits a strong, nearly linear positive association with citation counts. The U-shaped pattern implies that both highly disruptive and marginal patents—each at opposite ends of the mCD_5 spectrum—can attract significant attention within the innovation system.

Overall, the mCD index occupies a conceptual space that bridges the KI and CD indices. It shares the KI index's sensitivity to economic fluctuations, while also retaining the CD index's non-linear relationship with citation impact. These findings suggest that the mCD index captures a hybrid dimension of innovation, reflecting both the structural disruption of new knowledge and the broader influence conveyed through forward citations.

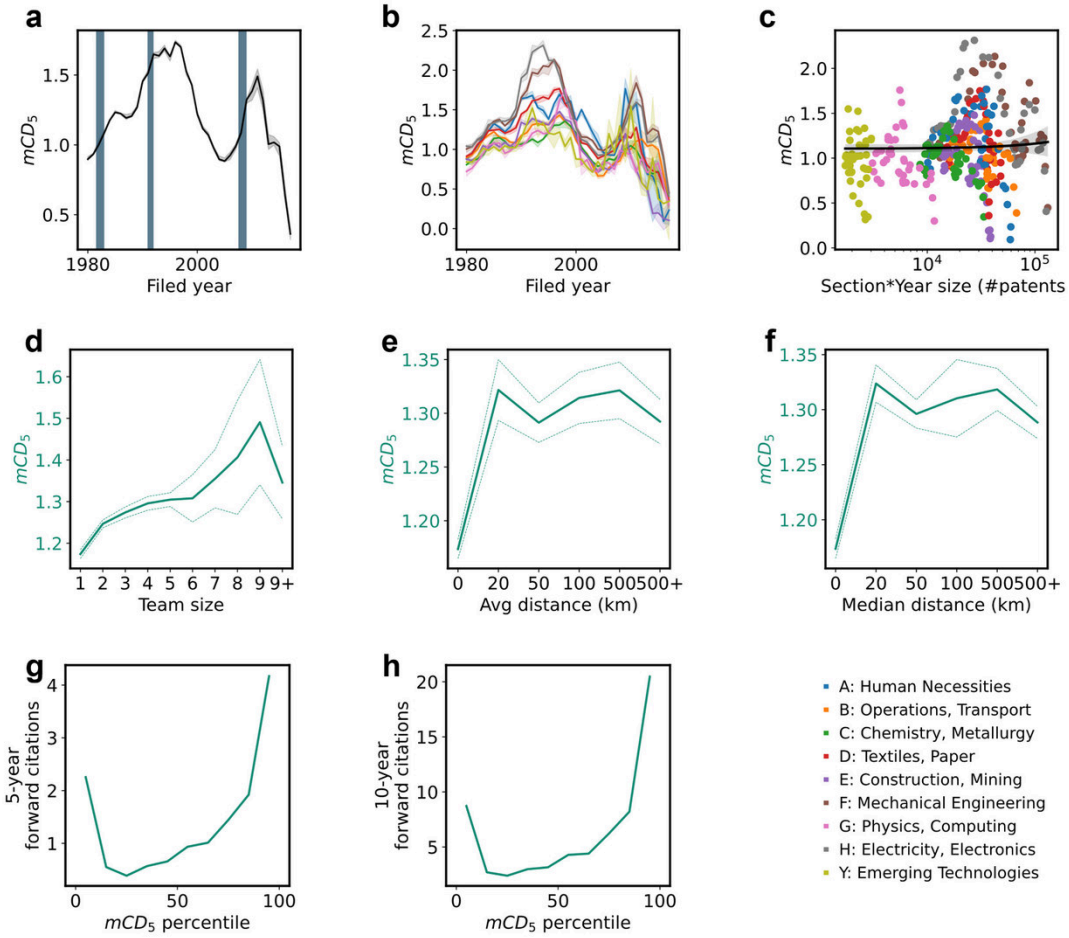


Fig. C1. Dynamic patterns of the mCD index. (a) The annual average mCD_5 from 1980 to 2017, showing trends that generally align with the expansions and contractions of the global economic cycle, similar to KI_5. Shaded blue regions indicate periods of economic recession as identified by NBER. (b) Temporal trends in mCD_5 remain consistent across different CPC sections. (c) mCD_5 does not show a significant correlation with section-year size. The plot displays the average mCD_5 against the number of patents per section-year, with each scatter point representing a specific section-year and colors corresponding to different CPC sections. (d) The relationship between team size and mCD_5, indicating that larger teams tend to have higher mCD_5 than smaller teams. (e–f) The relationship between collaboration distance and mCD_5, showing that remote teams generally have higher mCD_5 than on-site teams. (g–h) Five-year and ten-year forward citations, as well as the probability of a patent being classified as a "hit", do not display a consistent linear relationship with mCD_5. Instead, a U-shaped pattern emerges, suggesting that patents with very low or very high mCD_5 tend to receive more citations. Shaded areas or dotted lines in all panels represent 95 % bootstrapped confidence intervals. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Appendix D. Expert-selected technological milestones

