Beyond Patent Ownership: Learning About Technological Usefulness*

Jesus Gorrin† Rory Mullen‡

First version: June 7, 2025
This version: August 12, 2025

Abstract

Technology is central to economics, but current datasets lack the scale, scope, span, and specificity economists need. We apply natural language processing and positive-unlabeled machine learning to descriptions of U.S. patents and public firms over nearly three decades to create a firm-level technology dataset that is unmatched in its combination of these qualities. For the first time, we reveal the core technologies of non-patenting firms. We estimate the value these firms derive from technological innovation, and argue that stock market participants inefficiently process their technological information, enabling profitable trading strategies.

JEL: O32, G12, C55, O34, C81
Keywords: technology, patents, firm-level, positive-unlabeled learning, textual analysis, natural language processing, spillovers, momentum, non-patenting firms

*Working paper: comments welcome. Results are preliminary and subject to revision. We thank Vashisht Bhatt for excellent research assistance, Junxi Liu for contributions to an earlier version of the paper circulated under the title "Firm Technology Usage," and the Gillmore Centre for Financial Technology at the University of Warwick for financial support.

†Warwick Business School, University of Warwick, Jesus.Gorrin@wbs.ac.uk
‡Warwick Business School, University of Warwick, Rory.Mullen@wbs.ac.uk

1 Introduction

Technology is central to economics, driving everything from growth and business cycles (Solow, 1960; Kydland and Prescott, 1982) to labor markets and financial markets (Autor, Levy, and Murnane, 2003; Kogan, Papanikolaou, Seru, and Stoffman, 2017). Yet the majority of firms—those that do not patent—remain technologically veiled, obscuring how technology is used throughout much of the economy. Current technology datasets lack the scale, scope, span, and specificity that economists need, and often focus on the small minority of firms that own intellectual property rather than the broader universe of firms that use technology. While many technology datasets exist, each has important limitations.

In this paper, we construct a novel dataset linking firms to technologies using positive and unlabeled machine learning. Our firm-level technology dataset offers an unprecedented combination of scale, scope, span, and specificity. In scale, it covers all U.S. public firms and a sample of 50,000 U.S. utility patents per year, amounting to hundreds of millions of firm-patent pairs. In scope, it covers thousands of distinct technology categories curated by patent examiners with subject-specific expertise, representing all major areas of economically important innovation. In span, it covers nearly three decades of innovation, including the rise of internet and mobile computing, advances in biotechnology and medical technology, and the growth of renewable energy and electric vehicles. In specificity, it characterizes the usefulness of each individual patent to each firm as a continuous probability, allowing us to characterize each firm's technological associations in granular detail and by degrees. This contrasts with the available binary indicators—such as patent ownership or innovation survey responses—which frequently cover few firms or few technologies, lack specificity, or convey no variation in the strength of association.

Existing datasets typically rely on patent ownership or innovation surveys. Patent ownership datasets have several limitations. First, few firms receive patent grants—fewer than 1% of firms in the U.S. Census Bureau's Business Register (Graham, Marco, and Miller, 2015) and only 15.64% of CRSP firms in an average year. Second, patent-owning firms are atypical—they are unusually large and concentrate in manufacturing (Mezzanotti

and Simcoe, 2023). Third, patent ownership does not always signal technology use—firms often file patents for reasons other than protecting actively used technology, including patent blocking, use as bargaining chips in negotiations, and prevention of suits (Cohen, Nelson, and Walsh, 2000; Moore, 2005). Fourth, patent ownership does not imply exclusive access—intellectual property rights must be asserted through costly litigation (Lemley and Shapiro, 2005). Fifth, patent owners regularly grant access to others—around 40% of patents are embodied in commercial products (Argente et al., 2023) and around 4% are reassigned annually (Graham, Marco, and Myers, 2018). These limitations make patent ownership data unsuitable for many economic applications. Innovation surveys overcome some limitations but lack the specificity of patent data. For example, survey respondents may report *that* innovations occurred but rarely describe *which* specific innovations.

To overcome these limitations, we construct a new dataset that emphasizes technological usefulness use over patent ownership. We start from a simple premise: if the language that describes a patent is similar to the language that describes a firm, the patent is probably useful for the firm—regardless of who owns it. Building on this premise, we develop a novel methodology that combines natural language processing with techniques borrowed from an area of machine learning called positive and unlabeled learning. Positive and unlabeled learning is new to the economics literature, and allows us to view patent ownership as a positive signal that a patent is useful to the firm that owns it, without viewing the patent as useless to other firms. With positive and unlabeled learning, we train a classifier on positive usefulness labels exclusively, and use the classifier to predict the usefulness of patents to all firms, regardless of ownership.

We construct our dataset in two steps. First, we measure similarities in the language used to describe patents and firms. The patent descriptions come from a corpus of patent filings with the U.S. Patent and Trademark Office. The firm descriptions come from a corpus of annual reports filed with the U.S. Securities and Exchange Commission. These sources have each been extensively studied in the economics literature, but they have not been combined with the aim of constructing a comprehensive firm-level technology dataset covering non-patenting and patenting firms for use in economic research. We

represent each description numerically using traditional term frequencies and inverse document frequencies and using a modern natural language model built on the transformer architecture. We then compute similarity scores between these numerical representations for all firm-patent pairs. While the techniques we use in the first step of our methodology are now commonplace in the economics literature, we take an important second step that is novel.

