Identifying Firms with Stronger Innovative Ideas Using Machine Learning*

Ferhat Akbas[1], Furkan M. Çetin[2], and Egemen Genc[†1]

[1]College of Business Administration, University of Illinois at Chicago
[2]London School of Economics

Abstract

We propose a novel innovation measure based on the claim section of patents to identify firms that have a stronger pipeline of innovative ideas and a greater potential for successful innovation. Using advanced machine learning techniques, we compare the textual similarity and distinctiveness of the topical content of patent claim sections in a firm's innovation pipeline. Using various innovation productivity measures, we show that firms with a more distinct set of claims have better future innovation performance. We further examine the characteristics of innovation team employees in these firms. We find that firms with a distinct set of claims typically have more skilled innovators, characterized by greater gender equality, higher education levels, diverse skill sets, and longer tenure, making them better equipped to support innovation. Furthermore,firms with a more distinct set of claims have a higher operating performance suggesting that their this innovation success is reflected in their economic advancements and growth. However, the market only gradually incorporates this information into stock prices and these firms experience higher future abnormal stock returns.

This draft: September 2024

JEL classification: G12, G14, O32.

Keywords: Patent originality, patent productivity, firm profitability, firm performance

*We thank Stefano Cascino, Lauren Cohen, Joan Farre-Mensa, Dimitris Papanikolaou and workshop participants at the University of Illinois at Chicago, London School of Economics for their comments and suggestions. We also thank patent attorney Melih Okan for providing insights. All errors are ours.

[†]Ferhat Akbas: ferhat@uic.edu, Furkan Çetin: f.m.cetin@lse.ac.uk, Egement Genc: egenc@uic.edu

1 Introduction

Since the 1930s, it's been widely recognized that innovations boosting a firm's future cash flows are key to long-term economic growth (Solow, 1957; Arrow, 1962). According to the standard endogenous growth and technological progress models in economics, future innovation productivity is a function of current stock of ideas and the resources that are used to produce these new ideas (Bloom et al., 2020; Acemoglu, Akcigit, and Kerr, 2016). Accordingly, firms that effectively utilize their resources to generate and develop a rich stock of ideas are likely to have a stronger pipeline of innovation and are more likely to remain innovative. Therefore, evaluating a firm's pipeline of innovative ideas is vital for determining its capacity for future success in innovation.

However, the process of innovation is inherently complex and uncertain, and even substantial investments in research and development (R&D) can sometimes result in only modest outcomes. Additionally, recent research points to a trend of diminishing returns on innovation, mainly due to recent technological advances, challenging even seasoned innovators to generate fresh ideas and maintain their past levels of innovation (Bloom et al., 2020; Jones, 2009). Given these constraints and the critical role of innovation in sustaining strong economic growth, it is both challenging and essential to identify companies with promising new innovation ideas and allocate resources to them efficiently. However, evaluating a firm's pipeline of innovative ideas is difficult and requires a clear and concise measure to effectively quantify their quality.

In this study, we take a step in this direction and introduce a novel measure to quantify the quality of a firm's pipeline of innovative ideas and capacity to innovate successfully. Specifically, we employ state-of-the-art LLM based on Bidirectional Encoding Representation from Transformers (BERT) to generate high-dimensional embeddings that encapsulate the semantic meaning of patent claims. Our focus on the claim section of patents as a critical indicator of future innovation potential stems from its significance and unique role within company patents. Companies communicate their innovations through patents, which provide

them with exclusive legal rights and a competitive edge over the patented technology for a specified period (typically twenty years in the United States). As Ashtor (2019) points out, the claims section of a patent distinguishes itself from other parts by providing a precise description of the innovation, hence ensuring its distinctiveness across all patents. In fact, as acknowledged by patent agents, claims are the central aspects of patents, defining patents' primary contributions and exclusive rights granted by them. In cases of patent infringement, claims are the focal focus of scrutiny during court proceedings. Therefore, companies are highly motivated to invest considerable effort in meticulously crafting this section to safeguard their rights, precisely explain their contributions, and elucidate the creative advancements their work represents compared to prior endeavors.

While textual analysis of patents has been applied for some time, measuring the distinctiveness of a patent has remained a formidable challenge using traditional methods. Large Language Models (LLMs) offer a breakthrough in this domain, enabling us to quantify patent uniqueness by capturing the nuanced semantic meaning of text—a crucial aspect for patent analysis where word meanings are heavily context-dependent. We utilize this capability to create a novel metric for assessing textual distinctiveness among claims of patents granted to the same firm within the same technology group over the preceding five years. Using this metric, we then calculate the Innovation Pipeline Score (IPS) for each firm in each year.

We propose that companies capable of generating new ideas and poised to make novel contributions should have a portfolio of patents characterized by distinctive claims across their patents (firms with a higher IPS). These distinctive claims can serve as indicators of a robust innovation pipeline within the company, implying a higher likelihood of maintaining a wide range of ideas, diversifying their innovation efforts across various projects, and sustaining strong momentum in future innovations. Conversely, companies whose patent claims start to resemble each other over time, (Firms with a lower IPS) may be seen as depleting their reserve of fresh ideas, narrowing their focus to small refinements of similar concepts, and thereby experiencing a slowdown in their innovation performance. While it's possible that

these companies are concentrating on refining a single successful innovation, the repetitive nature of their claims could also indicate a decreasing capacity for innovation, particularly in today's competitive environment where continuous innovation efforts are indispensable.

Figure 1 illustrates the effectiveness of IPS measure in tracking the innovation trajectories of three distinct firms: Nokia, IBM, and Tesla. The figure highlights that, during the mid-1990s, IBM and Nokia maintained a similar IPS. However, a noticeable shift occurred around 2005, with Nokia beginning to produce more similar patent claims, its IPS score starts decreasing. This change coincided with Nokia's loss of competitiveness in the smartphone market, a significant reduction in its market share, and a marked decline in its stock prices. In contrast, IBM, a company well-known for its consistent innovation and prolific patent output, maintained a stable IPS over time, along with an increase in its stock value. This sustained IPS suggests IBM's ongoing generation of unique patents. Meanwhile, Tesla, which entered our analysis after going public in the post-2010 period, significantly ramped up its innovation efforts, focusing on groundbreaking concepts in autonomous vehicles and next-generation batteries. This focus led to a substantial increase in Tesla's IPS, indicating a robust and increasingly distinctive innovation pipeline.

To thoroughly evaluate the effectiveness of our measure in predicting firms' future innovation success, we explore the relationship between Innovation Pipeline Score (IPS) and their future innovation output. Our analysis reveals that firms characterized by a distinctive set of claims, i.e., high IPS, not only produce a greater number of patents and garner more citations per patent but also generate patents with higher economic value per patent. Specifically, our findings show that a one standard deviation increase in IPS results in a 6.3% increase in the number of patents per employee, a 1.1% increase in patent value per patent, and an 0.2% increase in patent citations per patent over the subsequent year. Furthermore, these effects persist over the following five years. These results strongly support the assertion that our Innovation Pipeline Score (IPS) serves as a powerful tool for real-time assessment of firms' innovation potential. Firms with unique patents consistently exhibit

both economically and statistically superior future innovation performance.

Endogenous growth models suggest that firms sustain a healthy flow of innovative ideas when a skilled team of inventors works on a diverse set of ideas. This implies that the superior innovation productivity of high IPS firms should be supported by the skilled inventors they employ. To explore this, we compare the inventor characteristics of firms with varying IPS scores. We use detailed employee data from Revelio Labs, a leading provider of labor market analytics, which allows us to analyze and compare several key characteristics of inventors. Using this novel data, we find that firms with a more distinct set of claims (higher IPS) have a more gender diverse inventor team, higher percentage of PhD educated inventors, longer inventor tenure, inventors with more diverse educational backgrounds, higher percentage of scientists, more diverse skill sets. These findings support the notion that high IPS firms, compared to low IPS firms, have relatively more skilled inventor teams, which enables them to innovate more successfully.

We next analyze whether the successes of these high IPS firms in innovation are reflected in their economic advancements and growth, as measured by future operating performance. Models of endogenous growth predict a positive relationship between a firm's growth and its success in innovation; firms that excel in innovative productivity are poised to experience significant growth. Corroborating this notion, our data reveals compelling evidence that firms with a high IPS outperform counterparts with comparable claims in their pipelines in terms of future metrics such as return on assets (ROA), return on equity (ROE), operating cash flow (OCF), and return on net operating assests (RNOA). In particular, we find that a one standard deviation increase in IPS results in 0.34% higher ROA, 0.80% higher ROE, 0.44% higher OCF, and 0.98% higher RNOA. Furthermore, these results are economically and statistically persistent over the following five years. These results, combined with our findings of improved future innovation success, support the hypothesis that high IPS firms not only succeed in future innovation endeavors but also see this success translated into tangible economic gains and growth, as indicated by forthcoming operating performance.

Given that a more unique collection of claims has a beneficial impact on business profitability, we shift our focus to future stock returns to examine whether the market correctly assess the value of these innovative activities and update prices accordingly. We find that, depending on the performance measure, a one-standard-deviation rise in IPS is associated with an increase in (abnormal) annual return of 0.7% to 0.9%. We examine two possible explanations for this return predictability. On the one hand, utilizing more unique patents may need greater risk-taking behavior, which raises the cost of equity and, as a result, carries a premium. Alternatively, investors may underestimate the favorable effect of high IPS on firm fundamentals and hence fail to properly reflect its value into prices. We find evidence that supports the latter undervaluation argument. Our analysis demonstrates a negative relationship between IPS in a given year and earnings surprises in the following two quarters. This finding shows that investors' earnings expectations do not appear to properly account for the real economic benefit of high IPS as investors are surprised when the true financial performance of the company is revealed at the earnings announcement. Furthermore, a risk based explanation would imply a higher expected price volatility. However, contrary to this prediction, we find that high IPS firms exhibit lower price volatility, posing a further challenge to the risk-based explanation. These findings support the idea that the market underreacts to information embedded in the IPS about a company's potential innovation success and progressively adjusts stock prices to reflect this information.

Our paper contributes significantly to the research on firms' innovation activities and their economic impacts. Firstly, most prior studies overlook the context of the innovating firm and focus solely on the novelty of the innovation in the context of earlier patents, irrespective of the patent holder. We take a different approach by concentrating on the search process within a firm, examining the economic consequences of the richness of its innovation strategies.

Secondly, our research extends beyond the conventional focus on a firm's exploration of new knowledge, known as search scope, as the primary driver of patent originality. This view typically considers the diversity in technological fields and the technological distance

from previous patents as indicators of the breadth of new knowledge underpinning a patent's uniqueness. However, the extent to which a firm leverages its existing knowledge base, termed 'search depth', can also contribute to the originality of innovation, as suggested by Katila and Ahuja (2002). For instance, a firm specializing in finance might innovate various products within that domain, an aspect often overlooked by traditional metrics that categorize all patents under the same technological umbrella. Similarly, a company applying fundamental technology across different sectors may score high on traditional originality metrics due to cross-sectoral activities, even though this approach is essentially exploitative. Our measure differs by directly analyzing the content of patent claims and is conceptually designed to capture originality driven by the firm's search focus, depth, or a combination of both. Furthermore, the prior literature has identified several problems stemming from the use of technology classes for patent comparison purposes. These issues include outdated categories, an inability to adapt to the dynamic nature of technological evolution, inconsistencies between the technology class and the patent content, and retrospect updates in classifications . To address these shortcomings, we utilize semantic search with a BERT based model trained on a substantial dataset of US and European patents. This advanced approach goes beyond the constraints of technology classes and traditional bag-of-words strategies. It comprehends the context-driven implications of words and phrases, allowing a deeper analysis of patent claims. This, in turn, enables us to better evaluate a patent's originality by recognizing nuanced meanings in its claims, setting our metric apart from traditional classification-based methods.

Our findings also adds to the literature that study the determinants of firm growth. We show that that firms with a diverse array of unique claims excel operationally and enjoy significant growth compared to peers who mimic existing claims and lack in unique innovation. Importantly, our predictive metric for imminent innovation in firms is independent of external market evaluations like stocks prices that may swayed by investor sentiment or firms-specific attributes such as visibility. Our findings thus complements and extends , by offering a forward-looking measure of innovation based on a thorough analysis of patent

claims, independent of market valuations.

Finally, our research contributes to the understanding of how markets value firm innovation, noting the gap in investors' ability to fully appreciate various innovation aspects like R&D spending, its conversion to sales, innovation efficiency, and exploration. Unlike other measures that require comparison across disparate firms or deep understanding of citations and technology classifications, our metric focuses on the patent claims section, the most economically crucial part of patents. Firms are motivated to make these claims clear to investors, and a thorough examination of these claims can offer investors a deeper understanding of the innovation's potential. The fact that our metric significantly predicts future stock prices suggests the challenges in investor assessments of innovation might be more profound than previously understood.

2 Data, Construction of Innovation Pipeline Score , and Descriptive Statistics

In this section, we explain our selection of patent claims and outline the methodology for calculating our metric that measures the textual similarity of a firm's patents. We also offer illustrative examples to highlight its unique features and capabilities.

2.1 Data

Our analysis focuses on publicly traded U.S. firms that appear at the intersection of the USPTO PatentsView, Compustat, and CRSP (Center for Research in Security Prices) databases. We extract stock-level information (such as returns and volume) from CRSP and accounting data from Compustat. To minimize the risk of look-ahead bias in accounting measures, we match monthly CRSP data with annual accounting data, ensuring a minimum six-month lag. This primary dataset comprises common shares (share codes 10 or 11) from firms listed on NYSE, NASDAQ, and AMEX. We exclude utility and financial firms, years with

negative book equity, and firm-month observations with stock prices under $1. Additionally, we require a firm to have been listed in Compustat for a minimum of two years before including it in our sample. We complement this dataset with information on institutional ownership from Thomson Reuters 13 Filings, analyst forecasts from the Institutional Brokers' Estimate System (I/B/E/S), and work force details from Revelio Labs, a leading provider of labor market analytics.

We extract all patent claims from the USPTO PatentsView database for patents issued between 1976 and 2020, consolidating each independent claim at the patent level. To match patents to their respective firms, we use the Kogan et al. (2017) (KPSS) dataset, which allows us to compile a sample of patents owned by public companies.[1] We then calculate the firm-level Innovation Pipeline Score (IPS) using this patent-related data, as elaborated in Section 6. Given that numerous patent-based metrics in our analysis require a minimum of five years of patent data, our final combined sample covers the period from 1982 to 2020.

2.2 Construction of Innovation Pipeline Score (IPS)

The initial step in creating a patent originality metric involves selecting pertinent information from patent documents. Each patent provides in-depth details about the invention, its technological area, inventors, the assignee, citations, and more. The specifics about the invention are mainly in textual form, often supplemented by drawings, and are divided into two sections: descriptions and claims.

Patent descriptions must provide enough information to replicate the technology, often encompassing prior art discussions and inventive process descriptions. This results in significant overlaps with descriptions in earlier patents (Ashtor, 2019). In contrast, claims are the only elements in a patent document that are uniquely distinct in each patent (Ashtor, 2019). Their uniqueness is crucial for several reasons: They legally define an invention and its technological boundaries, specifying the patent's scope and the exclusive rights associated

[1]The KPSS dataset includes a link between each patent and its corresponding CRSP "PERMNO" identifier. We are grateful to Kogan et al. (2017) for making their updated data accessible at GitHub.

with it.[2] Claims are also pivotal in enforcing patents and are central in infringement cases, underpinning the patent holder's legal rights. Additionally, the breadth and clarity of claims are crucial during the patent examination process, helping to determine the invention's novelty and compliance with patentability criteria. Claims are thus the most dependable means of determining the uniqueness of patents

The USPTO distinguishes between independent and dependent claims. An independent claim is a standalone claim without any reference to another claim and outlines all essential features and limitations necessary to define an invention. They are the core claims that define the boundary of patent protection. In contrast, a dependent claim builds on a previous claim by adding further limitations or specifics. Thus, dependent claims are narrower in scope compared to independent claims. To ensure comprehensive protection, a patent may have multiple independent claims that cover various aspects of the invention. A typical patent in our sample has three independent claims. These claims are usually arranged in order of scope with the first one being the broadest. Subsequent independent claims introduce additional unique features or applications of the invention, yet each is self-contained and independent of the others. In our analysis, we employ the text from all independent claims to construct our metric. However, our results are robust when focusing solely on the first independent claims of each patent.

We employ a state-of-the-art machine learning technique based on Bidirectional Encoder Representations from Transformers (BERT) for the analysis of patent claims. Unlike traditional approaches like the bag-of-words, which ignore the semantic and syntactic relationships between words, BERT comprehensively captures the context-specific meanings of words and phrases. This capability provides a more nuanced and precise analysis of patent claims. The bidirectional nature of BERT enables it to interpret the context of each word in relation to its surrounding words, both before and after, ensuring more accurate meaning comprehension. This aspect is particularly vital in patent analysis, where maintaining

[2]Specifically, there are independent claims, which highlight the unique features of the technology, and dependent claims, which relate to and elaborate on the independent claims.

semantic relationships across words is crucial.