I present the results concerning the expert-selected technological milestones (Kelly et al., 2021), showing that both the text-based KI index and the citation-based CD index effectively identify these breakthroughs. All KI values are above zero, and most CD values exceed zero, confirming their ability to capture historical breakthroughs. However, these indices may differ in their assessment of standard or more recent technological patents. This discrepancy likely stems from shifts in citation patterns, particularly in the use of backward citations.

Table D1
expert-selected technological milestones.

ID	Patent	Filed year	KI_5	KI_5 pct	CD_5	CD_5 pct
4399216	Richard Axel Co-transformation	1980	0.7881	98.8	0.8750	99.7
4464652	Apple Lisa Mouse	1982	0.1928	51.4	0.2329	71.6
4590598	Gordon Gould Laser	1984	0.2082	54.8	0.1000	48.7
4634665	Richard Axel Co-transformation	1983	0.8470	99.1		
4683195	Kary B. Mullis polymerase chain reaction	1986	0.6123	96.7	0.8293	99.3
4736866	Leder/Stewart transgenic (genetically modified) animals	1984	1.2071	99.8	0.2500	73.8
4744360	Patricia Bath Cataract Laserphaco Prob	1986	0.5427	94.9	0.1667	62.3
4799258	Donald Watts Davies Packet-switching technology	1985	0.5315	94.5	0.0896	45.7
4816397	Michael A. Boss recombinant antibodies	1984	0.8144	98.9	0.3333	81.6
4816567	Shmuel Cabilly immunoglobulins	1983	0.8299	99.0	0.0423	31.0
4838644	Ellen Ochoa Recognizing Method	1987	0.5300	94.4	0.3529	83.7

(*continued on next page*)

Table D1 (*continued*)

ID	Patent	Filed year	KI_5	KI_5 pct	CD_5	CD_5 pct
4889818	(several) polymerase chain reaction	1987	0.7960	98.8	0.1022	49.4
4965188	(several) polymerase chain reaction	1987	0.6907	97.9	0.0241	23.6
5061620	Ann Tsukamoto Method for isolating the human stem cell	1990	0.9320	99.3		
5071161	Georey L Mahoon Airbag	1989	0.3847	85.2	0.2500	73.8
5108388	Stephen L. Troke Laser Surgery Method	1987	0.6324	97.1		
5149636	Richard Axel Co-transformation	1988	0.9443	99.4	0.2500	73.8
5179017	Richard Axel Co-transformation	1991	1.2949	99.8		
5184830	Saturo Okada, Shin Kojo Compact Hand-Held Video Game System	1992	0.6760	97.7	0.4138	87.1
5194299	Arthur Fry Post-It Note	1986	0.4035	86.9		
5225539	Gregory P. Winter Chimeric, humanized antibodies	1991	1.2145	99.8		
5272628	Michael Koss Core Excel Function	1990	0.7884	98.8	0.2188	69.9
5747282	Mark H. Skolnick BRCA1 gene	1995	1.0145	99.5		
5770429	Nils Lonberg human antibodies from transgenic mice	1995	0.8224	98.9	0.0158	20.0
5837492	(several) BRCA2 gene	1996	0.5896	96.2		
5939598	(several) Transgenic mice	1992	1.3558	99.9	0.0370	29.0
5960411	Hartman/Bezos/Kaphan/Spiegel 1-click buying	1997	1.6221	100.0	0.1851	65.2
6230409	Patricia Billings Geobond	1998	0.3925	85.9	0.3000	78.7
6285999	Larry Page Google PageRank	1998	0.7074	98.1	0.2907	78.1
6331415	Shmuel Cabilly Antibody molecules	1988	0.7682	98.6		
6455275	Richard Axel Co-transformation	1995	0.7966	98.8	−0.0588	4.6
6574628	Robert Kahn, Vinton Cerf Packet-Switching Knowbot	1997	1.0832	99.6	0.0160	20.0
6955484	Nicholas D. Woodman Harness system for attaching camera to user	2004	0.1692	46.4	0.4167	87.3
6985922	Janet Emerson Bashen LinkLine	2001	0.4254	88.8	0.1151	52.3

Appendix E. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.respol.2025.105295.