In the second step, we use similarity scores to estimate probabilities that each firm finds each patent useful. We refer to these estimates as *usefulness probabilities*. Traditional supervised machine learning requires both positive and negative labels; in our setting, these labels would indicate that a firm finds a patent useful or useless, respectively. But these labels are not always available in our setting. When a patent is owned by a firm, we assume the firm is likely to find the patent useful, and assign a (possibly noisy) positive label to the pair. But for other firm patent pairs, we cannot immediately assign negative labels, because patented innovations may be useful to firms who do not own them. Lacking negative labels, we turn to positive and unlabeled machine learning techniques specifically designed to overcome this challenge.

Positive and unlabeled learning, while novel in the economics literature, is well-established in the machine learning literature. We adopt a classic two-stage procedure from this literature, which was proposed by Liu et al. (2002). In the first stage, we assign negative labels to a random selection of positive firm-patent pairs, and use these "spies" to estimate a probability threshold for identifying firm-patent pairs where the patent is reliably useless to the firm. We then train a second-stage classifier on positive and reliably-negative firm-patent pairs. The spy procedure has an intuitive appeal and imposes few assumptions on the labeling process.

Using this approach, we uncover surprising similarities between non-patenting and patenting firm types. After conditioning on industry and firm size, we find that within-type variation in technological associations substantially exceeds between-type variation. We also find difference: non-patenting firms associate with broader but shallower technological portfolios and exhibit higher rates of technological instability over time. The technological

associations we uncover have economic consequences: technology momentum portfolios of non-patenting firms generate monthly alphas of 1.97%, significantly exceeding strategies restricted to patenting firms. The superior returns appear to be driven by slower information diffusion about non-patenting firms' technological profiles. Event study evidence confirms that technological spillovers extend beyond patent ownershipfirms experience positive abnormal returns of approximately 0.9 basis points per useful patent over 30 days following patent announcements, even for patents they do not own. Together, these results demonstrate that technological associations matter for firm value, even for non-patenting firms.

Related Literature. We contribute to three strands of literature. First, we build on efforts to construct firm-level technology and innovation datasets. Early approaches used direct profiling methods. The UN Industrial Development Organization's Profiles of Manufacturing Plants documented equipment use across manufacturing plants worldwide in the late 1960s and early 1970s (United Nations, 1971), but these profiles lacked standardization, covered few plants per industry, and the series was discontinued after three editions. The U.S. Census Bureau's Survey of Manufacturing Technology in the late 1980s and early 1990s systematically surveyed over 10,000 U.S. manufacturing plants about their use and planned adoption of seventeen advanced technologies (see Dunne, 1994, for a description). However, the survey focused on a narrow set of technologies and was also discontinued after three editions.

Standardized innovation surveys emerged to provide broader, ongoing coverage of firm innovation activities. The E.U. Community Innovation Survey (launched in 1992) and U.S. National Science Foundation Business R&D and Innovation Survey (launched in 2008) now measure firm-level innovation under a common framework. These surveys represent a significant advance in scale and span over earlier profiling efforts. Yet their indicators lack scope and specificity, relying on firms' self-reported yes/no responses and coarse categorizations ("product" versus "process") that limit the utility of the data.

In parallel, a tradition developed around patent-based measures of firm technology. A

major NBER initiative launched in the 1980s created the first large-scale patent ownership dataset, with early contributions collected in Griliches (1987), a retrospective in Hall, Jaffe, and Trajtenberg (2001), and an update in Arora, Belenzon, and Sheer (2021). Patent datasets offer advantages in scale, scope, span, and specificity compared to survey approaches. However, patent data exclude the majority of firms—those that do not patent—creating significant coverage gaps. We contribute to this literature by providing highly-detailed technological profiles of non-patenting firms using positive-unlabeled machine learning. Unlike Bloom, Hassan, Kalyani, Lerner, and Tahoun (2021), who study 29 disruptive technologies, we study a comprehensive set of legacy and disruptive technologies.

Second, we contribute to a literature that uses the datasets described above to study how technology and innovation relate to firm productivity and performance. The literature is too large to survey here (Lerner and Seru, 2021, identify over 80 papers in top economics journals between 2005 and 2020 using patent data), but seminal contributions include: Griliches (1990), who surveys patents as economic indicators; Lerner (1994), who examines patent scope and firm value; Hall, Jaffe, and Trajtenberg (2005), who show patent citations better predict firm value than counts; Bloom, Schankerman, and Van Reenen (2013), who identify productivity spillovers from patented innovations; Kogan et al. (2017), who measure how innovation drives firm growth and aggregate productivity; and Akcigit and Kerr (2018), who examine how different innovation types affect firm productivity. We contribute by comparing, for the first time, the technological profiles of non-patenting and patenting firms. We find that the technology use of non-patenting firms rivals that of patenting firms, after conditioning on industry and firm size, with both groups exposed to similar shocks and spillovers.