As highlighted by Ashtor (2019), common NLP techniques frequently encounter limitations when applied to a broad range of technologies.. For instance, the word "cell" has different meanings in biology and computer science. Similarly, the term "receiver" in electronics might refer to completely different devices now compared to those in early radio patents, despite the fundamental concept remaining unchanged. To address this complexity, our method leverages the advanced capabilities of BERT, specifically a model trained on a comprehensive corpus of US and European patents. This use of a patent-trained BERT model allows us to discern the context-specific nuances of terms within their respective technological contexts and to ensure a more accurate interpretation of patent claims.

Furthermore, we employ Sentence-BERT (SBERT), an adaptation of BERT tailored for assessing textual similarity for long texts (Reimers and Gurevych, 2019). While BERT focuses on word-level context, SBERT extends this capability to the sentence and paragraph levels. This adaptation is crucial for effectively capturing the overall meaning of the lengthy and complex texts typical in patent documents. SBERT's design facilitates a more efficient and semantically rich analysis of patent claims, accurately reflecting their technological context. However, comparing patents across diverse technological domains within a firm would be problematic. For instance, comparing Google's patents in advertising technology with those in autonomous driving would not accurately reflect the firm's overall innovation capabilities.

To address this issue, we establish 132 technology clusters, aligning with the USPTO's number of technology classes, facilitating more meaningful patent comparisons. This approach mitigates the shortcomings of USPTO's technology classifications discussed later in Section 2.4. We utilize the FAISS k-Means algorithm to partition patents into these 132 clusters based on the cosine similarity between patent embeddings. FAISS (Facebook AI Similarity Search) is a highly efficient library designed for similarity search and clustering of dense vectors. The algorithm iteratively adjusts cluster centroids to minimize within-cluster variance, ensuring high semantic proximity among grouped patents. This process is repeated annually to capture

shifts in technological focus over time as new patents are added and clusters are updated.

For any given patent j granted to firm i in year t, we assign it to a cluster k. We then compute the cosine distance (as opposed to similarity) between patent j and all other patents granted to firm i within the previous five years in cluster k, including year t. This process gauges the semantic distinctiveness of patent j from the firm's earlier patents in the same technology class.

To illustrate our methodology, consider a hypothetical scenario where firm i holds ten patents $(P_{1,k}, ..., P_{10,k})$ issued from 2015 to 2020, all of them within cluster k. Additionally, in 2020, this firm is granted two additional patents in the same cluster, $P_{1,k,2020}$ and $P_{2,k,2020}$. To calculate the Innovation Pipeline Score (IPS) for firm i in 2020, we analyze the distances of both $P_{1,k,2020}$ and $P_{2,k,2020}$ from all earlier patents issued to the firm since 2015 within cluster k.

If we denote the distance between P_A and P_B as $d_{P_A - P_B}$, the individual Patent Uniqueness Scores, $IPS_{P_{1,2020}}$ and $IPS_{P_{2,2020}}$, are determined by calculating the mean of these distances:

$$IPS_{P_{1,k,2020}} = \text{Mean}\left([d_{P_{1,k,2020}-P_{1,k}}, d_{P_{1,k,2020}-P_{2,k}}, \ldots, d_{P_{1,k,2020}-P_{10,k}}]\right)$$

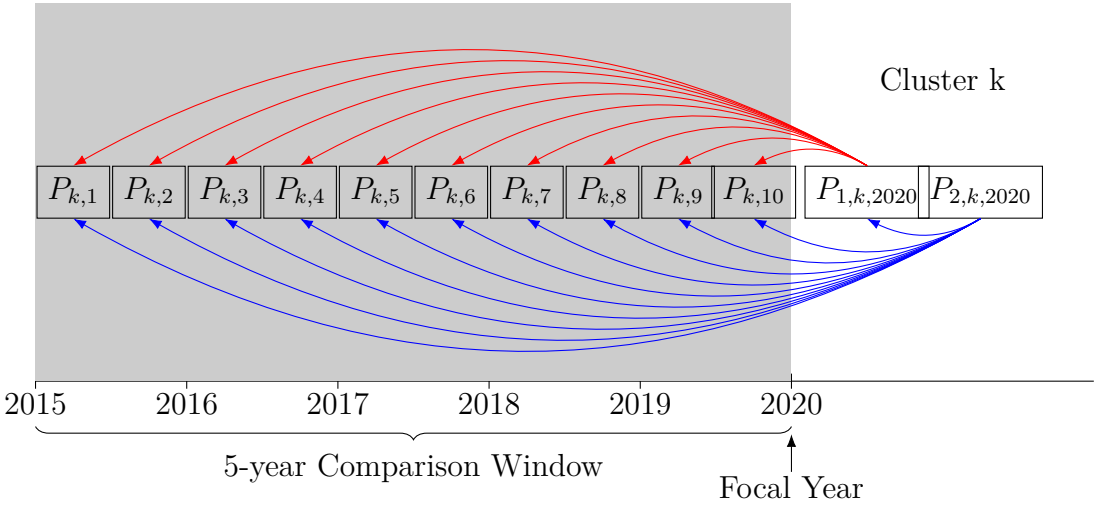
$$IPS_{P_{2,2020}} = \text{Mean}\left([d_{P_{2,k,2020}-P_{1,k}}, d_{P_{2,k,2020}-P_{2}}, \ldots, d_{P_{2,k,2020}-P_{10},k}, d_{P_{2,k,2020}-P_{1,k,2020}}]\right)$$

The overall IPS for firm i in 2020 is obtained by averaging these individual IPS scores. Additionally, we compute the distance of patent j to patents in class k across all firms over the last five years, providing a measure of its uniqueness within the broader technology class.

By utilizing distance measures instead of similarity, our IPS measure provides a more nuanced measure of innovation distinctiveness. Firms with higher average IPS are identified as having uniquely innovative patents, suggesting greater potential for future innovation. Conversely, firms with lower average IPS exhibit more similar innovation patterns, potentially indicating a decline in their innovation pipeline.

$$\textbf{IPS}_{i,2020} = Mean\left([IPS_{P_{1,2020}}, IPS_{P_{2,2020}}]\right)$$

Figure 1: Timeline for the IPS computation example



2.3 Descriptive Statistics

Figure 2 displays the distribution of our firm-level IPS measure. The average Innovation Pipeline Score across firms stands at 0.31, with the 5th and 95th percentiles being 0.16 and 0.49, respectively. The distribution demonstrates a right-skewness,characterized by a longer tail. This indicates that firms tend to develop new technologies by adapting or evolving their existing ones. The Cumulative Distribution Function (CDF) reveals that the likelihood of a score falling aboe 0.5 is quite low, highlighting a tendency for patents to be more similar than distinct. This pattern is in line with observations that many patents demonstrate low levels of subsequent innovation and receive few citations (Kogan et al., 2017). Additionally, this finding is consistent with the theory that generating innovative ideas is becoming increasingly difficult (Bloom et al., 2020). Corroborating this, Figure 3 indicates a decrease in the mean firm-level IPS to 0.4 in the period post-2000, suggesting a growing trend of technological similarity over time.

Table 1 presents a summary of the statistics for firm characteristics, along with their

correlations with IPS (Innovation Pipeline Score). IPS assesses a firm's patents each year against its patents from the previous five years. Consequently, our sample is limited to firms engaged in innovation activities within a specific year, resulting in a relatively smaller sample size. On average, our sample includes 1053 firms annually. Despite this, our sample accounts for over 68% of the total U.S. market capitalization, as firms with patent activities tend to be larger. Examining the correlations, we find that firms with less similar patent portfolios are generally larger, more established, more liquid, and exhibit a lower R&D expenditure-to-asset ratio.[3] Nonetheless, our IPS measure does not show a strong overall correlation with other firm characteristics.

2.4 Comparison of IPS and USPTO Technology Classes

Our approach differs from traditional social science research, which often relies on patent technology classes to understand the breadth of knowledge in patents. This classification system, while useful, has limitations due to the rapid and interdisciplinary nature of technological evolution. Patents spanning multiple fields make class assignments complex. For example, a recent application of machine learning by USPTO to classify artificial intelligence (AI) patents reveals their spread across various technology classes.[4] Similarly, Lerner et al. (2023) use machine learning to classify "financial" patents, revealing the constraints of technological classifications.[5] Furthermore, the pace of technological progress, especially in fields like AI and nanotechnology, often exceeds the classification system's adaptability, causing mismatches between classifications and actual patent content. Modifications in the classification system, such as the comprehensive 2013 revamp by the USPTO and EPO, can lead to the retroactive reassignment of patents to new classes, adding complexity

[3]The relatively lower R&D expenditure is somewhat unexpected and may be influenced by the size of the firms' assets.

[4]Online Appendix Table 1 displays the top fifteen Cooperative Patent Classification (CPC) subclasses identified in the analysis. G06F is the leading CPC subclass for AI patents, accounting for 423,397 patents. It is closely followed by a diverse range of subclasses, including H04L with 168,101 patents and A61B with 102,992 patents.

[5]Lerner et al. (2023) find that while most finance patents fall under G06Q, many related to blockchain and cryptocurrency were categorized under H04L, underscoring the need for a more refined approach.

to technology-based patent analysis over time.[6] Additionally, the broad nature of some classifications can dilute the comparison of patent originality.

Our methodology reveals unique insights that might be overlooked by standard classification. For example, it might be intuitive to assume that those within the same USPTO classification are exploitative, potentially indicating lower economic value. However, this can be misleading. Take the example of Amazon's patent 8,078,640 on high-efficiency binary encoding, classified under USPTO 707. Granted on December 13, 2011, this patent, filed on December 28, 2010, was the 67th in a series that Amazon had been developing since March 2010, with initial applications dating back to 2006. Despite being classified in a category with numerous preceding patents, this particular patent exhibits only a 40.7% similarity to Amazon's patents from the past five years. This places it in the fifth percentile of the IPS distribution, emphasizing its distinctiveness.

Moreover, the distribution of patents across different USPTO technology classes does not necessarily equate to heightened originality in a firm's innovation process. For example, despite Nokia's increased exploration in different USPTO classes post-2000, our analysis shows a rise in patent similarity from 2005. This suggests a growing homogeneity in Nokia's patents, coinciding with their declining competitive advantage in the smartphone market. In the Online Appendix, we present two Nokia patents from 2008; though classified differently, they share a significant similarity score of 0.80, highlighting the limitations of traditional classification systems in capturing true innovation diversity.

3 IPS and Future Innovation Performance

In this section, we assess how effectively our Innovation Pipeline Score (IPS) identifies firms with strong innovation pipelines and a high potential for successful innovation. We

[6]Even within the same classification system, technology classes can change over time. For example, in March 2005, USPTO announced a change in Class 260, stating, "It was decided that this class needed to be reclassified because its concepts did not necessarily address new technology and several of its subclasses were too difficult to search because of their size."

argue that the degree of similarity or divergence in the patent claims of firms can provide valuable insights into their potential for sustained innovation success. Initially, it is, however, unclear whether a lower or higher IPS indicates a greater likelihood of future innovation success.

On one hand, a low IPS, indicating unique patent claims, might correlate with a firm's innovative capabilities. These distinctive claims could indicate a strong pipeline characterized by a diverse and fresh range of ideas, suggesting a greater potential for maintaining innovation momentum. In contrast, a high IPS, pointing to similar claims, could suggest that a firm is reaching its creative boundaries, repeating similar patents due to a lack of novel ideas. This would imply a negative relationship between IPS and a firm's future innovation performance.

On the other hand, similarity in patent claims might indicate a strategic focus, potentially leading to increased innovation within that specialized domain. Firms adopting this strategy might consistently produce similar patents, dominate their market niche, and sustain successful innovation. This line of reasoning implies a positive relationship between IPS and a firm's future innovation performance.

To empirically investigate these arguments, we first examine whether IPS measure possesses predictive information about future innovation performance. Our approach employs both quantitative and qualitative measures of future innovation performance, including the number of granted patents and the private and scientific value of these patents. We utilize cross-sectional regressions of firms' future innovation, defined as the 1-year-ahead count of patents per employee (*PatentProd*) to capture the quantity of innovation, the 1-year-ahead total patent value per patent by using patent value estimates from Kogan et al. (2017) (*PatentValue*) to capture the private value of the innovation, or subsequent year total citations per patent (*Citations*) to capture the scientific value of the innovation, on our constructed Patent Similarity Score (*IPS*) as well as a vector of other control variables (X):

$$Innovation_{i,t+1} = \beta \times IPS_{i,t} + \gamma \times X_{i,t} + \delta_j + \pi_t + \epsilon_{i,t}$$

where, the unit of observation is at the firm i and year t level, with j indicating Fama-French industries for industry fixed effects.

Following the prior literature, we include a variety of control variables that have been found to be significant predictors of future innovation performance and also could be correlated with our IPS measure. Specifically, we added book-to-market ratio (bm), capital expenditure (CapEx), research and development expenses (R&D), number of prior granted patents (Patents), firm age (Age), patents to assets ration (PTA), advertising expenses (Advertising), 12 month returns (12m-return). As our sample period covers until 1976-2020, we used CPC classification to compute these measures. We also include Fama-French 48 industries and year fixed effects. To mitigate the outliers effect, all variables are winsorized at the 1% and 99% level. We also standardize all independent variables to zero mean and one standard deviation.

Table 2 shows that IPS measure is a significant predictor of future innovation productivity. Column 1 shows that one standard deviation decrease in patent similarity measure yields 6.3% more patents per employee in the next year after controlling for possible confounding factors.[7] Similarly, our measure has a significant predictive power on future patent values and citations. Economically, one standard deviation decrease in IPS increases next year's patent value per patent by 1.1% per patent and citation per patent by 0.2%.[8]

Our analysis extends beyond immediate future innovation, probing the predictive capacity of the IPS over a more extended horizon. As illustrated in Table 3, the IPS not only maintains its significance but also demonstrates an increasing economic magnitude when predicting future innovation outcomes over the next five years. This trend is particularly pronounced in the contexts of patent value and citations, offering compelling evidence that the IPS is a strong predictor of innovation quality.

In the literature, several innovation originality measures has been proposed to examine

[7]The point estimate in Column (1) of 0.05 is taken and normalized as *PatentProd* variable is logged. $e^{0.061} - 1 = 6.3\%$

[8]The point estimates in Column (3) and (4) are taken.

whether investors understand firms innovation strength and timely update the prices accordingly. For example, common innovation originality measures of Hirshleifer, Hsu, and Li (2018) and exploratory focus measures of Fitzgerald et al. (2021), Duysters et al. (2020), and Jaffe (1989) have been used to examine the relation between firms innovation and future stock prices. However, these studies do not examine the relation between their measure and future innovation productivity. Nonetheless, to ensure that our results are robust controlling for these alternative measures, we repeat our analysis controlling for these measures. According to the results in Table 4 , the relation between IPS and future innovation innovation measures is statistically and economically robust to controlling for these other measures. This finding shows that IPS effectively captures a unique dimension of strength in firms innovation pipelines and explains future innovation capacity beyond what can be explained with the existing measures.

Overall, the findings from Tables 2 and 3 reveal that firms with a more unique set of patent claims, as indicated by a higher IPS, are likely to be more innovative in the future. This underscores the effectiveness of IPS, our novel claim similarity-based metric, in identifying firms with higher innovation potential. This insight is particularly significant in the context of the dynamic nature of innovation and competitive market environments. Additionally, the growing economic significance of patent value and citations further confirms that IPS captures a subtle yet crucial aspect of innovation, closely linked to long-term value creation and market impact.

4 IPS and Firms' Innovation Team Characteristics

Our findings in previous section suggest that IPS is a strong predictor of future innovation performance. Hence, firms that have a higher amount of distinct claims are also the ones that successfully stay innovative in the future. However, achieving such a strong innovation pipeline would necessitate a strong innovation team in the firms' labor force. In other words,

if IPS accurately reflects a firm's ability to innovate, then the innovators' characteristics of firms with higher IPS should differ from those with lower IPS, with these characteristics being more conducive to stronger innovation.

To examine innovators' characteristic, we use employee data from Revelio Labs, a leading provider of labor market analytics. Revelio Labs aggregates information from over 380 million online public profiles and resumes of employees from more than 5,000 public companies, sourced from various job-related websites and social media platforms such as LinkedIn. This rich dataset includes employee gender, job tenure, education levels and fields, job categories (e.g., administrative vs. scientific roles), educational prestige (scored between -1 and 1 based on schools attended), and skills aggregated into 25 categories.