Data availability

The patent data used in this study is publicly available from PatentsView at https://patentsview.org/download/data-download-tables. The code for replication is accessible at https://github.com/AlexJieYang/RP_Text_vs_citations. For access to additional data, please contact the author directly.

References

Arthur, W.B., 2007. The structure of invention. Research Policy 36 (2), 274–287. https://doi.org/10.1016/j.respol.2006.11.005.

Arts, S., Hou, J., Gomez, J.C., 2021. Natural language processing to identify the creation and impact of new technologies in patent text: code, data, and new measures. Research Policy 50 (2), 104144. https://doi.org/10.1016/j.respol.2020.104144.

Arts, S., Melluso, N., Veugelers, R., 2025. Beyond citations: measuring novel scientific ideas and their impact in publication text. Rev. Econ. Stat. 1–33. https://doi.org/10.1162/rest_a_01561.

Azoulay, P., Graff Zivin, J.S., Manso, G., 2011. Incentives and creativity: evidence from the academic life sciences. Rand J. Econ. 42 (3), 527–554. https://doi.org/10.1111/j.1756-2171.2011.00140.x.

Becerra, A.Z., Kam, L., Kim, Y., Park, H., 2025. Identifying disruptive papers in pharmacy literature that changed clinical practice: 1954–2014. Am. J. Health-Syst. Pharm. 82 (6), 317–340.

Bornmann, L., Devarakonda, S., Tekles, A., Chacko, G., 2020. Are disruption index indicators convergently valid? The comparison of several indicator variants with assessments by peers. Quantitative Science Studies 1 (3), 1242–1259. https://doi.org/10.1162/qss_a_00068.

Bower, J.L., Christensen, C.M., 1995. Disruptive Technologies: Catching the Wave. J. Prod. Innov. Manag. 1, 75–76.

Chu, J.S.G., Evans, J.A., 2021. Slowed canonical progress in large fields of science. Proc. Natl. Acad. Sci. 118 (41), e2021636118. https://doi.org/10.1073/pnas.2021636118.

Cong, L.W., Liang, T., Zhang, X., Zhu, W., 2024. Textual Factors: A Scalable, Interpretable, and Data-Driven Approach to Analyzing Unstructured Information.

Deng, N., Gu, X., Fan, Y., Havlin, S., Zeng, A., 2025. The critical role of persistent disruption in advancing science. Nature Computational Science. https://doi.org/10.1038/s43588-025-00808-7.

Di Bona, G., Bellina, A., De Marzo, G., Petralia, A., Iacopini, I., Latora, V., 2025. The dynamics of higher-order novelties. Nat. Commun. 16 (1), 393. https://doi.org/10.1038/s41467-024-55115-y.

Esposito, C.R., 2023. The geography of breakthrough invention in the United States over the 20th century. Research Policy 52 (7), 104810. https://doi.org/10.1016/j.respol.2023.104810.

Fleming, L., 2001. Recombinant uncertainty in technological search. Manag. Sci. 47 (1), 117–132. https://doi.org/10.1287/mnsc.47.1.117.10671.

Fleming, L., Sorenson, O., 2001. Technology as a complex adaptive system: evidence from patent data. Res. Policy 30 (7), 1019–1039. https://doi.org/10.1016/S0048-7333(00)00135-9.

Fontana, M., Iori, M., Montobbio, F., Sinatra, R., 2020. New and atypical combinations: an assessment of novelty and interdisciplinarity. Research Policy 49 (7), 104063. https://doi.org/10.1016/j.respol.2020.104063.

Fortunato, S., Bergstrom, C.T., Boerner, K., Evans, J.A., Helbing, D., Milojevic, S., Petersen, A.M., Radicchi, F., Sinatra, R., Uzzi, B., Vespignani, A., Waltman, L., Wang, D., Barabasi, A.-L., 2018. Science of science. Science 359 (6379), eaao0185. https://doi.org/10.1126/science.aao0185.

Foster, J.G., Rzhetsky, A., Evans, J.A., 2015. Tradition and innovation in scientists' research strategies. Am. Sociol. Rev. 80 (5), 875–908. https://doi.org/10.1177/0003122415601618.

Funk, R.J., Owen-Smith, J., 2017. A dynamic network measure of technological change. Management Science 63 (3), 791–817. https://doi.org/10.1287/mnsc.2015.2366.