Third, we contribute to the growing literature on natural language processing and machine learning methods in economics and finance. Gentzkow, Kelly, and Taddy (2019) and Ash and Hansen (2023) survey recent work in economics, and Loughran and McDonald (2020b) and Kelly and Xiu (2023) survey recent work in finance. Papers studying firms' regulatory filings include Hoberg and Phillips (2016) on dynamic product-market industries,

Hoberg and Phillips (2018) on industry momentum, Lopez-Lira (2019) on risk factors for asset pricing, and Loughran and McDonald (2020a) and Cohen, Malloy, and Nguyen (2020) on firm complexity. Papers studying patent descriptions include Myers and Lanahan (2022) on R&D spillovers, Lerner et al. (2021) on patented financial technologies, Kogan et al. (2021) on technology and labor productivity, and Kakhbod et al. (2024) on innovation spillovers. To our knowledge, no prior work has linked regulatory filings with patent data to produce a comprehensive firm-level technology dataset covering non-patenting and patenting firms for use in economic research.

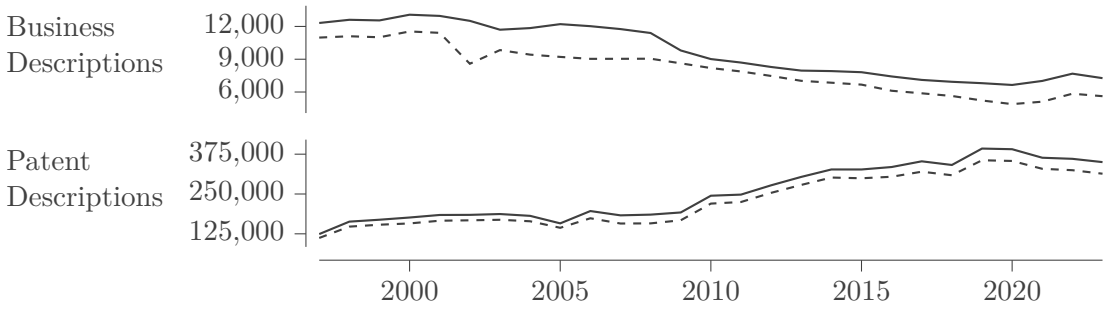
The paper is organized as follows. Section 2 describes our data sources. Section 3 describes our methodology for classifying patents as useful to firms. Section 4 compares the technological profiles of non-patenting and patenting firms. Section 5 studies a technology momentum investment strategy. Section 6 concludes.

2 Data Sources

Our study combines traditional financial data with textual data on public firms and patent grants in the United States. The financial data are CRSP daily and monthly stock files and Fama-French return factors (Fama, 2023). The textual data are business descriptions extracted from annual reports from the Electronic Data Gathering, Analysis, and Retrieval database (EDGAR) of the U.S. Securities and Exchange Commission (SEC) and patent descriptions from the PatentsView database of the U.S. Patent and Trademark Office (USPTO).

Figure 1 plots annual counts of business and patent descriptions. These descriptions offer broad coverage of technology users and technological innovations over an extended period. The SEC data describes the products, operations, and intellectual property of firms, while the USPTO data describes the novelty, function, and intended use of technology. These complementary perspectives motivate our basic hypothesis: that a patent is likely to be useful to a firm when the patent and firm are described in similar language—even if the firm does not own the patent.

Figure 1: Annual Business and Patent Description Counts



Notes. The figure shows annual counts of SEC business descriptions and USPTO patent descriptions. For business descriptions, the solid line plots the total number of available annual reports, while the dashed line plots the number of business descriptions we were able to extract. For patent descriptions, the solid line plots the total number of patent grants, while the dashed line plots the number of utility patent grants. We sample 50,000 utility patents randomly from each yearly total.

We provide details about business descriptions from SEC filings in Section 2.1 below, and about patent descriptions from the USPTO filings in Section 2.2. We describe the Cooperative Patent Classification (CPC) system used to categorize patents in Section 2.3, and we describe our approach to identifying patent owners in Section 2.4.

2.1 Business Descriptions

The Securities Exchange Act of 1934 mandates that certain firms register their securities and file annual reports with the SEC (U.S. Congress, 1934). The mandate applies to firms with securities registered under Section 12 of the Act, which includes firms listed on a national securities exchange as well as unlisted firms that own assets exceeding $10 million in value and issue equity securities held by more than 2,000 persons or more than 500 persons who are not accredited investors.[1]

The SEC requires annual reports to be filed on Forms 10-K and 20-F. Form 10-K filings must include non-financial information as outlined in Regulation S-K (U.S. SEC, 2013). Form 10-K is used by domestic U.S. issuers, while Form 20-F is used by foreign private issuers with shares listed on U.S. national securities exchanges and follows its

[1]The act is amended from time to time, with updated thresholds and other changes. Exemptions or modified requirements apply for certain types of securities and issuers, such as securities issued by banks, savings and loan associations, and religious, educational, or charitable organizations.

own form-specific disclosure requirements. Item 101 of Regulation S-K requires 10-K filing firms to disclose their dominant business segments and markets served, competitive conditions, principal products and services, material contracts and customer dependencies, government contracts, material government regulation and compliance costs, distribution methods, sources and availability of raw materials, research and development activities, and patents, trademarks, and other intellectual property.[2] Comparable disclosures are required in Form 20-F.[3] In prior research, Hoberg and Phillips (2016) have primarily focused on product market information contained in business descriptions, but Regulation S-K requires disclosures to contain operational and technological information that goes beyond product markets.