We matched this data with our innovator-firm sample for each firm-year observation. The matching process involved two steps: first, matching Revelio Labs firms with our sample firms, and second, matching innovators with Revelio Labs employees based on textual similarity between employee names in Revelio Labs and USPTO innovator names for the same firm-year. From this matched dataset, we extracted several key metrics that we believe can represent skill set of innovators. For example, we believe a better innovation necessitates a more inclusive team characteristic, as captured by innovator gender equality. Also, firms that creates and innovation environment that attracts and retains their employees for a long amount of time, as capture by employee tenure and scientist tenure, should have a higher innovation capacity. Furthermore, firms with higher level of innovator education, as captured by PhD percentage in the innovation team,and a higher diversity in education background of innovators, education degree diversity, firms that have higher percentage of scientist in their work force and a larger innovation team size, and skill set varity should be superior innovators.

In order to examine how innovator team characteristics by IPS, we group firms into five groups based on our IPS measure. In order to account for differences of team characteristics due to firm size or industry differences, we also use adjustments based on size and Fama-French

12 industry portfolios. According to the results in Table 6 firms with more distinct claims (high IPS) have higher female representation, longer employee tenure, a greater proportion of PhDs, more diverse educational backgrounds, a higher percentage of scientists, more varied skills within the inventor team, and larger innovation teams. These findings are inline with our conjecture that firms with higher IPS have innovation team characteristics that would support a better innovative pipeline. Overall, the results in Table 6 indicate that firms with higher IPS do, in fact, have a skilled team of innovators, which contributes to higher innovation productivity.

5 IPS as a Predictor of Future Profitability

In this section, we explore whether firms with high IPS also exhibit higher operational performance in the future. Theories of endogenous growth, such as those proposed by Romer (1990), Aghion and Howitt (1992), and Kogan et al. (2017), suggest that advancements at the technological frontier and efficient resource reallocation are key drivers of economic growth. Furthermore, Akcigit and Kerr (2018) shows that similar innovation in a technology cluster run into diminishing returns. These theories highlight the pivotal role of a firm's innovation success, as predicted by high IPS, in determining its growth and operational efficiency. Hence, in line with these endogenous growth models, IPS could also serve as a valuable forward-looking indicator of a firm's overall growth trajectory and operational success.

To test this prediction, we examine the relation between IPS and future operating performance. Specifically, we run the following cross-sectional regressions of an individual firm's future profitability, defined as either 1-year-ahead return on assets (ROA), 1-year-ahead return on equity (ROE) or 1-year-ahead operating cash flow (OCF), on a firm's patent similarity score (IPS) score as well as a vector of other control variables (X):

$$Profitability_{i,t+1} = \beta \times IPS_{i,t} + \gamma \times X_{i,t} + \delta_j + \pi_t + \epsilon_{i,t}$$

where, the unit of observation is at the firm i and year t level, with j indicating Fama-French industries for industry fixed effects. Similar to the previous section, we include a variety of control variables that have been found to be significant predictors of future operating performance (Fitzgerald et al., 2021; Hirshleifer, Hsu, and Li, 2018). We also include Fama-French 48 industries and year fixed effects. To mitigate the outliers effect, all variables are winsorized at the 1% and 99% level. We also standardize all independent variables to zero mean and one standard deviation.

Table 7 confirms our prediction. We find that there is a positive and significant relation between IPS and future operating performance of the firms. This is also economically significant as Column 1 shows that one standard deviation decrease in IPS associated with 0.34% increase in ROA after controlling for common factors and previous innovation originality and search measures. This is comparable to 0.16% estimate for exploitative patent ratio and 0.45% estimate for internal search proximity results in the literature Hirshleifer, Hsu, and Li (2018) and Fitzgerald et al. (2021). Similarly, Column 2 indicates one standard deviation decrease in IPS is associated with 0.80% increase in ROE, Column 3 shows 0.44% increase in OCF, and Column 4 shows 0.98% increase in RNOA. Furtherore, Table 9 shows that the result is statistically and economically robust to controlling for other common innovation originality measures of Hirshleifer, Hsu, and Li (2018) and exploratory focus measure of Fitzgerald et al. (2021), Duysters et al. (2020), and Jaffe (1989).

Finally, similar to our earlier observations on innovation, Table 8 demonstrates that the predictive power of the IPS with respect to operating performance extends over a significant duration. Specifically, the results indicate that the IPS's predictive power regarding future profitability metrics— ROA, ROE, OCF, and RNOA—remains both economically and statistically robust over a five-year span. This longitudinal consistency underscores the enduring relevance of the IPS as a predictor of future operating performance. Overall, our

findings in Sections 3 and 5 suggest that IPS is a significant predictor of future innovation success and this predictive power extends to the future operating performance of firms as well.

6 IPS and the Cross-section of Stock Returns

Our findings indicate that IPS offers crucial information to investors about a firm's future innovation and operational performance. The next logical step involves exploring how the market values this information and determining if it is reflected in stock prices. To achieve this, we will now turn our attention to stock prices and investigate whether the IPS can predict the cross-section of stock returns.

6.1 Future stock performance

We conduct a regression analysis to explore the relationship between the IPS and future stock returns. We regress monthly stock returns onto IPS from the previous year and a comprehensive set of controls known to predict stock returns. To ensure that our results remain unaffected by well-known stock characteristics and risk factors, we adjust returns using several methods. We use characteristic-adjusted returns developed by Daniel, Grinblatt, Titman, and Wermers (1997) ,referred to as DGTW returns, and risk-adjusted returns based on CAPM, Carhart's four-factor and Fama-French's six-factor models as our dependent variables.

To compute DGTW returns, we subtracted each firm's raw return from the return of a benchmark portfolio with similar size, market-to-book ratio, and momentum quintile. For risk adjustment, we regressed stock excess returns on factors over the past 36 months (with at least 24 months of valid returns) at the end of each month t and save the factor loadings. Next, to find the risk-adjusted return for each month t, we take the stock's excess return and subtract it from the required return implied by the factor model. The required return is equal

to the sum of the factor loadings times the corresponding factor realization. All independent variables in our analysis were winsorized at the 1% and 99% levels and standardized to have a mean of 0 and a standard deviation of 1. This standardization was performed to compare economic values across different variables and specifications more effectively.

Table 10 presents the results. Our analysis reveals a significant positive relationship between IPS and future stock returns after accounting for various well-established predictors of stock returns. Depending on the specific model and return adjustment method used, a one standard deviation increase in IPS is associated with a noteworthy increase in annual returns ranging between 65 and 94 bps. The inclusion of InnOrig and ExpFocus do not significantly affect the predictive power of IPS. These findings suggest that firms with a more unique set of claims (indicated by a high IPS) tend to have higher future returns, in addition to demonstrating superior innovation and operational performance. The observation that IPS predicts both future profitability and stock returns raises an important question: Does this predictive relationship reflect rational pricing, compensating for risk, or is it a result of irrational mispricing? Such a distinction requires further analysis, which we undertake in the next section.

6.2 Underlying mechanism: Mispricing or Risk?

The risk-based explanation, rooted in rational behavior, posits that high future innovation success, as indicated by a high Patent Similarity Score (IPS), may be linked to unknown risk factors associated with firms exhibiting higher expected productivity (Kogan et al., 2017). In this view, current stock prices of firms with high IPS levels are not mispriced. Instead, the higher average future returns observed for these firms are interpreted as a compensation for the greater risk inherent in their more distinctive set of claims.

Alternatively, the positive relationship between IPS and stock returns could be due to irrational mispricing related to a firm's IPS score. In this scenario, investors fail to fully comprehend IPS's information content for a firm's innovation and operational performance.

As a result, they underreact to positive (negative) signals conveyed by a higher (lower) IPS. This underreaction could lead to a delayed adjustment of the stock price in response to the information provided by the IPS. Consequently, stock prices are likely to continue moving in the direction suggested by IPS regarding future profitability until this underreaction is eventually corrected.

We conduct three distinct analyses to test our hypothesis. First, as shown in Figure 4, we find that the return predictability from the IPS lasts for six months before becoming insignificant. This suggests that, unlike the lasting impact of innovation on productive performance, the influence of IPS on returns is shorter-lived, with the market quickly processing the information and adjusting prices accordingly. This aligns with the underreaction argument, as posited by Chambers, Jennings, and Thompson (2002), and Hirshleifer, Hsu, and Li (2018), who assert that mispricing-based return predictability should be rectified swiftly. However, a risk-based explanation would imply longer-term return predictability, in line with the duration of innovation success.

Secondly, considering that investors have limited attention and resources to evaluate every firm, we hypothesized that if mispricing is the cause, return predictability would vary based on stock visibility and complexity in valuation. We use three key proxies to measure investor attention and information uncertainty: firm size, return volatility, and the number of analysts following. Smaller firms, those with higher return volatility, and those with less analyst coverage are often less visible and harder to value (Zhang, 2006). Therefore, we would anticipate stronger underreaction or information asymmetry among these firms. To investigate this, we categorize stocks into two groups based on firm size, volatility, or analyst coverage. For each criterion, the top and bottom groups were formed from the upper and lower 30% of stocks, respectively, based on that specific measure. We then replicated the analysis in Table 10 using this subset of stocks. Our findings, reported in Table 11 using both raw returns and risk-adjusted returns based on Carhart's four-factor model, reveal that the relationship between IPS and future returns is significantly more pronounced in smaller firms,

those with higher return volatility, and those with lower analyst coverage. These results suggest that the visibility of a firm and the uncertainty in its valuation significantly influence our results, supporting the mispricing hypothesis and challenging a risk-based explanation.

Finally, we delve into the relationship between the Patent Similarity Score (IPS) and future earnings surprises. According to a risk-based explanation, rational investors should anticipate future profitability and stock returns as indicated by the IPS and thus should not be caught off guard when a firm achieves this projected performance. Conversely, the undereraction argument suggest that investors might not fully process the predictive information of the IPS, leading them to underestimate the higher profitability it signals, as highlighted by La Porta et al. (1997). This scenario would result in investors being unexpectedly surprised with the subsequent earnings performance following a high or low IPS.

To formally test this hypothesis, we conduct a regression analysis of realized earnings surprises for each of the four quarters in year $t+1$ against the IPS from year t, alongside lagged earnings surprises and other relevant firm-level controls. We utilize two approaches to measure earnings surprises: analyst forecasts and return-based measures.

The analyst forecast-based measure, referred to as SUE (Standard Unexpected Earnings), is calculated as the actual earnings per share (EPS) minus the expected EPS, adjusted by the stock price. The expected EPS here is defined as the median of analysts' forecasts in the 90 days leading up to the earnings announcement in each quarter, as outlined by Livnat and Mendenhall (2006). The return-based measure, termed CAR (Cumulative Abnormal Returns), captures the cumulative abnormal returns from the day of the announcement (day 0) to one day after (day 1).

To measure abnormal returns around earnings announcements, we applied two methodologies. The first method involves adjusting daily stock returns using returns from six portfolios categorized based on size and book-to-market (BM) ratios (2x3 sorting). The second method employs a one-factor model (CAPM), using daily returns from the month preceding the announcement to estimate the coefficients, which are then used to calculate daily abnormal

returns surrounding the earnings announcement dates.

Table 12 reports our estimates. All independent variables are winsorized at the 1% and 99% levels and standardized to have a mean value of 0 and standard deviation of 1. The results show a positive significant relationship between IPS and earnings surprises. Notably, a one standard deviation increase in IPS results in a 9.4 basis point increase in the two-day CAR based on CAPM. These results complement our earlier findings, where we reported a 37.6 basis point increase in quarterly FF6 alpha for each standard deviation increase in IPS in Table 11. This suggests that approximately all of the quarterly abnormal returns are realized around the time of earnings announcements.

Overall, our analysis indicates that firms characterized by a more unique set of claims, as denoted by a high IPS, are associated with higher future returns. This outcome remains consistent across various firm-specific and risk-based controls. Furthermore, our findings point towards the return prediction being primarily driven by investor underreaction to the information provided by IPS, rather than being attributable to a risk-based explanation. This highlights the critical role of investor behavior in interpreting and reacting to IPS-related information.

7 Conclusion

We introduce a new metric to assess the quality of a company's pipeline of innovative ideas and its ability to innovate effectively. Utilizing the Bidirectional Encoding Representation from Transformers (BERT), we generate high-dimensional embeddings that capture the semantic content of patent claims. This metric evaluates the uniqueness of the claim sections of patents, offering a quantitative assessment of the quality and potential of a company's current innovative initiatives. We suggest that companies with the capacity to generate novel ideas and make significant contributions are likely to have a patent portfolio marked by unique claims (reflected in a higher Innovation Pipeline Score, or IPS). These distinctive claims

indicate a strong innovation pipeline, suggesting a broader spectrum of ideas, diversification across various projects, and sustained momentum in future innovations.

We use our measure to empirically examine the relation between our measure and firms' future innovation success. Our findings demonstrate that a higher IPS, indicative of more distinct patent portfolios, is associated with higher future innovation outputs, including an increase in the number of patents, their economic value, and citations. This validates our novel measure in identifying firms with a strong pipeline for future innovations, underscoring the importance of distinctiveness in patent claims as a predictor of sustained innovation success.

We also demonstrate that firms innovator team characteristics that would support a better innovative pipeline. These firms have relatively higher female representation, longer employee tenure, higher proportion of PhDs, more diverse educational backgrounds, higher percentage of scientists, more skills variety within the innovator team, and larger innovator teams.

Furthermore, our research sheds light on the economic implications of innovation, showing that firms with higher IPS not only excel in innovation metrics but also translate their innovative capabilities into superior operational and economic performance, such as ROA, ROE, OCF, and RNOA. However, despite the predictive power of IPS, our analysis reveals that investors underreact to IPS information and gradually incorporate it into stock prices as we find a significant negative relationship between IPS and future stock returns.

Our findings suggest that firms with a strong pipeline of patents with distinct claims tend to stay innovative in the future and turn this innovative efforts to operation profitability. Hence, our measure successfully provides a metric to identify future innovators. These findings are inline with the endogenous growth models that future innovation can be predicted by strength of current portfolio of innovative ideas.

References

Acemoglu, Daron, Ufuk Akcigit, and William R. Kerr (Oct. 11, 2016). "Innovation Network". In: *Proceedings of the National Academy of Sciences* 113.41, pp. 11483–11488. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1613559113.

Aghion, Philippe and Peter Howitt (1992). "A Model of Growth through Creative Destruction". In: *Econometrica* 60.2, pp. 323–351. DOI: 10.2307/2951599.

Akcigit, Ufuk and William R. Kerr (Mar. 20, 2018). "Growth through Heterogeneous Innovations". In: *Journal of Political Economy* 126.4, pp. 1374–1443. ISSN: 0022-3808. DOI: 10.1086/697901.

Arrow, Kenneth (1962). "Economic Welfare and the Allocation of Resources for Invention. The Rate and Direction of Inventive Activity: Economic and Social Factors". In: *N. Bureau*.

Ashtor, Jonathan H. (2019). "Investigating Cohort Similarity as an Ex Ante Alternative to Patent Forward Citations". In: *Journal of Empirical Legal Studies* 16.4, pp. 848–880. ISSN: 1740-1461. DOI: 10.1111/jels.12237.

Bernstein, Shai (2015). "Does Going Public Affect Innovation?" In: *The Journal of Finance* 70.4, pp. 1365–1403. ISSN: 1540-6261. DOI: 10.1111/jofi.12275.

Bloom, Nicholas et al. (Apr. 2020). "Are Ideas Getting Harder to Find?" In: *American Economic Review* 110.4, pp. 1104–1144. ISSN: 0002-8282. DOI: 10.1257/aer.20180338.

Chambers, Dennis, Ross Jennings, and Robert B. Thompson (June 1, 2002). "Excess Returns to R&D-Intensive Firms". In: *Review of Accounting Studies* 7.2, pp. 133–158. ISSN: 1573-7136. DOI: 10.1023/A:1020217817156.

Daniel, Kent et al. (1997). "Measuring Mutual Fund Performance with Characteristic-Based Benchmarks". In: *The Journal of Finance* 52.3, pp. 1035–1058. ISSN: 1540-6261. DOI: 10.1111/j.1540-6261.1997.tb02724.x.

Duysters, Geert et al. (Oct. 2020). "What Drives Exploration? Convergence and Divergence of Exploration Tendencies among Alliance Partners and Competitors". In: *Academy of Management Journal* 63.5, pp. 1425–1454. ISSN: 0001-4273, 1948-0989. DOI: 10.5465/amj.2017.1409.

Fitzgerald, Tristan et al. (Feb. 2021). "Innovation Search Strategy and Predictable Returns". In: *Management Science* 67.2, pp. 1109–1137. ISSN: 0025-1909. DOI: 10.1287/mnsc.2019.3480.

Giczy, Alexander V., Nicholas A. Pairolero, and Andrew A. Toole (Apr. 1, 2022). "Identifying Artificial Intelligence (AI) Invention: A Novel AI Patent Dataset". In: *The Journal of Technology Transfer* 47.2, pp. 476–505. ISSN: 1573-7047. DOI: 10.1007/s10961-021-09900-2.