Gates, A.J., Ke, Q., Varol, O., Barabási, A.L., 2019. Nature's reach: narrow work has broad impact. Nature 575 (7781), 32–34. https://doi.org/10.1038/d41586-019-03308-7.

Grunvald, M.W., Williams, M.D., Rao, R.D., O'Donoghue, C.M., Becerra, A.Z., 2021. 100 disruptive publications in breast cancer research. Asian Pacific Journal of Cancer Prevention: APJCP 22 (8), 2385.

Hall, B.H., Harhoff, D., 2012. Recent research on the economics of patents. Annual Review of Economics 4, 541–565. https://doi.org/10.1146/annurev-economics-080511-111008.

Hall, B.H., Jaffe, A.B., 2018. Measuring science, technology, and innovation: a review. Annals of Science and Technology Policy 2 (1), 1–74. https://doi.org/10.1561/110.00000005.

Higham, K., de Rassenfosse, G., Jaffe, A.B., 2021. Patent quality: towards a systematic framework for analysis and measurement. Research Policy 50 (4), 104215. https://doi.org/10.1016/j.respol.2021.104215.

Hofstra, B., Kulkarni, V.V., Munoz-Najar Galvez, S., He, B., Jurafsky, D., McFarland, D.A., 2020. The diversity-innovation paradox in science. Proc. Natl. Acad. Sci. 117 (17), 9284–9291. https://doi.org/10.1073/pnas.1915378117.

Horen, S.R., Hansdorfer, M.A., Kronshtal, R., Dorafshar, A.H., Becerra, A.Z., 2021. The most disruptive publications in craniofacial surgery (1954–2014). Journal of Craniofacial Surgery 32 (7), 2426–2430.

Jones, B.F., Wuchty, S., Uzzi, B., 2008. Multi-university research teams: shifting impact, geography, and stratification in science. Science 322 (5905), 1259–1262. https://doi.org/10.1126/science.1158357.

Jurowetzki, R., Hain, D.S., Wirtz, K., Bianchini, S., 2025. The private sector is hoarding AI researchers: what implications for science? AI Soc. 1–8.

Kelly, B., Papanikolaou, D., Seru, A., Taddy, M., 2021. Measuring technological innovation over the long run. American Economic Review: Insights 3 (3), 303–320. https://doi.org/10.1257/aeri.20190499.