We use the SEC's annual index files to identify relevant filings; these files record an identifier, the filing date, form type, and company name for each filing.[4] We then scrape the full text of relevant filings and extract business descriptions from these. Over the period 1997 to 2023, we achieve an average annual extraction rate of 82%. We obtain business descriptions for a total of 29,807 unique firms or an average of 8,048 unique firms per year over the sample period.

2.2 Patent Descriptions

The Patent Act of 1952 governs patenting in the United States, authorizing the USPTO to issue patents for inventions that are new, useful, and non-obvious (U.S. Congress, 1952). Usefulness is broadly defined under U.S. patent law: the invention must provide a specific, substantial, and credible utility, including a practical application in industry or research, solving a real-world problem, or performing a useful function (USPTO, 2013). Patent applications must include a description that establishes the usefulness of the innovation,

[2]For smaller reporting firms, the disclosure requirements are simplified. They must still report their principal products, markets, and competitive conditions, government regulations, environmental compliance, key suppliers, and material customer dependencies, though the level of detail required is reduced (U.S. SEC, 2013, Item 101(h)).

[3]Form 10-K business descriptions appear in Item 1. Form 20-F business descriptions appear in Item 4, which provides an extensive overview with information on the company's operations, products, markets, raw materials, important dependencies, competitive position, regulatory context, organization structure, and property, plant, and equipment.

[4]The SEC maintains a complete list of all filings on its full-index web page (link).

that is detailed enough to enable any person skilled in the relevant field to make and use the invention. As such, patent descriptions must provide enough detail for reviewers—and later, for researchers like ourselves—to infer whom the invention will benefit and the context in which it will be applied.

We obtain digital records of patent filings made available on the USPTO's PatentsView platform, covering all filings from 1976 to present. The records include detailed patent descriptions and metadata including patent title, assignee, application date and grant date, and Cooperative Patent Classification (CPC) codes. We focus on utility patents granted between 1997 and 2023, totaling 6,226,101 filings over the sample period. For our main analysis, we randomly sample 50,000 patents per year, representing over 20% of the yearly average of 230,596 utility patents in the PatentsView dataset.

2.3 Cooperative Patent Classification

The USPTO has a congressional mandate to maintain a classification system for patents, to facilitate prior art searches and examination (U.S. Congress, 1836; U.S. Congress, 2011). The Cooperative Patent Classification (CPC) system currently serves this purpose; it builds on the International Patent Classification (IPC) system, adopting identical rules and principles but extending the number of classification codes to more than 260,000 (Simmons, 2014; EPO and USPTO, 2017; Lobo and Strumsky, 2019).[5] Patents are frequently reclassified, and PatentsView maintains both current and historical CPC classifications. We use the current classification to ensure consistency over our sample period, in line with previous studies (Strumsky, Lobo, and Van der Leeuw, 2012; Lobo and Strumsky, 2019).

Following IPC principles, the CPC classifies patents according to either intrinsic function or particular application (WIPO, 2024, paragraphs 81-87), and classifications often, but not always, cross industry boundaries. For example, patents for mixing and

[5]The CPC emerged from a collaboration with the European Patent Office (EPO) and was formally adopted by the USPTO in 2015, replacing a legacy system, the United States Patent Classification (USPC), that had been in use since 1836 (Simmons, 2014). The CPC is revised multiple times per year to keep pace with technological change, and Notices of Change are published regularly to the USPTO website (link).

Table 1: Descriptive Statistics for Levels of the CPC

CPC Level	Number of Categories	Average Annual Patents per Category					
		Median	Min	Max	IQR	Skewness	Kurtosis
Section	8	5,360	356	14,017	5,630	0.48	−1.03
Class	121	99	1	6,576	282	4.43	21.35
Subclass	577	21	1	3,991	62	9.15	109.50
Group	3,887	3	1	1,075	7	11.03	195.72

Notes. The table describes four levels of the Cooperative Patent Classification (CPC) system: Section, Class, Subclass, and Group. For each level, we report the number of categories, and the minimum, median, and maximum number of patents per category, as well as the interquartile range (IQR), skewness, and kurtosis, as averages of yearly statistics over the sample period, 1997 to 2023. The skewness and kurtosis measures indicate distributional asymmetry and tail heaviness, respectively, particularly at finer levels of classification.