Hirshleifer, David, Po-Hsuan Hsu, and Dongmei Li (Mar. 1, 2013). "Innovative Efficiency and Stock Returns". In: *Journal of Financial Economics* 107.3, pp. 632–654. ISSN: 0304-405X. DOI: 10.1016/j.jfineco.2012.09.011.

— (July 1, 2018). "Innovative Originality, Profitability, and Stock Returns". In: *The Review of Financial Studies* 31.7, pp. 2553–2605. ISSN: 0893-9454. DOI: 10.1093/rfs/hhx101.

Jaffe, Adam B. (Apr. 1, 1989). "Characterizing the "Technological Position" of Firms, with Application to Quantifying Technological Opportunity and Research Spillovers". In: *Research Policy* 18.2, pp. 87–97. ISSN: 0048-7333. DOI: 10.1016/0048-7333(89)90007-3.

Jones, Benjamin F. (Jan. 2009). "The Burden of Knowledge and the "Death of the Renaissance Man": Is Innovation Getting Harder?" In: *Review of Economic Studies* 76.1, pp. 283–317. ISSN: 00346527, 1467937X. DOI: 10.1111/j.1467-937X.2008.00531.x.

Katila, Riitta and Gautam Ahuja (2002). "Something Old, Something New: A Longitudinal Study of Search Behavior and New Product Introduction". In: *The Academy of Management Journal* 45.6, pp. 1183–1194. ISSN: 0001-4273. DOI: 10.2307/3069433. JSTOR: 3069433.

Kogan, Leonid et al. (May 1, 2017). "Technological Innovation, Resource Allocation, and Growth". In: *The Quarterly Journal of Economics* 132.2, pp. 665–712. ISSN: 0033-5533. DOI: 10.1093/qje/qjw040.

La Porta, Rafael et al. (1997). "Good News for Value Stocks: Further Evidence on Market Efficiency". In: *The Journal of Finance* 52.2, pp. 859–874. ISSN: 1540-6261. DOI: 10.1111/j.1540-6261.1997.tb04825.x.

Lerner, Josh et al. (Mar. 15, 2023). *Financial Innovation in the 21st Century: Evidence from U.S. Patents*. DOI: 10.2139/ssrn.3871977. URL: https://papers.ssrn.com/abstract=3871977 (visited on 05/09/2023). Pre-published.

Livnat, Joshua and Richard R. Mendenhall (2006). "Comparing the Post–Earnings Announcement Drift for Surprises Calculated from Analyst and Time Series Forecasts". In: *Journal of Accounting Research* 44.1, pp. 177–205. ISSN: 1475-679X. DOI: 10.1111/j.1475-679X.2006.00196.x.

Manso, Gustavo, Benjamin Balsmeier, and Lee Fleming (Sept. 22, 2021). "Heterogeneous Innovation over the Business Cycle". In: *The Review of Economics and Statistics*, pp. 1–50. ISSN: 0034-6535. DOI: 10.1162/rest_a_01104.

Reimers, Nils and Iryna Gurevych (Aug. 27, 2019). *Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks*. DOI: 10.48550/arXiv.1908.10084. arXiv: 1908.10084 [cs]. URL: http://arxiv.org/abs/1908.10084 (visited on 09/02/2023). Pre-published.

Romer, Paul M. (Oct. 1990). "Endogenous Technological Change". In: *Journal of Political Economy* 98 (5, Part 2), S71–S102. ISSN: 0022-3808. DOI: 10.1086/261725.

Solow, Robert M. (1957). "Technical Change and the Aggregate Production Function". In: *The Review of Economics and Statistics* 39.3, pp. 312–320. ISSN: 00346535, 15309142. DOI: 10.2307/1926047. JSTOR: 1926047.

Zhang, X. Frank (2006). "Information Uncertainty and Stock Returns". In: *The Journal of Finance* 61.1, pp. 105–137. ISSN: 1540-6261. DOI: 10.1111/j.1540-6261.2006.00831.x.

Appendix A Variable Definitions

Variable	Definition and Calculation	Source
$PSS_{i,t}$	Average pairwise similarity score of the patent claims that are granted to firm i in between years $t-1$ and $t-5$	USPTO
$InnOrig_{i,t}$	Innovation originality is defined as the average patent citation diversity based on CPC subgroups across all patents granted to the firm i over the past five years from year t, following Hirshleifer, Hsu, and Li (2018)	USPTO
$ISP_{i,t}$	Following Manso, Balsmeier, and Fleming (2021), Internal Search Proximity (ISP) is a proxy for exploitative focus and it measures the degree of overlap between patents granted to the firm i in year t and the existing patent portfolio held by the same firm up to year $t-1$ based on CPC classification.	USPTO
$EPR_{i,t}$	Explorative Patent Ratio (EPR) is defined as the ratio of explorative patents to total number of patents granted to firm i in year t. Following the literature, a patent is determined explorative if at least 80% of its citations are new citations for the firm.	USPTO
FirmExp $_{i,t}$	Following Duysters et al. (2020), Firm Exploration is the inverse of a normalized Herfindahl index that captures the diversity of unique patent classes at year t based on patents granted to firm i for in the past five years. For S_r is the share of patent class r in firm i's patent classes, and N is the number of distinct patent classes, $FirmExp_{i,t} = \frac{N}{N-1}\left(1 - \sum_{r=1}^{N} S_r^2\right)$	USPTO
$InnovEff_{i,t}$	Innovative efficiency is the patents-based measure developed in Hirshleifer, Hsu, and Li (2013). It is patents granted scaled by R&D capital (the 5-year cumulative R&D expenses assuming an annual depreciation rate of 20%)	USPTO and Compustat
$PatentProd_{i,t}$	Natural logarithm of one plus the total number of patents granted per employee for firm i in year t: $\log\left(1 + \frac{\sum_{\tau \in t} Patent_{i,\tau}}{Number\ of\ Employees_{i,t}}\right)$	USPTO and KPSS.
$PatentValue_{i,t}$	Natural logarithm of one plus the total value of patents per patent granted to firm i in year t: $\log\left(1 + \frac{\sum_{\tau \in t} PatentValue_{i,\tau}}{Number\ of\ Patents_{i,t}}\right)$	USPTO and KPSS.

Continued on next page

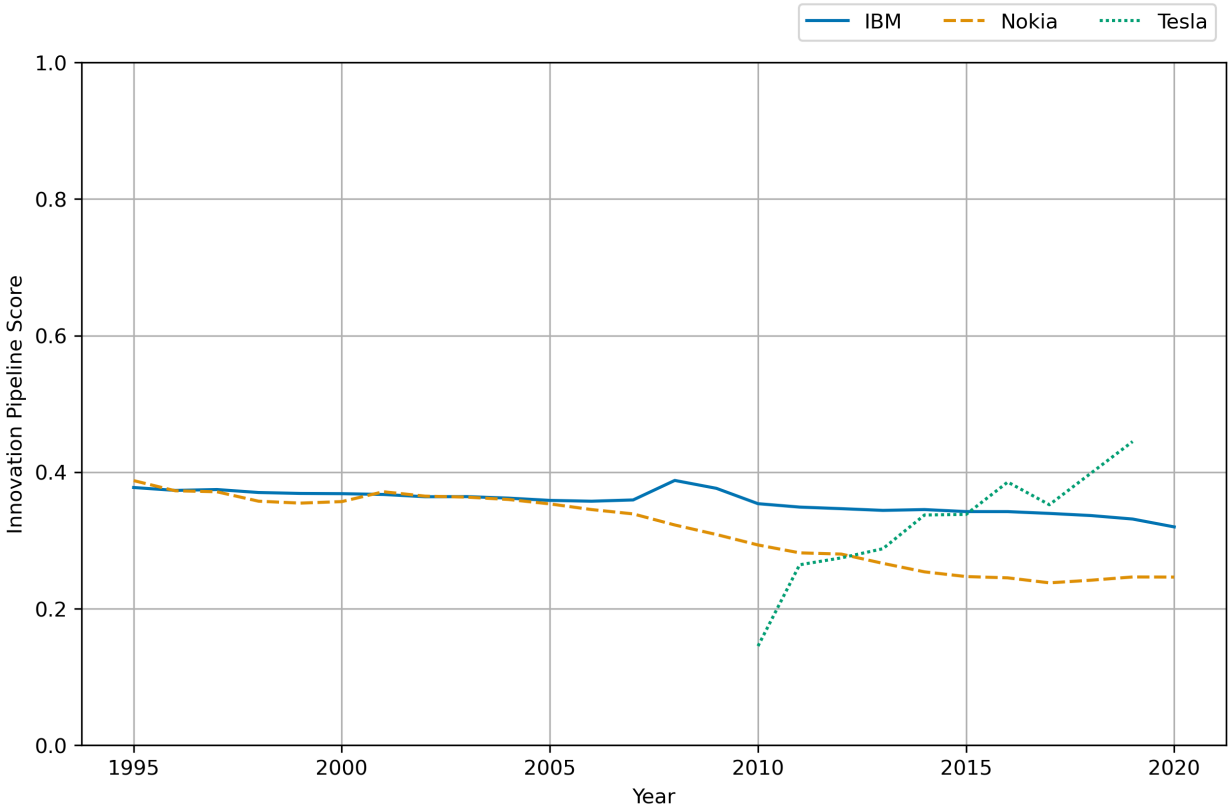
Variable	Definition and Calculation	Source
$\text{Patents}_{i,t}$	Total number of patents granted to firm i in year t divided by lagged assets	USPTO and Compustat
$\text{Citation}_{i,t}$	Natural logarithm of one plus the total number of cohort adjusted cites per patent granted to firm i in year t similar to Bernstein (2015): $\log\left(1+\frac{\sum_{\tau\in t} PatentCites_{i,\tau}}{Number\ of\ Patents_{i,t}}\right)$	USPTO and KPSS
$\text{SUE}_{i,t}$	Standardized unexpected earnings: $\frac{MedianForecast_{i,t}-Actual_{i,t}}{Price_{i,t}}$	Compustat, CRSP, I/B/E/S
$\text{RNOA}_{i,t}$	Return on Net Operating Assets for firm i in year t	Compustat
$\text{ROA}_{i,t}$	Return on Assets for firm i in year t	Compustat
$\text{ROE}_{i,t}$	Return on assets for firm i in year t is income before extraordinary items divided by lagged shareholder's equity	Compustat
$\text{OCF}_{i,t}$	Return on Assets for firm i in year t	Compustat
$\text{BM}_{i,t}$	Book-to-Market for firm i in year t is book equity of fiscal year t divided by market equity at the end of year t.	Compustat
$\text{Size}_{i,t}$	Natural log of market capitalization for firm i in year t is number of shares outstanding times share price times adjustment factor for the firm in that year.	CRSP
$\text{Age}_{i,t}$	Age of firm i in year t is approximated by the difference between year t and the time firm i firm appear in Compustat database.	Compustat
$\text{CapEx}_{i,t}$	Capital Expenditure divided by lagged assets	Compustat
$\text{R\&D}_{i,t}$	Research and Development Expense divided by lagged assets	Compustat
$\text{Advertising}_{i,t}$	Advertising Expense divided by lagged assets	Compustat
$\text{SGA}_{i,t}$	Selling, General, and Administrative Expenses for firm i in year t divided by lagged assets	Compustat

Continued on next page

Variable	Definition and Calculation	Source
$\text{Illiquidity}_{i,t}$	Illiquidity for firm i in year t is the Amihud measure	Compustat
$\text{12m-Return}_{i,t}$	The cumulative return over the preceding 12-months	CRSP
$\text{IO}_{i,t}$	Total shares held by institutions at time t divided by the shares outstanding at time t	Thomson Reuters
$\text{Female}_{i,t}$	The percentage of female innovators in a firm at time t	Revelio Labs
$\text{Tenure}_{i,t}$	The number of years an innovator work in a firm	Revelio Labs
$\text{Phd}_{i,t}$	The number of years an innovator work in a firm at time t	Revelio Labs
Education $\text{Field}_{i,t}$	Degree field of the innovator	Revelio Labs
$\text{Scientist}_{i,t}$	The percentage of scientist innovators	Revelio Labs
Scientist $\text{Prestige}_{i,t}$	The prestige score of the institutions that innovator attended ranging from -1 to 1	Revelio Labs
No of $\text{Innovators}_{i,t}$	The total number of innovators in a firm at time t	Revelio Labs
$\text{Skill Count}_{i,t}$	The number of skill categories determined using a clustering algorithm	Revelio Labs

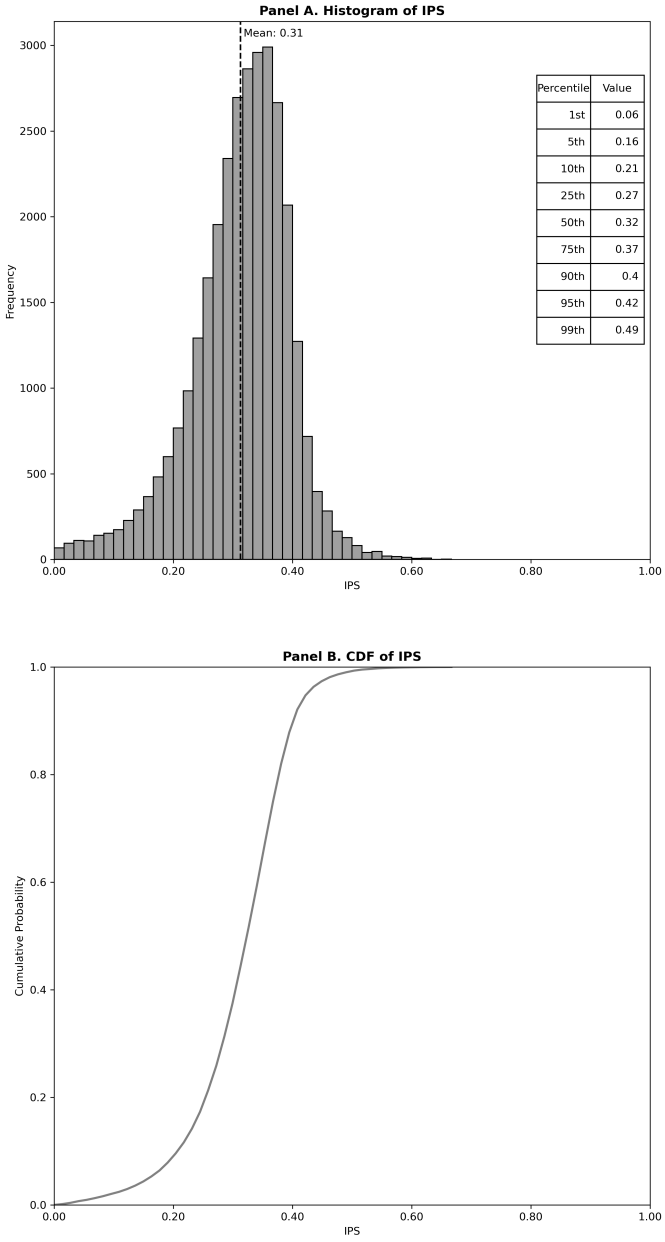
Appendix B Figures and Tables

Figure 1: The Innovation Pipeline Score (IPS) Trends of IBM, Nokia, and Tesla



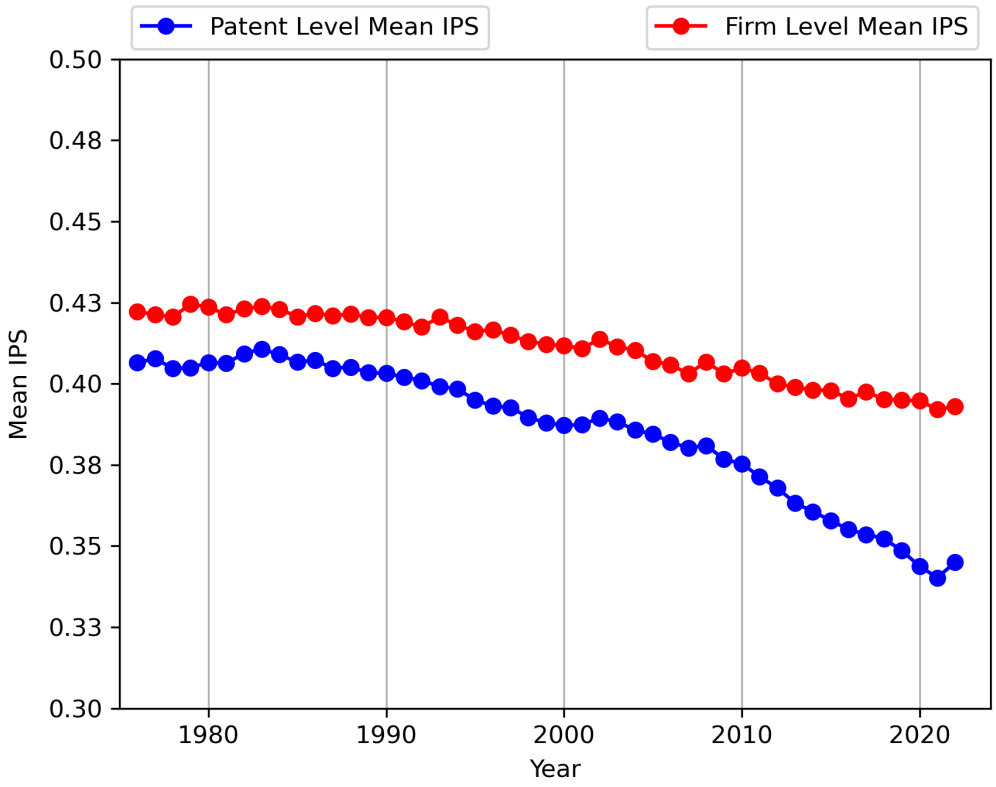
This figure shows the time-series plot of Innovation Pipeline Score (IPS) for IBM, Nokia, and Tesla. IBM maintains a consistent IPS, indicative of sustained innovation and a growing stock value. In contrast, Nokia shows an increase in IPS around 2005, correlating with a decline in market share and stock prices, suggesting a period of producing more similar claims and a diminished competitive edge in the smartphone market. Meanwhile, Tesla, entering the sample post-2010, displays a significant IPS reduction, illustrating its focus on pioneering developments in autonomous vehicles and next-generation batteries, indicative of a burgeoning and distinct innovation pipeline.