Kogan, L., Papanikolaou, D., Seru, A., Stoffman, N., 2017. Technological innovation, resource allocation, and growth*. Q. J. Econ. 132 (2), 665–712. https://doi.org/10.1093/qje/qjw040.
Kuhn, T.S., 1962. Historical structure of scientific discovery. Science 136 (3518), 760–764.
Leahey, E., Lee, J., Funk, R.J., 2023. What types of novelty are Most disruptive? Am. Sociol. Rev. 88 (3), 562–597. https://doi.org/10.1177/00031224231168074.
Lee, S., Kim, T., Yoon, J., Youn, H., 2024. Increasing Volume and Decreasing Disruption in US Case Law arXiv preprint.
Leibel, C., Bornmann, L., 2024. What do we know about the disruption index in scientometrics? An overview of the literature. Scientometrics 129, 601–639. https://doi.org/10.1007/s11192-023-04873-5.
Leydesdorff, L., Bornmann, L., 2021. Disruption indices and their calculation using web-of-science data: indicators of historical developments or evolutionary dynamics? J. Informet. 15 (4), 101219. https://doi.org/10.1016/j.joi.2021.101219.
Lin, Y., Frey, C.B., Wu, L., 2023. Remote collaboration fuses fewer breakthrough ideas. Nature 623 (7989), 987–991. https://doi.org/10.1038/s41586-023-06767-1.
Macher, J.T., Rutzer, C., Weder, R., 2024. Is there a secular decline in disruptive patents? Correcting for measurement bias. Research Policy 53 (5), 104992. https://doi.org/10.1016/j.respol.2024.104992.
Manso, G., Balsmeier, B., Fleming, L., 2023. Heterogeneous innovation over the business cycle. Rev. Econ. Stat. 105 (5), 1224–1236. https://doi.org/10.1162/rest_a_01104.
Mokyr, J., 1990. Punctuated equilibria and technological progress. Am. Econ. Rev. 80 (2), 350–354.
Mukherjee, S., Romero, D.M., Jones, B., Uzzi, B., 2017. The nearly universal link between the age of past knowledge and tomorrow's breakthroughs in science and technology: the hotspot. Science Advances 3 (4). https://doi.org/10.1126/sciadv.1601315.
Nielsen, M.W., Andersen, J.P., 2021. Global citation inequality is on the rise. Proc. Natl. Acad. Sci. 118 (7), e2012208118. https://doi.org/10.1073/pnas.2012208118.
Park, M., Leahey, E., Funk, R.J., 2023. Papers and patents are becoming less disruptive over time. Nature 613 (7942), 138–144. https://doi.org/10.1038/s41586-022-05543-x.
Shinichi, H., Matsui, A., 2024. A Deep Learning Approach to Quantitative Analysis of Japanese Artworks: Exploring the Latent Features of Ukiyo-e Artists. SSRN.
Strumsky, D., Lobo, J., 2015. Identifying the sources of technological novelty in the process of invention. Research Policy 44 (8), 1445–1461. https://doi.org/10.1016/j.respol.2015.05.008.
Sun, B., Kolesnikov, S., Goldstein, A., Chan, G., 2021. A dynamic approach for identifying technological breakthroughs with an application in solar photovoltaics. Technol. Forecast. Soc. Chang. 165, 120534. https://doi.org/10.1016/j.techfore.2020.120534.
Uzzi, B., Mukherjee, S., Stringer, M., Jones, B., 2013. Atypical combinations and scientific impact. Science 342 (6157), 468–472. https://doi.org/10.1126/science.1240474.
Wang, J., Veugelers, R., Stephan, P., 2017. Bias against novelty in science: a cautionary tale for users of bibliometric indicators. Research Policy 46 (8), 1416–1436. https://doi.org/10.1016/j.respol.2017.06.006.
Wang, S., Ma, Y., Mao, J., Bai, Y., Liang, Z., Li, G., 2023. Quantifying scientific breakthroughs by a novel disruption indicator based on knowledge entities. J. Assoc. Inf. Sci. Technol. 74 (2), 150–167. https://doi.org/10.1002/asi.24719.
Wu, L.F., Wang, D.S., Evans, J.A., 2019. Large teams develop and small teams disrupt science and technology. Nature 566 (7744), 378+. https://doi.org/10.1038/s41586-019-0941-9.
Wuchty, S., Jones, B.F., Uzzi, B., 2007. The increasing dominance of teams in production of knowledge. Science 316 (5827), 1036–1039. https://doi.org/10.1126/science.1136099.
Yang, A.J., 2025. Understanding the relationship between interdisciplinary knowledge and disruption in science. J. Inf. Sci. https://doi.org/10.1177/01655515251330614, 01655515251330614.
Yang, A.J., Deng, S., 2024. Dynamic patterns of the disruptive and consolidating knowledge flows in Nobel-winning scientific breakthroughs. Quantitative Science Studies 1–17. https://doi.org/10.1162/qss_a_00323.
Yang, A.J., Deng, S., Wang, H., Zhang, Y., Yang, W., 2023a. Disruptive coefficient and 2-step disruptive coefficient: novel measures for identifying vital nodes in complex networks. J. Informet. 17 (3), 101411. https://doi.org/10.1016/j.joi.2023.101411.
Yang, A.J., Hu, H., Zhao, Y., Wang, H., Deng, S., 2023b. From consolidation to disruption: a novel way to measure the impact of scientists and identify laureates. Inf. Process. Manag. 60 (5), 103420. https://doi.org/10.1016/j.ipm.2023.103420.
Yang, A.J., Gong, H., Wang, Y., Zhang, C., Deng, S., 2024a. Rescaling the disruption index reveals the universality of disruption distributions in science. Scientometrics 129 (1), 561–580. https://doi.org/10.1007/s11192-023-04889-x.
Yang, A.J., Zhao, S.X., Deng, S., 2024b. Revisiting delayed recognition in science: a large-scale and comprehensive study. J. Inf. Sci. https://doi.org/10.1177/01655515241244462.
Yang, A.J., Wang, F., Shi, Y., Zhang, Y., Wang, H., Deng, S., 2025a. Beyond surface correlations: reference behavior mediates the disruptiveness-citation relationship. Journal of Data and Information Science. https://doi.org/10.2478/jdis-2025-0029, 20250029–20250029.
Yang, A.J., Yan, X., Hu, H., Hu, H., Kong, J., Deng, S., 2025b. Are disruptive papers more likely to impact technology and society? J. Assoc. Inf. Sci. Technol. 76 (3), 563–579. https://doi.org/10.1002/asi.24947.