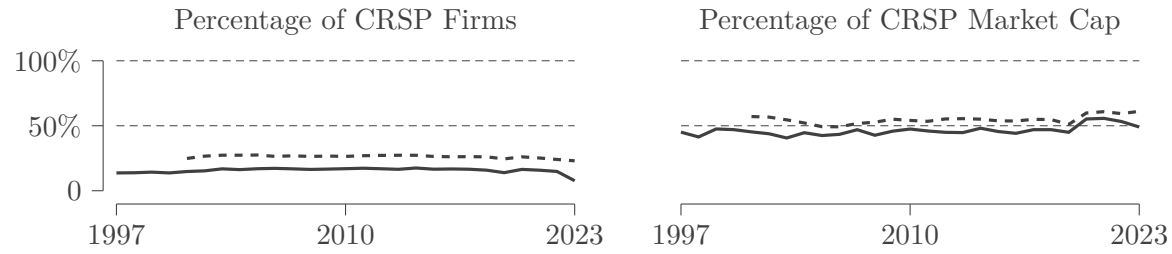
agitation are grouped together, regardless of whether they are used for washing clothes, mixing paint, or churning butter (Simmons, 2014). Similarly, valves characterized by aspects of their construction are grouped together, independent of the fluid they control; however, a valve specially for use in a heart, as a particular application, would be classified separately (WIPO, 2024, paragraph 85).[6]

The CPC organizes patents into a hierarchy with four main levels: Sections, Classes, Subclasses, and Groups, progressing from broad to narrow categories. Table 1 characterizes these levels. Our sample of 50,000 patents per year includes patents from 8 Sections, 121 Classes, 577 Subclasses, and 3,887 Groups.[7] Lower levels of the CPC are highly skewed with a few large and many smaller categories. For example, the skewness coefficient rises from 0.48 for Sections to 11.03 for Groups, and the ratio of maximum to median annual patent count rises from 2.62 for Sections to 324.86 for Groups. Similar skew also arises in other patent classification systems and may result from stochastic growth and category-splitting dynamics of the system over time (Lafond and Kim, 2019).

[6]The example from Simmons (2014) refers to the legacy USPC system, but carries over to the CPC, where patents for mixing and agitation are classified under Subclass B10F (link). Likewise, the example of values from the WIPO's IPC manual carries over to the CPC, where, for instance, valves with pivoted discs or flaps are classified under F16K 1/18 (link) and heart valves are classified under A61F 2/24 (link).

[7]The CPC system distinguishes between Main Groups (denoted by classification codes ending in /00) and Subgroups (with additional digits after the slash). Throughout this paper, "Group" refers exclusively to Main Groups, which number 3,887 in our sample. While the complete CPC system contains over 260,000 categories when including all Subgroups, we focus on Main Groups as they provide sufficient granularity for our analysis.

Figure 2: Prevalence of Patent Owners in CRSP



Notes. The figures show the number and market capitalization of patent-owning firms relative to the number and market capitalization of SEC filing firms in CRSP monthly data from 1997 to 2023. The left figure shows the number of patent owners divided by the number of CRSP SEC filing firms each year, while the right figure shows the total market capitalization of patent owners divided by the total market capitalization of CRSP SEC filing firms each year. Solid lines plot values for patent owners defined as firms with at least one patent granted in a given year, while dashed lines plot values for patent owners defined as firms with at least one patent granted in a backward-looking five-year rolling window.

2.4 Patent Ownership

We identify patent owners by matching disambiguated assignee names from USPTO patent grants with company names from SEC annual reports using natural language processing. Our approach essentially follows the methodology pioneered by Bound, Cummins, Griliches, Hall, Jaffe, et al. (1982), while adopting recent refinements introduced by Kogan et al. (2017) and Arora, Belenzon, and Sheer (2021). While we adopt a similar name-matching approach to establish initial patent-to-owner links, our work departs from this literature by treating these links as a first step that is followed by a machine learning classification step, rather than an end goal.

We pre-process patent assignee names and company names by standardizing cases, removing non-standard characters and punctuation, and stripping any suffixes that do not aid in matching. We then use exact and fuzzy matching techniques to identify potential matches between the pre-processed names, considering common substrings and other similarity metrics. Initial matches are supplemented with a set of manually curated matches for the most active patenting firms, and the results are iteratively refined to ensure reliability. Only high-confidence matches are retained, with ambiguous cases flagged for manual review.

We identify 712.48 unique patenting firms in an average year and 19,237 in total

over the period 1997 to 2023, matching 9,653.00 out of 50,000 sampled patents in an average year and 260,631 patents in total. Only 15.64% of CRSP firms receive patent grants in an average year, and while these patenting firms are often large, they account for just 46.29% of CRSP market capitalization—less than half.[8] Even among patenting firms, concentration is high: while the average firm receives 13.55 patents annually, the yearly maximum reaches 1,078.37 patents. Figure 2 shows that the low share of patenting firms has remained remarkably stable from 1997 to 2023. These patterns highlight the importance of expanding technology research beyond patent ownership to include the majority of firms that use technology but do not own patents.