Figure 2: IPS Distribution



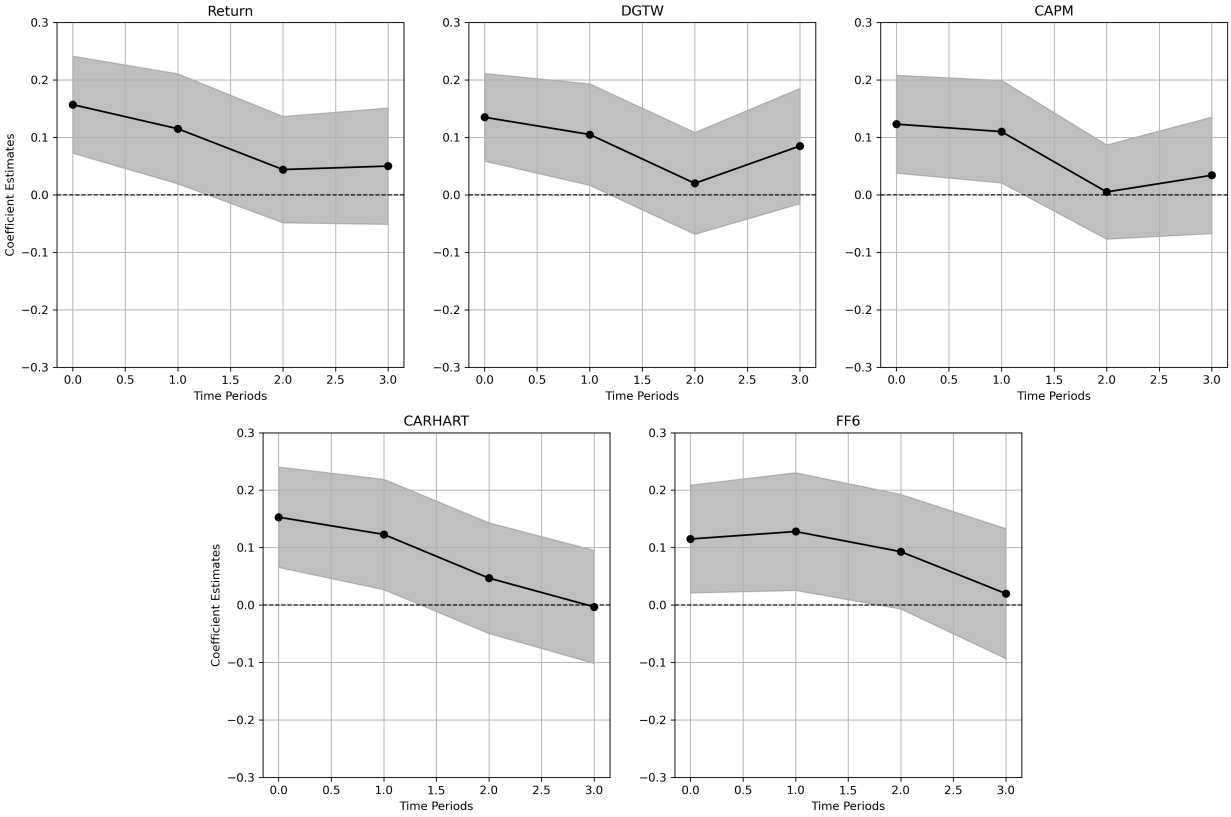
This figure illustrates the distribution of the IPS measure. Panel A displays a histogram plot of the measure, accompanied by percentile values at the 1st, 5th, 10th, 25th, 50th, 75th, 90th, 95th, and 99th levels, as well as the mean values. In Panel B, the cumulative distribution plot of the IPS is provided.

Figure 3: IPS scores over time



This figure presents the annual mean values of IPS scores at the patent and firm level. We group the IPS scores by year and calculate their mean values. There is a significant increase in the mean IPS values over time.

Figure 4: Coefficients on Similarity



This figure shows the coefficient estimates of the Patent Similarity Score (IPS) for quarterly return regressions. It examines raw returns (return), characteristic-adjusted returns (DGTW), risk-adjusted returns based on the CAPM, Carhart's four-factor, and Fama-French's six-factor models as dependent variables. Across all models, there is a consistent negative relationship between IPS and returns, which diminishes over subsequent quarters and becomes statistically insignificant after the second quarter.

Table 1: Summary Statistics and Correlations with IPS

	Mean	Median	Std. Dev	Pearson	Spearman
IPS	0.3128	0.3233	0.0824		
Productivity Measures:					
PatentProd	16.4244	3.7164	34.9082	-0.1139	-0.1565
InnovEff	37.3837	6.0000	103.4401	0.1386	0.2628
PatentValue	9.1260	2.8425	19.0709	0.0830	0.2132
Citation	0.0805	0.0000	0.2718	0.0262	0.1472
Profitability Measures:					
ROA	0.0582	0.1193	0.2254	0.1313	0.1287
ROE	-0.0243	0.0822	0.3844	0.1164	0.1340
OCF	550.2386	36.1980	1765.4060	0.1026	0.2559
RNOA	0.0631	0.1427	0.4612	0.1331	0.1478
Other Measures:					
Size (in $M)	6736.2370	658.9186	20864.5700	0.0880	0.2429
BM	0.5474	0.4451	0.4123	0.0379	0.0733
CapEx	0.0510	0.0387	0.0444	0.0466	0.0982
R&D	0.0964	0.0501	0.1287	-0.1159	-0.1354
Age	22.7203	16.5000	19.7249	0.2578	0.2885
Conglomerate	0.3071	0.0000	0.4613	0.1779	0.2158
Patent Stocks	1.1715	0.4749	2.1128	-0.0400	-0.0141
SGA	0.2855	0.2463	0.2205	-0.1301	-0.1385
Advertising	0.0115	0.0000	0.0299	0.0031	-0.0255
Momentum	0.1635	0.0840	0.5632	-0.0004	0.0410
IO	0.5477	0.5676	0.2921	0.0791	0.0633
IVol	0.0255	0.0196	0.0191	-0.1549	-0.1914
InnOrig	7.7191	6.1853	5.4481	-0.0952	-0.0740
ISP	0.7636	0.8479	0.2366	-0.1014	-0.1301
EPR	0.6196	0.6667	0.3179	0.2204	0.2005
FirmExp	0.2127	0.1702	0.1701	-0.3519	-0.4314

This table shows the mean, median, and standard deviations of various firm characteristics. The definitions of each variable is provided in Appendix A. The table also reports the Pearson and Spearmen correlations of each firm characteristics with the IPS measures.

Table 2: IPS and Future Innovation

Dep_{t+1}	PatentProd	InnovEff	PatentValue	Citation
IPS	0.061***	0.009***	0.011***	0.002***
	(8.40)	(4.27)	(3.82)	(3.10)
Dep_t	0.861***	0.171***	1.006***	0.011***
	(54.06)	(33.57)	(92.80)	(10.69)
Bm	-0.031***	-0.008**	-0.034***	-0.002
	(-3.17)	(-2.36)	(-4.55)	(-1.53)
CapEx	0.007	0.011***	0.001	0.003**
	(0.46)	(3.13)	(0.27)	(2.23)
R&D	0.075***	-0.015***	0.005	-0.002***
	(7.22)	(-7.04)	(0.80)	(-3.12)
Age	0.017***	-0.003*	0.025***	0.004***
	(3.40)	(-1.98)	(5.17)	(3.54)
Patents	-0.007	0.017***	-0.012***	0.000
	(-0.95)	(7.23)	(-6.46)	(0.61)
Advertising	-0.003	-0.002*	0.006*	0.008***
	(-0.39)	(-1.75)	(2.01)	(3.84)
12m-Return	-0.019	-0.002	0.108***	-0.001
	(-1.47)	(-1.58)	(6.90)	(-1.12)
Conglomerate	-0.003	0.008*	0.009**	0.009***
	(-0.26)	(2.00)	(2.05)	(4.52)
Observations	26047	23652	23983	26057
R^2	0.71	0.61	0.93	0.17
Industry × Time FEs	Yes	Yes	Yes	Yes

This table shows the panel regressions of log patent counts deflated by number of employees, patent value, and adjusted citations per patent in year $t+1$ onto Innovation Pipeline Score (IPS) and other control variables from 1983 to 2019. The control variables (defined in Appendix A) are book-to-market ratio (BM), capital expenditure (CapEx), research and development (R&D), firm age (Age), patents to assets ratio (Patents), advertising expenses (Advertising), 12 month returns (12m-return),the lagged values of the dependent variables. All regressions include industry and year fixed effects. We cluster standard errors by both industry and year to account for possible serial correlation and heteroskedasticity.

Table 3: IPS and Future Innovation Over the Years

PatentProd				
Year 1	Year 2	Year 3	Year 4	Year 5
0.061***	0.057***	0.062***	0.069***	0.068***
(8.40)	(6.93)	(7.69)	(8.28)	(6.63)

InnovEff				
Year 1	Year 2	Year 3	Year 4	Year 5
0.009***	0.007***	0.008***	0.012***	0.010***
(4.27)	(3.18)	(3.42)	(4.71)	(4.09)

PatentValue				
Year 1	Year 2	Year 3	Year 4	Year 5
0.011***	0.019***	0.028***	0.032***	0.035***
(3.82)	(3.49)	(4.04)	(4.23)	(3.55)

Citation				
Year 1	Year 2	Year 3	Year 4	Year 5
0.002***	0.004***	0.005***	0.005***	0.005***
(3.10)	(5.17)	(5.49)	(5.09)	(4.77)
Note:		*p<0.1; **p<0.05; ***p<0.01		

This table shows the panel regressions of log patent counts deflated by number of employees, patent value, and adjusted citations per patent each year from $t+1$ to $t+5$ onto Innovation Pipeline Score (IPS) and other control variables. The sample runs from 1983 to 2019. The control variables (defined in Appendix A) are book-to-market ratio (BM), capital expenditure (CapEx), research and development (R&D), firm age (Age), patents to assets ratio (Patents), advertising expenses (Advertising), 12 month returns (12m-return),the lagged values of the dependent variables. All regressions include industry and year fixed effects. We cluster standard errors by both industry and year to account for possible serial correlation and heteroskedasticity.

Table 4: IPS and Alternative Metrics to Predict Future Innovation

	PatentProd					InnovEff				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
IPS	0.061***	0.063***	0.077***	0.040***	0.050***	0.009***	0.009***	0.011***	0.006***	0.007***
	(8.53)	(8.83)	(11.02)	(5.72)	(7.26)	(4.35)	(4.21)	(4.90)	(2.78)	(2.99)
InnOrig	0.010				-0.032***	0.003				-0.001
	(1.38)				(-3.81)	(1.64)				(-0.57)
ISP		0.306***			0.320***		0.038***			0.046***
		(11.77)			(12.59)		(5.77)			(6.61)
EPR			-0.084***		-0.066***			-0.008***		-0.004**
			(-11.18)		(-8.37)			(-4.68)		(-2.31)
FirmExp				-0.143***	-0.172***				-0.023***	-0.026***
				(-11.07)	(-13.16)				(-6.02)	(-6.14)
Observations	26047	25436	25775	25864	25360	23652	23136	23413	23499	23063
R^2	0.7099	0.7169	0.7138	0.7123	0.7246	0.6107	0.6184	0.6120	0.6126	0.6218
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Industry × Time FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

	PatentValue					Citation				
	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)
IPS	0.011***	0.012***	0.013***	0.009***	0.009***	0.002***	0.003***	0.003***	0.001	0.002***
	(3.81)	(3.96)	(4.38)	(2.95)	(2.98)	(3.50)	(4.21)	(5.15)	(1.68)	(3.72)
InnOrig	-0.001				-0.007**	0.004***				0.003***
	(-0.74)				(-2.26)	(4.92)				(3.20)
ISP		0.063***			0.075***		0.029***			0.029***
		(5.88)			(6.67)		(6.99)			(6.55)
EPR			-0.007***		-0.003			-0.006***		-0.003***
			(-2.74)		(-1.07)			(-7.14)		(-3.20)
FirmExp				-0.020***	-0.030***				-0.006***	-0.008***
				(-3.87)	(-5.02)				(-3.73)	(-4.33)
Observations	23983	23471	23755	23819	23419	26057	25493	25784	25874	25417
R^2	0.9286	0.9301	0.9292	0.9291	0.9304	0.1679	0.1714	0.1691	0.1675	0.1746
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Industry × Time FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Note:										*p<0.1; **p<0.05; ***p<0.01

This table shows the innovation results in Table 2 by controlling with alternative measures: (1) the innovation originality measure (InnOrig) of Hirshleifer, Hsu, and Li (2018), (2) the internal search proximity (ISP) measure of Manso, Balsmeier, and Fleming (2021) (ExpFocus), (3) the ratio of explorative patents to total number of patents granted to firm i in year t. (4) the inverse of a normalized Herfindahl index that captures the diversity of unique patent classes following Duysters et al. (2020). The control variables (defined in Appendix A). We cluster standard errors by both industry and year to account for possible serial correlation and heteroskedasticity.

Table 5: Summary Statistics of Innovator Characteristics and Correlations with IPS

	Mean	Median	Std. Dev	Pearson	Spearman
IPS	0.3128	0.3233	0.0824		
Inventor Statistics (Revelio):					
Female (%)	0.1306	0.0928	0.1440	0.0149	0.0432
Tenure	6.0109	5.4091	3.4859	0.1943	0.2571
PhD (%)	0.2719	0.1794	0.2761	-0.0380	0.0178
Education Field	5.0097	4.0000	3.8157	0.2050	0.2181
Scientist (%)	0.2000	0.0690	0.2701	-0.0623	0.0126
Scientist Tenure	5.9315	5.0000	4.5630	0.1247	0.2039
Scientist Prestige	0.5239	0.5614	0.2565	-0.0136	-0.0594
No of Innovators	225.8759	29	1006.24	0.0790	0.2259
Skill Count	19.3071	21	5.4571	0.1580	0.1856

This table shows the mean, median, and standard deviations of various firm characteristics. The definitions of each variable is provided in Appendix A. The table also reports the Pearson and Spearmen correlations of each firm characteristics with the IPS measures.