Table 2 compares non-patenting to patenting firms by industry group and size class. Industry groups are defined by SIC codes: Resource (0100–1799 and 4900–4999), Manufacturing (2000–3999), Service (4000–4899, 5000–5999, and 7000–8999), and Finance (6000–6399 and 6411). The table excludes real estate, holding companies, public administration, and firms with missing SIC codes. Size classes are based on market capitalization across all industries within each year: Large Cap (top 10%), Mid Cap (next 20%), Small Cap (bottom 70%), and Private (firms with SEC filings but no CRSP data). The table reports firm counts, market capitalization, and industry shares for each category.

Patenting activity varies substantially across industries. Patenting firms account for 74.50% of manufacturing market capitalization, 45.35% of services, and only 18.53% of finance. Over one in three manufacturing firms patent, compared to one in ten service firms and fewer than one in fifty finance firms. Large firms dominate patenting across all industries, accounting for 48.90% of total market capitalization while representing only 2.42% of all firms. These patterns show that focusing only on patenting firms excludes most firms, particularly in service-oriented industries. As we show later, however, non-patenting and patenting firms have surprisingly similar technological profiles within industry and size groups.

[8]The numbers rise to 25.07% of CRSP firms and 54.08% of CRSP market capitalization when firms with patents granted in five-year rolling windows are included. For comparison, Lee et al. (2019) use firm-patent matches from Kogan et al. (2017) and report an annual average of 956 patent owners accounting for 52.56% of CRSP market capitalization over the period 1963 to 2012, similar to what we find.

Table 2: Comparison of Non-Patenting and Patenting Firms by Industry and Size

Industry Group	Size Class	Firm Count		Firm Size		Industry Share	
		NP	P	NP	P	NP	P
Finance	Large	61.67	8.74	29.20	54.42	61.80	18.12
	Mid	125.26	2.42	3.05	3.57	13.74	0.38
	Small	523.96	3.40	0.32	0.57	5.93	0.08
	Private	414.22	2.44	—	—	—	—
	All	1,125.11	16.48	3.03	35.19	81.47	18.53
Service	Large	87.41	40.37	24.35	59.62	37.38	41.96
	Mid	237.52	43.93	3.07	3.28	12.35	2.64
	Small	799.89	98.37	0.37	0.42	4.92	0.75
	Private	828.33	38.70	—	—	—	—
	All	1,953.15	221.37	2.92	16.47	54.65	45.35
Manufacture	Large	49.48	121.52	29.79	46.06	17.67	66.22
	Mid	155.44	169.41	3.05	3.21	5.36	6.09
	Small	736.52	471.56	0.29	0.38	2.47	2.19
	Private	653.26	126.11	—	—	—	—
	All	1,594.70	888.59	2.30	8.06	25.50	74.50
Resource	Large	31.33	7.19	15.19	24.60	50.28	18.33
	Mid	68.52	7.62	3.13	2.99	22.37	2.39
	Small	162.78	13.74	0.36	0.39	6.13	0.59
	Private	415.93	10.41	—	—	—	—
	All	678.56	38.67	2.72	6.60	78.78	21.22
All	Large	229.89	177.81	25.51	49.54	32.38	48.90
	Mid	586.74	222.81	3.08	3.23	9.65	3.88
	Small	2,223.15	586.81	0.33	0.38	3.89	1.30
	Private	3,121.19	193.07	—	—	—	—
	All	6,160.96	1,180.52	2.71	10.10	45.92	54.08

Notes. The table compares non-patenting firms (NP) and patenting firms (P) by industry group and size class. Firm count is the number of firms of each type. Firm size is the average market capitalization (in millions) for firms of each type. Industry share is the total market capitalization for firms of each type expressed as a percentage of the total market capitalization of each industry group. These statistics are computed by industry group and size class. Industry groups are defined by SIC four-digit codes: Resource includes SIC 0100–1799 and 4900–4999; Manufacture includes SIC 2000–3999; Service includes SIC 4000–4899, 5000–5999, and 7000–8999; and Finance includes SIC 6000–6399 and 6411. We exclude firms with SIC codes 6400–6410, 6412–6499, 6500–6599, 6700–6799, 9000–9999, and firms with missing SIC codes from the table. Size classes are determined by market capitalization across all industries within each year: Large Cap (top 10%), Mid Cap (next 20%), Small Cap (bottom 70%), and Private (firms with SEC filings that do not appear in CRSP). We define patenting firms as firms with patents granted in five-year rolling windows.

3 Estimating Usefulness Probabilities

Using textual similarity scores between business and patent descriptions, we train a binary classifier to estimate the probability that a given patent would, in principle, be useful to a given firm—even if the firm does not own the patent. We refer to the estimates as *usefulness probabilities*. For training, we use a two-stage positive and unlabeled machine learning technique. Despite the growing importance of machine learning in economics and finance, we believe our paper is the first in these fields to use positive and unlabeled learning.

We interpret patent ownership as a (possibly noisy) signal of usefulness. Under this assumption, we assign positive usefulness labels to firm-patent pairs where the firm owns the patent. All remaining firm-patent pairs, which constitute the vast majority of pairs, are left unlabeled. We then identify "reliably negative" pairs from the unlabeled set and assign them negative labels in the first stage. In the second stage, we use logistic regression to train a binary classifier on the positive and reliably negative labeled data. We evaluate our model on a test set of firm-patent pairs that were not used in training, using a performance metric appropriate for positive and unlabeled learning. This approach allows us to bring formal statistical tools to the patent usefulness classification problem, revealing for the first time the technological profiles of non-patenting firms.