Table 6: IPS and Innovator Characteristics

Raw

IPS Quintiles	Female (%)	Tenure	PhD (%)	Education Field	Scientist (%)	Scientist Tenure	Scientist Prestige	Team Size	Skill Count
1	13.38%	5.40	28.59%	3.20	21.17%	5.20	0.54	37.50	16.95
2	12.26%	5.47	24.48%	4.49	17.76%	5.34	0.52	137.12	19.35
3	12.37%	5.81	25.42%	5.44	17.74%	5.56	0.52	305.48	20.21
4	12.58%	6.45	26.37%	5.90	18.33%	6.32	0.52	374.11	20.48
5	14.97%	7.10	31.24%	5.82	25.16%	6.98	0.53	283.98	19.44
Difference	1.58%	1.70	2.65%	2.60	3.99%	1.78	-0.01	246.48	2.49
t-stat	(2.20)	(11.36)	(2.81)	(40.27)	(2.02)	(18.39)	(-0.99)	(13.28)	(16.62)

Industry Adjusted

IPS Quintiles	Female (%)	Tenure	PhD (%)	Education Field	Scientist (%)	Scientist Tenure	Scientist Prestige	Team Size	Skill Count
1	-0.41%	0.28	3.03%	-0.40	2.16%	-0.13	0.02	-63.63	0.89
2	-0.41%	0.28	0.83%	0.79	1.19%	-0.05	0.00	21.12	2.86
3	-0.06%	0.53	2.09%	1.73	1.64%	0.04	-0.01	190.38	3.67
4	0.03%	0.93	2.70%	2.21	1.37%	0.53	0.00	266.72	3.99
5	0.74%	1.32	4.71%	2.20	3.91%	0.90	0.02	201.42	3.37
Difference	1.16%	1.04	1.68%	2.60	1.75%	1.02	0.00	265.05	2.47
t-stat	(2.53)	(9.49)	(3.26)	(55.54)	(1.74)	(12.47)	(0.37)		

Size Adjusted

IPS Quintiles	Female (%)	Tenure	PhD (%)	Education Field	Scientist (%)	Scientist Tenure	Scientist Prestige	Team Size	Skill Count
1	13.03%	5.27	28.31%	3.41	18.86%	5.17	0.53	43.05	17.18
2	10.80%	5.57	24.87%	4.30	18.61%	5.26	0.53	130.17	18.90
3	11.34%	5.94	26.59%	4.70	18.55%	5.67	0.51	238.98	19.06
4	12.82%	6.46	27.57%	4.91	19.80%	6.29	0.52	256.41	19.23
5	16.55%	7.14	36.03%	4.75	28.04%	7.19	0.53	190.62	18.03
Difference	3.52%	1.87	7.72%	1.34	9.18%	2.02	0.00	147.57	0.85
t-stat	(2.00)	(6.14)	(5.28)	(9.30)	(4.49)	(9.69)	(0.26)	(10.12)	(4.38)

This table shows the work force characteristics in each IPS quintile. The definitions of each variable is provided in Appendix A.

Table 7: IPS and Short-Term Profitability

Dep_{t+1}	ROA	ROE	OCF	RNOA
IPS	0.342***	0.799***	0.438***	0.975***
	(3.80)	(4.17)	(4.90)	(3.97)
Dep_t	17.807***	21.925***	12.864***	64.630***
	(63.04)	(30.17)	(36.56)	(20.64)
BM	-0.439***	-3.653***	-0.676***	-2.822***
	(-4.55)	(-8.99)	(-4.73)	(-5.84)
CapEx	0.456***	-0.209	1.070***	-0.548
	(3.30)	(-0.81)	(7.94)	(-1.50)
R&D	-1.658***	-6.269***	-2.563***	-6.319***
	(-7.25)	(-13.03)	(-10.21)	(-8.70)
Age	0.418***	1.746***	0.427***	2.063***
	(8.13)	(10.45)	(6.90)	(7.06)
Patents	-0.319***	-0.867***	-0.526***	-1.342***
	(-5.28)	(-3.90)	(-6.56)	(-5.10)
Advertising	0.405***	0.845***	0.377***	1.150***
	(3.74)	(3.32)	(3.98)	(2.88)
12M Return	1.477***	4.013***	0.795***	3.534***
	(8.76)	(8.72)	(5.70)	(6.85)
Conglomerate	-0.138	-0.057	-0.287*	-0.633
	(-1.30)	(-0.16)	(-1.96)	(-1.68)
Observations	26,265	25,696	26,265	25,790
R^2	0.8022	0.5566	0.6623	0.6150
Industry × Time FEs	Yes	Yes	Yes	Yes
Note:				*p<0.1; **p<0.05; ***p<0.01

This table shows the panel regressions of profitability ratios in year $t+1$ onto Innovation Pipeline Score (IPS) and other control variables in year t from 1983 to 2019. The profitability ratios are return on assets (ROA), return on equity (ROE), and the operating cash flows (OCF). ROA is income before extraordinary items plus interest expenses scaled by lagged total assets. ROE is computed as income before extraordinary items plus interest expenses by lagged book equity. OCF is defined as income before extraordinary items plus depreciation less changes in working capital. The control variables (defined in Appendix A) are book-to-market ratio (BM), capital expenditure (CapEx), research and development (R&D), firm age (Age), patents to assets ratio (Patents), advertising expenses (Advertising), 12 month returns (12m-return),the lagged values of the dependent variables, the innovation originality measure of Hirshleifer, Hsu, and Li (2018), and the exploitation search focus measure of Manso, Balsmeier, and Fleming (2021). All regressions include industry and year fixed effects. We cluster standard errors by both industry and year to account for possible serial correlation and heteroskedasticity.

Table 8: IPS and Long-Term Profitability

ROA				
Year 1	Year 2	Year 3	Year 4	Year 5
0.342***	0.538***	0.512***	0.448***	0.432***
(3.80)	(4.41)	(3.75)	(3.39)	(2.97)

ROE				
Year 1	Year 2	Year 3	Year 4	Year 5
0.799***	1.392***	1.293***	1.140***	0.976***
(4.17)	(5.72)	(4.64)	(3.68)	(3.38)

OCF				
Year 1	Year 2	Year 3	Year 4	Year 5
0.438***	0.465***	0.550***	0.424***	0.422***
(4.90)	(4.25)	(4.13)	(3.30)	(2.83)

RNOA				
Year 1	Year 2	Year 3	Year 4	Year 5
0.975***	1.410***	1.267***	1.394***	1.199***
(3.97)	(4.24)	(3.12)	(3.82)	(3.22)
Note:		*p<0.1; **p<0.05; ***p<0.01		

This table shows the panel regressions of profitability ratios in years from t+1 to t+5 onto Innovation Pipeline Score (IPS) and other control variables in year t from 1983 to 2019. The profitability ratios are return on assets (ROA), return on equity (ROE), and the operating cash flows (OCF). ROA is income before extraordinary items plus interest expenses scaled by lagged total assets. ROE is computed as income before extraordinary items plus interest expenses by lagged book equity. OCF is defined as income before extraordinary items plus depreciation less changes in working capital. The control variables omitted from the table for brevity are book-to-market ratio (BM), capital expenditure (CapEx), research and development (R&D), firm age (Age), patents to assets ration (Patents), advertising expenses (Advertising), 12 month returns (12m-return),the lagged values of the dependent variables, the innovation originality measure of Hirshleifer, Hsu, and Li (2018), and the exploitation search focus measure of Manso, Balsmeier, and Fleming (2021) (ExpFocus).The control variables are defined in Appendix A. All regressions include industry and year fixed effects. We cluster standard errors by both industry and year to account for possible serial correlation and heteroskedasticity.

Table 9: IPS and Alternative Metrics to Predict Future Profitability

	ROA					ROE				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
IPS	0.003***	0.004***	0.004***	0.003***	0.003***	0.008***	0.009***	0.009***	0.007***	0.007***
	(3.68)	(4.34)	(4.04)	(3.23)	(3.36)	(3.99)	(4.73)	(4.68)	(3.75)	(3.83)
InnOrig	-0.002**				-0.003***	-0.006*				-0.008***
	(-2.08)				(-3.22)	(-1.97)				(-2.78)
ISP		0.003***			0.003***		0.004			0.006**
		(2.96)			(3.84)		(1.62)			(2.31)
EPR			-0.000		0.000			0.000		0.001
			(-0.46)		(0.24)			(0.12)		(0.25)
FirmExp				-0.004***	-0.006***				-0.010***	-0.014***
				(-2.91)	(-4.08)				(-3.36)	(-3.76)
Observations	26265	25643	25993	26082	25567	25696	25107	25431	25519	25032
R^2	0.8023	0.8012	0.8019	0.8023	0.8011	0.5567	0.5567	0.5588	0.5576	0.5582
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Industry × Time FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

	OCF					RNOA				
	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)
IPS	0.004***	0.005***	0.004***	0.004***	0.004***	0.009***	0.010***	0.011***	0.008***	0.007***
	(4.75)	(5.58)	(5.05)	(4.58)	(4.87)	(3.81)	(4.48)	(4.18)	(3.20)	(3.25)
InnOrig	-0.004***				-0.006***	-0.007**				-0.010***
	(-3.66)				(-4.50)	(-2.61)				(-3.47)
ISP		0.004***			0.005***		0.009**			0.012***
		(4.16)			(4.93)		(2.49)			(3.56)
EPR			0.000		0.000			-0.001		0.001
			(0.19)		(0.33)			(-0.28)		(0.32)
FirmExp				-0.003**	-0.006***				-0.015***	-0.022***
				(-2.08)	(-3.35)				(-3.97)	(-5.58)
Observations	26265	25643	25993	26082	25567	25790	25173	25518	25607	25097
R^2	0.6627	0.6600	0.6611	0.6628	0.6590	0.6152	0.6140	0.6143	0.6147	0.6146
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Industry × Time FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note: *p<0.1; **p<0.05; ***p<0.01

This table shows the innovation results in Table 7 by controlling with alternative measures. The control variables (defined in Appendix A). We cluster standard errors by both industry and year to account for possible serial correlation and heteroskedasticity.

Table 10: IPS and Returns

	Raw Return	DGTW	CAPM	CARHART	FF6
IPS	0.091***	0.084***	0.065**	0.090***	0.094***
	(3.25)	(3.27)	(2.34)	(3.08)	(2.92)
Size	-0.091	-0.065	-0.016	-0.013	0.030
	(-0.96)	(-0.87)	(-0.19)	(-0.17)	(0.38)
BM	0.227***	0.048	0.189***	0.116**	0.114**
	(4.08)	(1.03)	(3.53)	(2.18)	(2.09)
Momentum	-0.031	-0.055	-0.056	-0.020	-0.140
	(-0.13)	(-0.35)	(-0.27)	(-0.10)	(-0.70)
Ret	-0.448***	-0.477***	-0.510***	-0.608***	-0.840***
	(-3.52)	(-5.57)	(-4.20)	(-4.62)	(-5.55)
CapExp	-0.020	-0.008	-0.056	-0.028	-0.007
	(-0.51)	(-0.21)	(-1.43)	(-0.67)	(-0.15)
R&D	0.152	0.176**	0.113	0.077	0.144*
	(1.62)	(2.21)	(1.28)	(0.94)	(1.81)
Io	-0.082	-0.099**	-0.110*	-0.160***	-0.182***
	(-1.36)	(-2.18)	(-1.87)	(-2.88)	(-3.12)
Age	-0.001	-0.002	-0.000	-0.001	-0.004**
	(-0.28)	(-0.88)	(-0.21)	(-0.64)	(-2.01)
Patent	-0.068	-0.043	-0.017	-0.024	0.011
	(-1.61)	(-1.14)	(-0.40)	(-0.56)	(0.25)
Advertising	0.094***	0.067**	0.094***	0.057*	0.021
	(2.92)	(2.54)	(2.91)	(1.70)	(0.56)
Roe	0.213***	0.211***	0.285***	0.248***	0.159***
	(3.86)	(4.29)	(5.45)	(4.57)	(2.90)
Illiquidity	0.115**	0.060	0.200***	0.130***	0.118**
	(2.27)	(1.59)	(4.46)	(3.04)	(2.50)
Ivol	0.006	0.070	-0.068	0.049	0.107
	(0.04)	(0.69)	(-0.59)	(0.56)	(1.17)
Observations	318376	316780	316791	316791	316791
R^2	0.1876	0.0127	0.0421	0.0152	0.0145
Industry × Time FEs	Yes	Yes	Yes	Yes	Yes
Note:					*p<0.1; **p<0.05; ***p<0.01

This table shows the yearly return regressions in each year $t+1$ on the Innovation Pipeline Score (IPS) and other control variables in year t from 1983 to 2019. Column (1) uses raw returns, column (2) uses characteristic-adjusted returns (DGTW), column (3) uses risk-adjusted returns based on the CAPM, columns (4) uses Carhart's four-factor, and columns (5) uses Fama-French's six-factor models as dependent variables. The control variables (defined in Appendix A are return on assets (ROA), book-to-market ratio (BM), capital expenditure (CapEx), research and development (R&D), firm age (Age), patents to assets ration (Patents), advertising expenses (Advertising), institutional ownership (IO), 12 month returns (12m-return),the lagged values of the dependent variables, innovation efficiency (Innov_Eff), the innovation originality measure of Hirshleifer, Hsu, and Li (2018), and the exploitation search focus measure of Fitzgerald et al. (2021) (ExpFocus). All regressions include industry and year fixed effects. We cluster standard errors by both industry and year to account for possible serial correlation and heteroskedasticity.

Table 11: IPS and Cross-Section of Returns

Panel A: Based on raw returns

	Size		Analyst		IO	
	Small	Big	Low	High	Low	High
IPS	0.096***	0.077*	0.103***	0.067	0.125***	0.054
	(2.63)	(1.69)	(2.97)	(1.40)	(3.18)	(1.36)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Industry × Time FEs	Yes	Yes	Yes	Yes	Yes	Yes
Observations	144548	131307	163854	145715	144332	164473
R^2	0.2429	0.3733	0.2397	0.3281	0.2485	0.3120

Panel B: Based on Carhart

	Size		Analyst		IO	
	Small	Big	Low	High	Low	High
IPS	0.097**	0.040	0.108***	0.039	0.091**	0.057
	(2.40)	(0.84)	(2.89)	(0.78)	(2.11)	(1.38)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Industry × Time FEs	Yes	Yes	Yes	Yes	Yes	Yes
Observations	143625	130825	162811	145145	163886	143625
R^2	0.0928	0.1495	0.0880	0.1206	0.1081	0.0882

Note: *p<0.1; **p<0.05; ***p<0.01

This table shows the effect of IPS on the cross-section of returns. We take the top and bottom 30th percentiles based on size, number of analyst following, and Ivol. Panel A shows the results for the raw returns and Panel B for the Carhart returns. All regressions include the same control variables in Table 10, industry and year fixed effects. We cluster standard errors by both industry and year to account for possible serial correlation and heteroskedasticity.

Table 12: IPS and Market Reaction around Earnings Announcements

Dep_{t+1}:	SUE	CAR1	CAR2
IPS	0.012**	0.094***	0.095***
	(2.36)	(3.06)	(3.05)
Dep_t	0.198***	0.052	0.050
	(11.26)	(1.34)	(1.26)
ROA	0.025**	0.450***	0.471***
	(2.43)	(8.66)	(8.97)
BM	-0.016	0.107***	0.115***
	(-1.25)	(2.69)	(2.92)
12m-Return	0.032***	-0.092**	-0.082*
	(3.70)	(-2.08)	(-1.82)
Age	-0.002	-0.040	-0.029
	(-0.47)	(-1.20)	(-0.89)
CapEx	-0.014**	-0.107***	-0.115***
	(-2.30)	(-3.09)	(-3.20)
R&D	0.030***	0.192***	0.184***
	(2.92)	(3.69)	(3.44)
Patents	-0.032***	-0.103***	-0.099***
	(-3.99)	(-3.06)	(-2.97)
Advertising	0.006	0.016	0.014
	(1.15)	(0.44)	(0.37)
Conglomerate	0.009	0.028	0.044
	(0.82)	(0.43)	(0.66)
Observations	63944	92671	92666
R^2	0.1045	0.0458	0.0466
Industry × Time FEs	Yes	Yes	Yes
Note:	*p<0.1; **p<0.05; ***p<0.01		

This table shows the panel regressions of earnings surprises in each quarter of year $t+1$ on the Innovation Pipeline Score (IPS) and other control variables in year t from 1983 to 2019. Columns (1) and (2) use earnings surprises based on the seasonal random walk model of Livnat and Mendenhall (2006), while columns (3) and (4) use analyst forecast. The control variables (defined in Appendix A are return on assets (ROA), book-to-market ratio (BM), capital expenditure (CapEx), research and development (R&D), firm age (Age), patents to assets ration (Patents), advertising expenses (Advertising), institutional ownership (IO), 12 month returns (12m-return),the lagged values of the dependent variables, innovation efficiency (Innov_Eff), the innovation originality measure of Hirshleifer, Hsu, and Li (2018), and the exploitation search focus measure of (ExpFocus). All regressions include industry and year fixed effects. We cluster standard errors by both industry and year to account for possible serial correlation and heteroskedasticity.

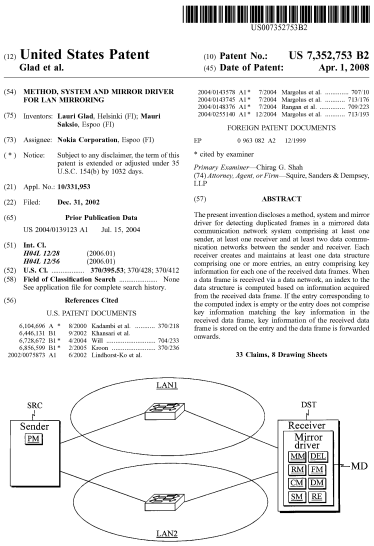
Table 13: IPS and Volatility

Dep_{t+1}:	Raw Return Volatility	CAPM		CARHART	
		Market β	Ivol	Market β	Ivol
IPS	-0.053	0.054	-0.084	-0.505	-0.081
	(-0.54)	(0.01)	(-0.89)	(-0.10)	(-0.86)
Controls	Yes	Yes	Yes	Yes	Yes
Industry × Time FEs	Yes	Yes	Yes	Yes	Yes
Observations	22234	22231	22231	22231	22231
R^2	0.7559	0.5731	0.7508	0.2825	0.7474
Note:					*p<0.1; **p<0.05; ***p<0.01

This table shows the panel regressions of volatility measures in year $t+1$ onto Innovation Pipeline Score (IPS) and other control variables in year t from 1983 to 2019. In Column (1), we use the volatility of daily returns in year $t+1$. Column (2) and (3) are the market beta and idiosyncratic volatility based on CAPM, computed using daily returns in year $t+1$. Column (4) and (5) use four-factor model (market, hml, smb, and momentum) to compute market beta and idiosyncratic volatility. All regressions include the same control variables in 10, industry and year fixed effects. We cluster standard errors by both industry and year to account for possible serial correlation and heteroskedasticity.

Online Appendix

O.A. Figure 1: Nokia Patent Examples



US007352753B2

(12) **United States Patent**
Glad et al.

(10) **Patent No.: US 7,352,753 B2**
(45) **Date of Patent: Apr. 1, 2008**

(54) **METHOD, SYSTEM AND MIRROR DRIVER FOR LAN MIRRORING**

(75) Inventors: **Lauri Glad**, Helsinki (FI); **Mauri Sakslo**, Espoo (FI)

(73) Assignee: **Nokia Corporation**, Espoo (FI)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 1052 days.

(21) Appl. No.: **10/331,953**

(22) Filed: **Dec. 31, 2002**

(65) **Prior Publication Data**

US 2004/0139123 A1 Jul. 15, 2004

(51) **Int. Cl.**
H04L 12/28 (2006.01)
H04L 12/56 (2006.01)

(52) **U.S. Cl.** **370/395.53**; 370/428; 370/412

(58) **Field of Classification Search** ..... None
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

6,104,696 A * 8/2000 Kadambi et al. 370/218
6,446,131 B1 * 9/2002 Khanzai et al.
6,728,672 B1 * 4/2004 Will 704/233
6,836,559 B1 * 2/2005 Kwon 370/236
2002/0075873 A1 6/2002 Lindhorst-Ko et al.
2004/0143578 A1 * 7/2004 Margolus et al. 707/10
2004/0147345 A1 * 7/2004 Margolus et al. 713/176
2004/0148376 A1 * 7/2004 Rangan et al. 709/223
2004/0255140 A1 * 12/2004 Margolus et al. 713/193

FOREIGN PATENT DOCUMENTS

EP 0 963 082 A2 12/1999

* cited by examiner

Primary Examiner—Chirag G. Shah
(74) *Attorney, Agent, or Firm*—Squire, Sanders & Dempsey, L.L.P.

(57) **ABSTRACT**

The present invention discloses a method, system and mirror driver for detecting duplicated frames in a mirrored data communication network system comprising at least one sender, at least one receiver and at least two data communication networks between the sender and receiver. Each receiver creates and maintains at least one data structure comprising one or more entries, an entry comprising key information for each one of the received data frames. When a data frame is received via a data network, an index to the data structure is computed based on information acquired from the received data frame. If the entry corresponding to the computed index is empty or the entry does not comprise key information matching the key information in the received data frame, key information of the received data frame is stored on the entry and the data frame is forwarded onwards.

33 Claims, 8 Drawing Sheets

(a) First page of the patent 7,352,753

creating and maintaining at least one data structure comprising one or more entries, an entry comprising key information for each one of the received data frames;

receiving a data frame via a data communication network;

computing an index to the data structure based on information acquired from the received data frame, the index pointing to an entry in the data structure; and when the entry corresponding to the computed index is empty or the entry does not comprise key information matching the key information in the received data frame,

storing the key information of the received data frame on the entry; and

forwarding the data frame to an upper software layer;

when the entry corresponding to the computed index comprises the key information matching the key information in the received data frame, the method further includes,

discarding the received data frame; and

deleting the corresponding key information from the entry,

wherein the mirrored data communications network system includes at least one sender, at least one receiver and at least two redundant physically separate data communication networks between the sender and receiver.

2. The method according to claim **1**, wherein in the computing step: computing the index to the data structure based on at least two pieces of the following information:

a sender address;

a receiver address;

a data frame identifier; or

a checksum.

3. The method according to claim **1**, wherein key information within an entry comprises at least two of the following:

a sender address;

a receiver address;

a data frame identifier;

a checksum; or

a time stamp.

4. The method according to claim **2**, wherein the data frame identifier is incorporated in one of a mirror hearer, a virtual local area network identifier (VID) field in the virtual local area network (VLAN) tag or 802.2 standard network access protocol (SNAP) org field.

5. The method according to claim **2**, wherein the checksum is a cyclic redundancy checksum in the frame check sequence field.

6. The method according to claim **1**, wherein the data structure is a hash table.

7. The method according to claim **1**, wherein the index to the data structure is computed by a hash function.

8. The method according to claim **1**, wherein the method further comprises:

deleting timed-out pieces of key information from the data structure.

9. The method according to claim **8**, wherein the method further comprises:

deleting every key information comprising an older time stamp than a predetermined threshold value.

10. The method according to claim **1**, wherein the communication networks are local area networks.

11. A system for detecting duplicated frames in a mirrored data communication network system, the system comprising:

a creating and maintaining unit configured to create and maintain at least one data structure comprising one or more entries, an entry comprising key information for each one of the received data frames;

a receiving unit configured to receive a data frame;

a computing unit configured to compute an index to the data structure based on information acquired from the received data frame, the index pointing to an entry in the data structure;

a storing unit configured to store the key information of the received data frame on the entry, when the entry corresponding to the computed index is empty or the entry does not comprise key information matching the key information in the received data frame;

a forwarding unit configured to forward the data frame to an upper software layer;

a discarding unit configured to discard the received data frame, when the entry comprises the key information matching the key information in the received data frame; and

a deleting unit configured to delete key information from an entry, when the entry comprises the key information matching the key information in the received data frame,

wherein the mirrored data communication network system includes at least one sender, at least one receiver and at least two redundant physically separate data communication networks between the sender and receiver.

12. The system according to claim **11**, wherein the computing unit is configured to compute the index to the data structure based on at least two pieces of the following information:

a sender address;

a receiver address;

a data frame identifier; or

a checksum.

13. The system according to claim **11**, wherein key information within an entry comprises at least two of the following:

a sender address;

a receiver address;

a data frame identifier;

a checksum; or

a time stamp.

14. The system according to claim **11**, wherein the system further comprises a placing unit configured to place a data frame identifier in one of a mirror hearer, a virtual local area network identifier (VID) field in the virtual local area network (VLAN) tag or 802.2 standard network access protocol (SNAP) org field.

15. The system according to claim **11**, wherein the system further comprises a reading unit configured to read a data frame identifier from one of a mirror hearer, a virtual local area network identifier (VID) field in the virtual local area network (VLAN) tag or 802.2 standard network access protocol (SNAP) org field.

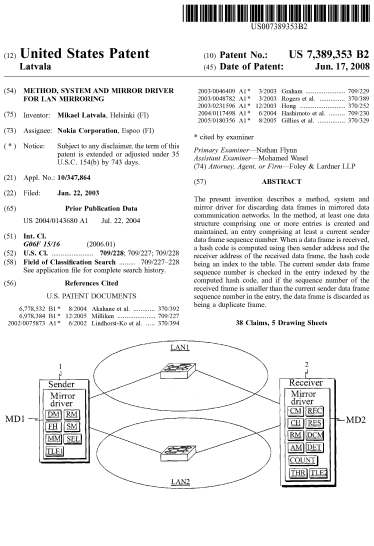
16. The system according to claim **12**, wherein the checksum is a cyclic redundancy checksum in the frame check sequence field.

17. The system according to claim **11**, wherein the data structure is a hash table.

18. The system according to claim **11**, wherein the index to the data structure is computed by a hash function.

19. The system according to claim **11**, wherein the deleting unit is configured to delete timed-out pieces of key information from the data structure.

(b) Claims page of the patent 7,352,753



US007389353B2

(12) **United States Patent**
Latvala

(10) **Patent No.: US 7,389,353 B2**
(45) **Date of Patent: Jun. 17, 2008**

(54) **METHOD, SYSTEM AND MIRROR DRIVER FOR LAN MIRRORING**

(75) Inventor: **Mikael Latvala**, Helsinki (FI)

(73) Assignee: **Nokia Corporation**, Espoo (FI)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 743 days.

(21) Appl. No.: **10/347,864**

(22) Filed: **Jan. 22, 2003**

(65) **Prior Publication Data**

US 2004/0143680 A1 Jul. 22, 2004

(51) **Int. Cl.**
G06F 15/16 (2006.01)

(52) **U.S. Cl.** **709/228**; 709/227; 709/228

(58) **Field of Classification Search** 709/227-228
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

6,778,532 B1 * 8/2004 Akahane et al. 370/392
6,978,384 B1 * 12/2005 Milliken 709/227
2002/0075873 A1 * 6/2002 Lindhorst-Ko et al. 370/394
2003/0040409 A1 * 3/2003 Graham 709/229
2003/0048782 A1 * 3/2003 Rogers et al. 370/389
2003/0231596 A1 * 12/2003 Hong 370/252
2004/0117498 A1 * 6/2004 Hashimoto et al. 709/230
2005/0180356 A1 * 8/2005 Gillies et al. 370/329

* cited by examiner

Primary Examiner—Nathan Flynn
Assistant Examiner—Mohamed Wasel
(74) *Attorney, Agent, or Firm*—Foley & Lardner LLP

(57) **ABSTRACT**

The present invention describes a method, system and mirror driver for discarding data frames in mirrored data communication networks. In the method, at least one data structure comprising one or more entries is created and maintained, an entry comprising at least a current sender data frame sequence number. When a data frame is received, a hash code is computed using then sender address and the receiver address of the received data frame, the hash code being an index to the table. The current sender data frame sequence number is checked in the entry indexed by the computed hash code, and if the sequence number of the received frame is smaller than the current sender data frame sequence number in the entry, the data frame is discarded as being a duplicate frame.

38 Claims, 5 Drawing Sheets

(c) First page of patent 7,389,353

communication networks LAN1, LAN2 between the sender **1** and the receiver **2**. A data frame sent by the sender **1** is sent to the receiver **2** via both LANs LAN1 and LAN2. The sender and receiver are e.g. computers, routers, switches etc.

According to the present invention, the sender comprises at least one data structure TLE1 comprising one or more entries, an entry comprising at least receiver identification information and a current receiver data frame sequence number of the receiver, means for finding FM out the current receiver data frame sequence number based on the receiver address information, means for marking MM the data frame to be sent to the receiver via the data communication networks with the current receiver data frame sequence number, means for duplicating DM the data frame, means for sending SM the data frame to the receiver via at least two data communication networks, and means for incrementing RM the current receiver data frame sequence number.

In this embodiment, the sender further comprises means for selecting SEL a sequence number to be used as the current receiver data frame sequence number if the sender is starting or restarted.

In one embodiment of FIG. **5**, the current receiver data frame sequence numbers are incorporated in one of a mirror hearer, a VID field in the VLAN tag or 802.2 SNAP org field.

According to the present invention, the receiver comprises at least one data structure TLE2 comprising one or more entries, an entry comprising at least sender and receiver identification information and a current sender data frame sequence number of the sender, means for receiving REC data frames, means for computing CM a hash code using the sender address and the receiver address of the received data frame, the hash code being an index to the data structure, means for checking CH the current sender data frame sequence number in the entry indexed by the computed hash code, and means for discarding DCM the data frame, wherein the received frame is discarded if the sequence number of the received frame is smaller than the current sender data frame sequence number in the entry.

In this embodiment, means for discarding DCM are arranged to discard the data frame if the sequence number of the received data frame is equal or greater than the current sender data frame sequence number in the entry incremented by a predetermined window size. The receiver further comprises a discard counter COUNT for each sender, wherein the discard counter COUNT is arranged to count discarded data frames of a sender, a predetermined discard threshold value THR, means for determining DET that the sender is not synchronized when the discard counter exceeds the predetermined discard threshold value, means for resetting RES the discard counter of the sender, means for accepting AM the next packet from the sender automatically, and means for incrementing RM the current sender data frame sequence number.

In one embodiment, means for accepting AM the data frame are arranged to accept the data frame if the sequence number of the received data frame is equal or greater than the current sender data frame sequence number and equal or smaller than the current sender data frame sequence number in the entry incremented by a predetermined window size.

The aforementioned means are implemented in a mirror driver (mirror driver MD1 in the sending side and mirror driver MD2 in the receiving side).

The aforementioned means, data structure and variables are implemented in a known manner preferably with software and/or hardware, and therefore they are not described in more detail.

One of the advantages of the present invention is that an unsynchronized receiver does not require manual maintenance interventions. A further advantage is that the sender does not have to be informed when the receiver has gone out of synchronization. Furthermore, it does not have to change its behavior. The receiver gets in synchronization immediately after the failure of the receiver node has happened. The time required by the receiver to get in synchronization after a failure of sender node is dependent only on the discardThreshold value: the lower the value, the sooner the receiver gets back in synchronization.

It is obvious to a person skilled in the art that with the advancement of technology, the basic idea of the invention may be implemented in various ways. The invention and its embodiments are thus not limited to the examples described above, instead they may vary within the scope of the claims.

The invention claimed is:

1. A method for sending duplicated frames in a mirrored data communication network system comprising at least one sender node and at least one receiver node and at least two data communication networks between a sender node and a receiver node, wherein the method comprises the steps of:

creating and maintaining at least one data structure comprising one or more entries, an entry comprising at least a current receiver data frame sequence number, wherein a single current receiver data frame sequence number is maintained for each receiver node, and a current sender data frame sequence number is maintained for each sender node;

finding out the current receiver data frame sequence number based on the receiver address information;

marking the data frame to be sent to the receiver node via the data communication networks with the current receiver data frame sequence number;

duplicating the data frame;

sending the data frame to the receiver node via at least two data communication networks; and

at least one of updating the receiver data frame sequence number within the entry and putting the receiver node into a starting phase to resynchronize the receiver node with the sender node if the receiver node has become unsynchronized with the sender node.

2. The method according to claim **1**, wherein if the sender is starting or restarted, the method further comprises the step of:

selecting a sequence number to be used as the current receiver data frame sequence number.

3. The method according to claim **1**, wherein the current receiver data frame sequence numbers are incorporated in one of a mirror hearer, a VID field in the VLAN tag or 802.2 SNAP org field.

4. A method for receiving duplicated frames in a mirrored data communication network system comprising at least one sender node and at least one receiver node and at least two data communication networks between a sender node and a receiver node, wherein the method comprises the steps of:

creating and maintaining at least one data structure comprising one or more entries, an entry comprising at least a current sender data frame sequence number;

receiving a data frame;

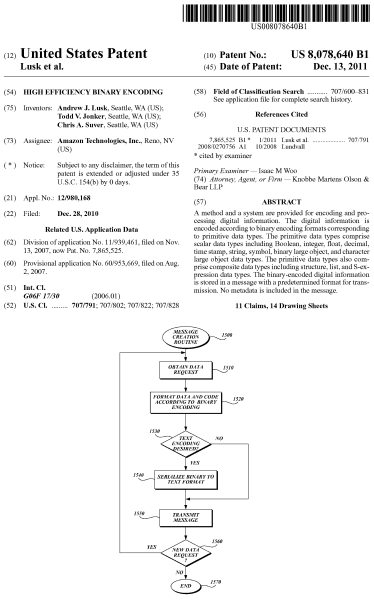
utilizing at least a sender address and a receiver address of the received data frame to compute a hash code, the hash code being an index to the at least one data structure;

checking the current sender data frame sequence number in the entry indexed by the computed hash code; and

(d) Claims page of patent 7,389,353

This figure shows the first and last pages of Nokia's patents 7,352,753 and 7,389,353. Even though these patents are in USPC classes (one is in 370 and the other is 709), their similarity is 91%.

O.A. Figure 2: Amazon Patent Examples



US008078640B1

(12) **United States Patent**
Lusk et al.

(10) **Patent No.: US 8,078,640 B1**
(45) **Date of Patent: Dec. 13, 2011**

(54) **HIGH EFFICIENCY BINARY ENCODING**

(75) Inventors: **Andrew J. Lusk**, Seattle, WA (US); **Todd V. Jonker**, Seattle, WA (US); **Chris A. Suwer**, Seattle, WA (US)

(73) Assignee: **Amazon Technologies, Inc.**, Reno, NV (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

(21) Appl. No.: **12/980,168**

(22) Filed: **Dec. 28, 2010**

Related U.S. Application Data

(62) Division of application No. 11/939,461, filed on Nov. 13, 2007, now Pat. No. 7,865,525.