3.1 Measuring Document Similarity

The core features used by our classifier to predict usefulness are textual similarity scores between business descriptions and patent descriptions. To obtain these features, we vectorize the text of descriptions using three methods: term frequency (TF), term frequency-inverse document frequency (TF-IDF), and Sentence-BERT (SBERT). TF and TF-IDF are traditional methods that focus on word frequencies within and across documents. In contrast, SBERT is a modern pre-trained transformer model that produces embeddings (numerical vectors) capturing the contextual and semantic meanings of text.

For each of the three document vectorization methods, we compute the cosine similarity

between all possible pairs of business descriptions and patent descriptions. We use the resulting similarity scores as the core features in our positive and unlabeled learning classifier. We estimate the classifier separately for each year, considering only contemporaneous pairs of business and patent descriptions. We discuss document vectorization in more detail in Appendix A.1, and cosine similarity scoring in Appendix A.2. In addition to the similarity scores, we include indicators for SIC Divisions and CPC Sections, as well as textual characteristics as control features in some specifications of our classifier.

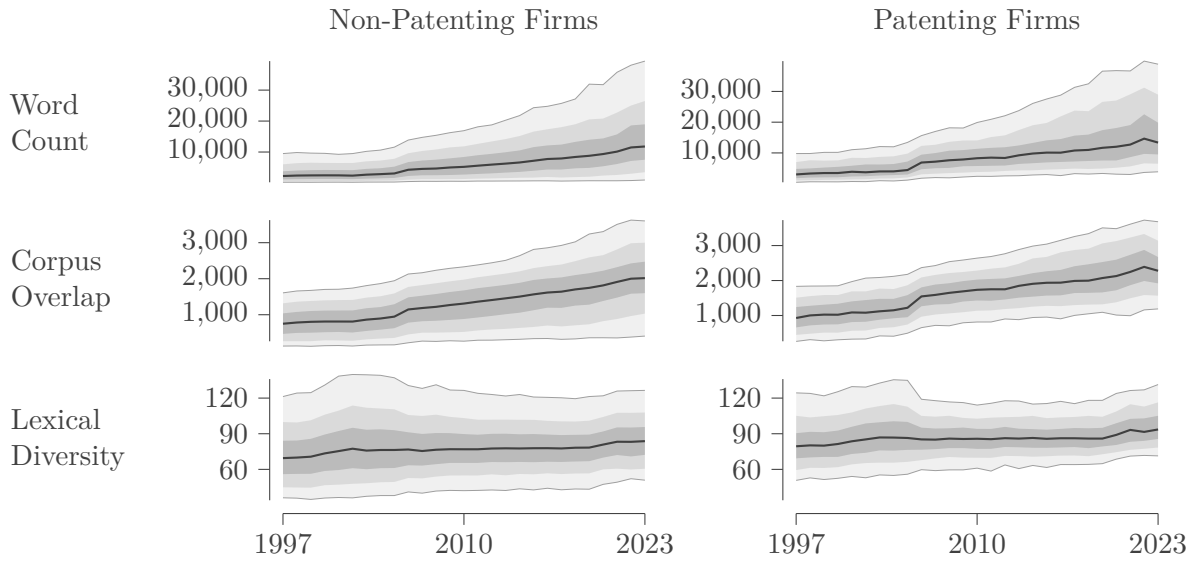
Including document-level textual characteristics as control features is important because cosine similarity can be sensitive to superficial textual properties such as document length, writing style, or the presence of non-technological language (Brown and Tucker, 2011). Without accounting for these factors during the classification step, we risk conflating true technological relevance with stylistic or structural artifacts of the documents.

The three textual characteristics we examine are word count (the number of words after removing standard English stop words), corpus overlap (the number of unique words from one corpus that appear more than once in the other corpus), and lexical diversity (measured using the MTLD metric of McCarthy and Jarvis 2010). Figure 3 shows the cross-sectional distributions of these characteristics over time. Unlike traditional firm-level characteristics like market capitalization and industry assignment, which differ markedly between non-patenting and patenting firms as Table 2 shows, these textual characteristics show more moderate differences.