(60) Provisional application No. 60/953,669, filed on Aug. 2, 2007.

(51) **Int. Cl.**
G06F 17/30 (2006.01)

(52) **U.S. Cl.** **707/791**; 707/802; 707/822; 707/828

(58) **Field of Classification Search** 707/600–831
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

7,865,525 B1 * 1/2011 Lusk et al. 707/791
2008/0270756 A1 10/2008 Landvall

* cited by examiner

Primary Examiner — Isaac M Woo
(74) *Attorney, Agent, or Firm* — Knobbe Martens Olson & Bear LLP

(57) **ABSTRACT**

A method and a system are provided for encoding and processing digital information. The digital information is encoded according to binary encoding formats corresponding to primitive data types. The primitive data types comprise scalar data types including Boolean, integer, float, decimal, time stamp, string, symbol, binary large object, and character large object data types. The primitive data types also comprise composite data types including structure, list, and S-expression data types. The binary-encoded digital information is stored in a message with a predetermined format for transmission. No metadata is included in the message.

11 Claims, 14 Drawing Sheets

(a) First page of the patent 8,078,640

application on the other server. Those skilled in the art will appreciate that scripts are used to perform many data processing functions in business applications, such as conversion of formats, database queries, display formatting, etc. The method terminates at block **1660**.

While illustrative embodiments have been illustrated and described, it will be appreciated that various changes can be made therein without departing from the spirit and scope of the invention.

The embodiments of the invention in which an exclusive property or privilege is claimed are defined as follows:

1. A computer-readable medium having computer-executable components encoded thereon for communicating digital information, wherein the computer-executable components, in response to execution, cause a computing device to:

receive binary-encoded digital information in a multi-record message having:

a length of message field for representing a length of the multi-record message;

a binary format identification field for identifying binary data encoding formats; and

a number of value fields for representing component information parts, the number of value fields being based on the length of the multi-record message, wherein each of the value fields has a value type descriptor field, a length of data field, and a representation field; and

decode the component information parts of the binary-encoded digital information according to the binary data encoding formats to retrieve the digital information, each binary data encoding format corresponding to one of a plurality of primitive data types;

wherein the binary-encoded digital information is decoded to retrieve the digital information based on the value fields.

2. The computer-readable medium of claim **1**, wherein the plurality of primitive data types include scalar data types and composite data types;

wherein the scalar data types include a Boolean data type, an integer data type, a real data type, a decimal data type, a timestamp data type, a string data type, a symbol data type, a binary large object data type, and a character large object data type; and

wherein the composite data types include a struct data type, a list data type, and an S-expression data type.

3. The computer-readable medium of claim **2**, wherein the primitive data types include an annotation data type for providing additional type information about the scalar data types.

4. The computer-readable medium of claim **3**, wherein the binary data encoding format corresponding to the annotation data type comprises a type descriptor subfield, a length descriptor subfield, an annotation list field, a value type descriptor subfield, a value length descriptor subfield, and a value content field.

5. The computer-readable medium of claim **2**, wherein the primitive data types include an annotation data type for providing additional type information about the composite data types.

6. The computer-readable medium of claim **5**, wherein the binary data encoding format corresponding to the annotation data type comprises a type descriptor subfield, a length descriptor subfield, an annotation list field, a value type descriptor subfield, a value length descriptor subfield, and a value content field.

7. A method of communicating digital information, the method comprising:

receiving a multi-record message including binary-encoded digital information, the multi-record message comprising:

a length of message field for representing a length of the multi-record message;

a binary format identification field for identifying binary data encoding formats; and

a number of value fields for representing component information parts, the number of value fields being based on the length of the multi-record message, wherein each of the value fields has a value type descriptor field, a length of data field, and a representation field;

decoding the component information parts of the binary-encoded digital information to retrieve the digital information based on the value fields; and

processing the retrieved digital information.

8. The method of claim **7**, wherein the binary-encoded digital information is generated according to binary data encoding formats, each binary data encoding format corresponding to one of a plurality of primitive data types.

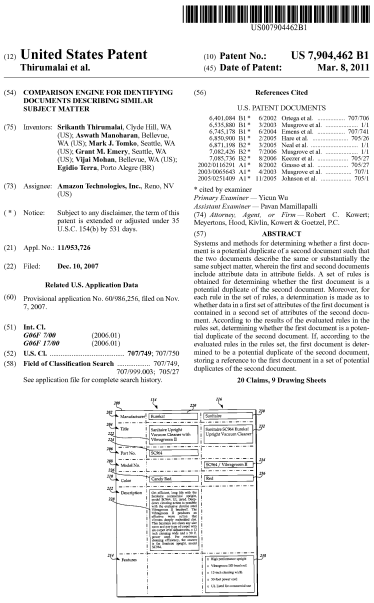
9. The method of claim **8**, wherein the plurality of primitive data types include scalar data types including a Boolean data type, an integer data type, a real data type, a decimal data type, a timestamp data type, a string data type, a symbol data type, a binary large object data type, and a character large object data type; and wherein the plurality of primitive data types further include composite data types including a struct data type, a list data type, and an S-expression data type.

10. The method of claim **9**, wherein the primitive data types include an annotation data type for providing additional type information about the scalar data types.

11. The method of claim **9**, wherein the primitive data types include an annotation data type for providing additional type information about the composite data types.

* * * * *

(b) Claims page of the patent 8,078,640



US007904462B1

(12) **United States Patent**
Thirumalai et al.

(10) **Patent No.: US 7,904,462 B1**
(45) **Date of Patent: Mar. 8, 2011**

(54) **COMPARISON ENGINE FOR IDENTIFYING DOCUMENTS DESCRIBING SIMILAR SUBJECT MATTER**

(75) Inventors: **Srikanth Thirumalai**, Clyde Hill, WA (US); **Aswath Manoharan**, Bellevue, WA (US); **Mark J. Tomko**, Seattle, WA (US); **Grant M. Emery**, Seattle, WA (US); **Vijai Mohan**, Bellevue, WA (US); **Egidio Terra**, Porto Alegre (BR)

(73) Assignee: **Amazon Technologies, Inc.**, Reno, NV (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 531 days.

(21) Appl. No.: **11/953,726**

(22) Filed: **Dec. 10, 2007**

Related U.S. Application Data

(60) Provisional application No. 60/986,256, filed on Nov. 7, 2007.

(51) **Int. Cl.**
G06F 7/00 (2006.01)
G06F 17/00 (2006.01)

(52) **U.S. Cl.** **707/749**; 707/750

(58) **Field of Classification Search** 707/749, 707/999.005; 705/27
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

6,401,084 B1 * 6/2002 Ortega et al. 707/706
6,535,880 B1 * 3/2003 Musgrove et al. 1/1
6,745,178 B1 * 6/2004 Emens et al. 707/741
6,850,901 B1 * 2/2005 Hunt et al. 705/26
6,947,930 B2 * 9/2005 Anick et al. 1/1
7,082,426 B2 * 7/2006 Musgrove et al. 1/1
7,085,730 B2 * 8/2006 Koenig et al. 705/27
2002/0116291 A1 * 8/2002 Grasso et al. 705/27
2003/0065643 A1 * 4/2003 Musgrove et al. 707/1
2005/0251499 A1 * 11/2005 Johnson et al. 705/1

* cited by examiner
Primary Examiner — Yicun Wu
Assistant Examiner — Pavan Mamillapalli
(74) *Attorney, Agent, or Firm* — Robert C. Kowert; Meyertons, Hood, Kivlin, Kowert & Goetzel, P.C.

(57) **ABSTRACT**

Systems and methods for determining whether a first document is a potential duplicate of a second document such that the two documents describe the same or substantially the same subject matter, wherein the first and second documents include attribute data in attribute fields. A set of rules is obtained for determining whether the first document is a potential duplicate of the second document. Moreover, for each rule in the set of rules, a determination is made as to whether data in a first set of attributes of the first document is contained in a second set of attributes of the second document. According to the results of the evaluated rules in the rules set, determining whether the first document is a potential duplicate of the second document. If, according to the evaluated rules in the rules set, the first document is determined to be a potential duplicate of the second document, storing a reference to the first document in a set of potential duplicates of the second document.

20 Claims, 9 Drawing Sheets

(c) First page of patent 7,904,462

ond product description, store a reference to the first product description in a set of potential duplicates for the second product description in a data store.

2. The computer-implemented system of claim **1**, wherein the evaluation of the set of rules comprises an evaluation in a direction from the first product description to the second product description and an evaluation in a direction from the second product description to the first product description.

3. The computer-implemented system of claim **2**, wherein the evaluation of the set of rules is performed in a symmetrical manner.

4. The computer-implemented system of claim **2**, wherein the evaluation of the set of rules is performed in a strictly symmetrical manner.

5. The computer-implemented system of claim **2**, wherein the evaluation of the set of rules is performed in an asymmetrical manner.

6. A computer-implemented system for determining whether a first product description is a potential duplicate of a second product description, wherein the first and second product descriptions are structured such that they include attributes that store attribute data in attribute fields, the system comprising:

a processor;

a memory coupled to the processor; and

computer executable code stored at least in part in the memory, wherein the computer executable code, upon execution, causes the computer-implemented system to:

obtain a set of rules for evaluating whether the first product description is a potential duplicate of the second product description;

obtain data from a first set of attributes of the first product description, the first set of attributes being determined according the set of rules;

obtain data from a second set of attributes of the second product description, the second set of attributes being determined according to the set of rules, and wherein attribute fields within the first set of attributes and attribute fields within the second set of attributes are not identical;

evaluate the set of rules with regard to the data from the first set of attributes and the data from the second set of attributes; and

in response to the evaluation of the set of rules indicating that the first product description is a potential duplicate of the second product description, store a reference to the first product description in a set of potential duplicates for the second product description in a data store.

7. The computer-implemented system of claim **6**, wherein the computer executable code, upon execution, causes the computer-implemented system to evaluate the data from the first set of attributes with regard to the data from the second set of attributes according to a generalized Jaccard algorithm.

8. The computer-implemented system of claim **6**, wherein the computer executable code, upon execution, causes the computer-implemented system to evaluate the set of rules the set of rules according to a disjunction of an evaluation of a plurality of rules in the set of rules.

9. The computer-implemented system of claim **8**, wherein a rule of the plurality of rules comprises a set of rule conditions and wherein the computer executable code, upon execution, causes the computer-implemented system to evaluate the rule according to a conjunction of an evaluation of each condition within the set of rule conditions.

10. A computer-implemented method for determining whether a first document is a potential duplicate of a second document such that the two documents describe the same or substantially the same subject matter, and wherein the first and second documents include attributes that store attribute data in attribute fields, the method comprising:

obtaining a rule, wherein the rule associates a first set of attributes in the first document with a second set of attributes in the second document, and wherein attribute fields within the first set of attributes is not equivalent to attribute fields within the second set of attributes;

determining whether data from the first set of attributes is contained in the second set of attributes;

determining whether data from the second set of attributes is contained in the first set of attributes; and

storing a reference to the first document in a set of potential duplicates of the second document in response to determining that data from the first set of attributes is contained in the second set of attributes and data from the second set of attributes is contained in the first set of attributes.

11. A tangible computer-readable medium bearing computer-executable instruction which, when executed on a computing device having a processor and a memory, carry out a method for determining whether a first document is a potential duplicate of a second document such that the two documents describe the same or substantially the same subject matter, and wherein the first and second documents include attributes that store attribute data in attribute fields, the method comprising:

obtaining a rule, wherein the rule associates a first set of attributes in the first document with a second set of attributes in the second document, and wherein attribute fields from the first set of attributes are not the same as attribute fields from the second set of attributes;

determining whether data from the first set of attributes is included within the second set of attributes;

determining whether data from the second set of attributes is included within the first set of attributes;

storing a reference to the first document in a set of potential duplicates of the second document in response to determining that data from the first set of attributes is included within the second set of attributes and data from the second set of attributes is included within the first set of attributes.

12. An article of manufacture comprising a computer readable storage medium having program instructions stored thereon that, in response to execution by a computer system, cause the computer system to perform operations including:

receiving a first product description associated with a first product;

receiving a second product description associated with a second product;

selecting a rule set that defines a comparison between at least one token within a first set of attribute fields of the first product description and at least one token within a second set of attribute fields of a second product description, wherein the first set of attribute fields is different from the second set of attribute fields;

evaluating the rule set in a first direction from the first set of attribute fields to the second set of attribute fields;

evaluating the rule set in a second direction from the second set of attribute fields to the first set of attribute fields; and

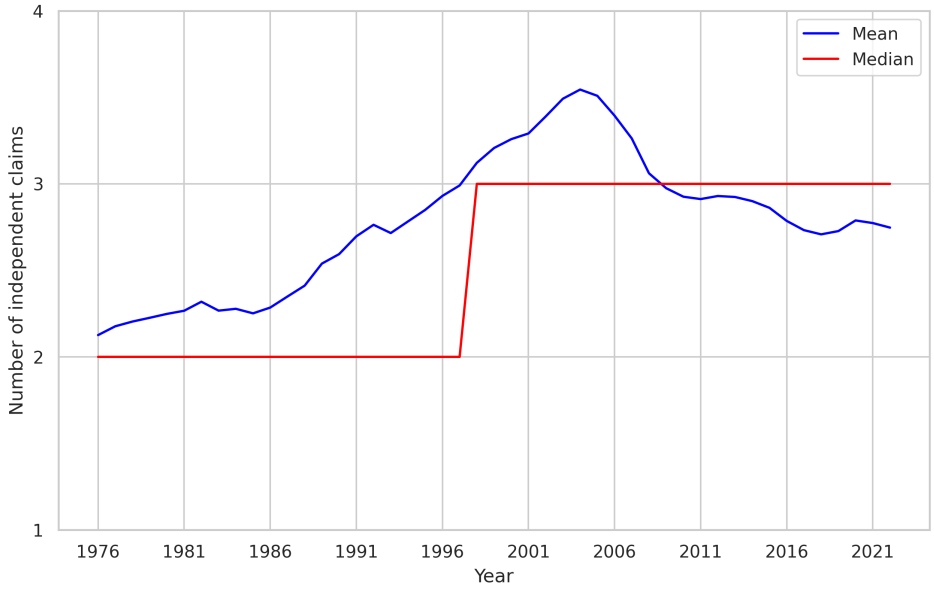
determining whether the first product is a duplicate of the second product in response to both evaluation operations.

13. The article of manufacture of claim **12**, wherein the second set of attribute fields is larger than the first set of attribute fields.

(d) Claims page of patent 7,904,462

This figure shows the first and last pages of Amazon's patents 7,904,462 and 8,078,640. Even though both patents are in USPC class 707, their similarity is 49%.

O.A. Figure 3: Number of Independent Claims Over Time



This figure shows the mean and median numbers of independent claims for each grant year from 1976 to 2022.

O.A. Table 1: Fifteen CPC Subclasses for AI Patents

CPC Subclass	CPC Subclass Title	Patent Count
G06F	Electric Digital Data Processing	423397
H04L	Transmission Of Digital Information, E.G. Telegraphic Communication	168101
G06T	Image Data Processing Or Generation, In General	139828
G06Q	Information And Communication Technology [Ict] Specially Adapted For Administrative, Commercial, Financial, Managerial Or Supervisory Purposes; Systems Or Methods Specially Adapted For Administrative, Commercial, Financial, Managerial Or Supervisory Purposes, Not Otherwise Provided For	133395
H04N	Pictorial Communication, E.G. Television	108312
A61B	Diagnosis; Surgery; Identification	102992
G06K	Graphical Data Reading ; Presentation Of Data; Record Carriers; Handling Record Carriers	85045
G06N	Computing Arrangements Based On Specific Computational Models	53506
H04W	Wireless Communication Networks	40935
G10L	Speech Analysis Or Synthesis; Speech Recognition; Speech Or Voice Processing; Speech Or Audio Coding Or Decoding	35424
G01N	Investigating Or Analysing Materials By Determining Their Chemical Or Physical Properties	23885
G05B	Control Or Regulating Systems In General; Functional Elements Of Such Systems; Monitoring Or Testing Arrangements For Such Systems Or Elements	21356
B60W	Conjoint Control Of Vehicle Sub-Units Of Different Type Or Different Function; Control Systems Specially Adapted For Hybrid Vehicles; Road Vehicle Drive Control Systems For Purposes Not Related To The Control Of A Particular Sub-Unit	17469
G05D	Systems For Controlling Or Regulating Non-Electric Variables	14936
G16H	Healthcare Informatics, I.E. Information And Communication Technology [Ict] Specially Adapted For The Handling Or Processing Of Medical Or Healthcare Data	14909

This table presents the top fifteen Cooperative Patent Classification (CPC) subclasses in which United States Patent and Trademark Office (USPTO) officials have classified artificial intelligence (AI) patents. (Giczy, Pairolero, and Toole, 2022).