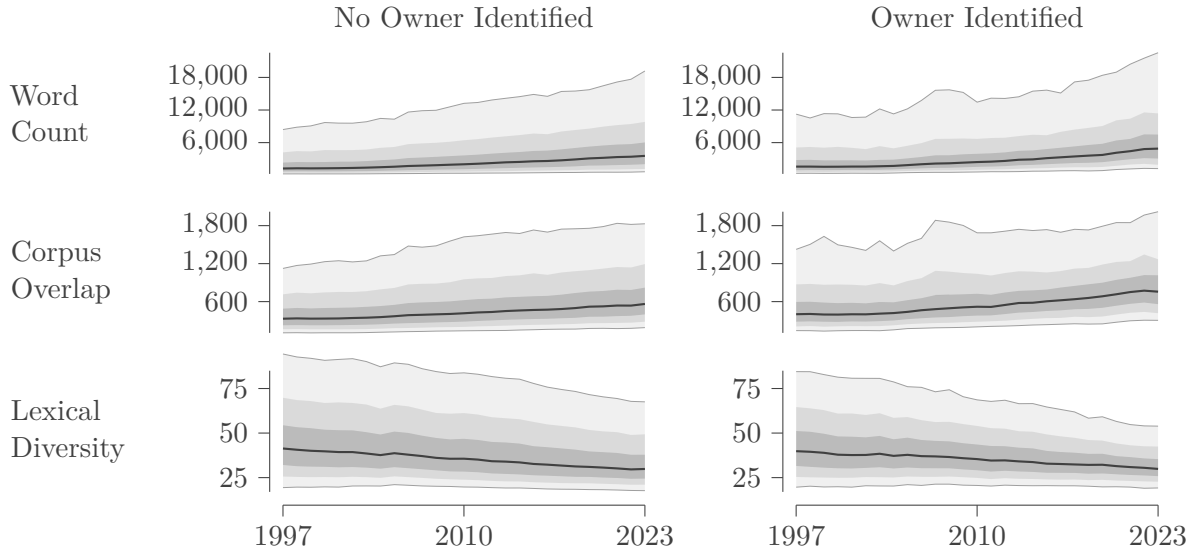
Panel 3a shows textual characteristics of business descriptions for non-patenting firms (left) and patenting firms (right). Averaging across years, we find cross-sectional median word counts of 5,700.7 for non-patenting firms and 7,895.2 for patenting firms—a difference of 38.5%. Corpus overlap medians are 1,309.1 versus 1,632.3—a difference of 24.7%. Lexical diversity medians are 77.0 versus 85.9—a difference of 11.6%. The plots show broadly similar interquartile ranges and 90-10 percentile ranges for the cross-sectional distributions of all three metrics for both firm types, and interquartile ranges within firm type are nearly twice the difference in medians across firm types for all three metrics. In other words, differences within type are much larger than differences across types.

Figure 3: Business and Patent Description Characteristics

(a) Business Descriptions



(b) Patent Descriptions



Notes. The plots in Panels 3a and 3b show the cross-sectional distribution of characteristics for business and patent descriptions, respectively, from 1997 to 2023. Black lines shows the median and shaded regions show the 25-75, 90-10, and 92.5-97.5 quantile ranges of values across descriptions each year. In Panel 3a, the left column shows the characteristics of business descriptions for non-patenting firms, while the right column shows the same characteristics for patenting firms. In Panel 3b, the left column shows the characteristics of patent descriptions for patents with no identified firm owner (these patents may be owned by universities, government, or firms that do not file annual reports with the SEC, including some foreign firms), while the right column shows the same characteristics for patents with identified owners. The plotted characteristics are word count, corpus overlap, and lexical diversity. We define word count as the number of words (including repeated occurrences) in a given description after removing standard English stop words, corpus overlap as the number of unique words in a given description from one corpus that appear more than once anywhere in the second corpus, and lexical diversity as the MTLD metric of McCarthy and Jarvis (2010).

Panel 3b shows textual characteristics of patent descriptions for patents without identified owners (left) and with identified owners (right) in our sample of firms.[9] Averaging across years, we find median word counts of 2,146.4 for patents with identified owners and 2,672.4 for those without—a 24.5% difference. Corpus overlap medians are 424.1 versus 539.1—a 27.1%. Lexical diversity medians are 35.3 versus 35.1—a −0.6% difference. Additionally, as with business descriptions, the cross-sectional distributions are broadly similar for patents with and without identified owners. Furthermore, interquartile ranges within patent type are nearly twice the difference in medians across patent types for all three metrics, indicating again that differences within type are much larger than differences across types.

Business descriptions are substantially longer, contain greater corpus overlap, and exhibit higher lexical diversity than patent descriptions. Both document types show gradual increases in median word count and corpus overlap over time. From 1997 to 2023, business description word counts grew 386.4% for non-patenting firms and 337.4% for patenting firms, while patent word counts grew 155.6% and 156.2% for patents without and with identified owners, respectively. Lexical diversity trends diverged: business descriptions grew by 19.8% (non-patenting) and 18.0% (patenting), while patent descriptions shrunk by -27.1% (no identified owner) and -26.9% (identified owner), respectively.

Importantly, these patterns hold consistently across both non-patenting and patenting firm types, indicating that textual characteristics are not driven primarily by firm type. This similarity matters for our empirical approach: our training data consists of patenting firms paired with patents that have identified owners—though we emphasize that most pairs are patenting firms that do not own the specific paired patent. If non-patenting firms or patents without identified owners had substantially different textual characteristics than the pairs we use for training, our classifier might perform poorly when applied to the broader universe. The evidence in Figure 3 suggests this is not a major concern—the textual features we use for classification are comparable across all firm and patent types,

[9]Patents without identified owners in our sample of firms may be owned by private individuals, governments, or firms that do not file annual reports with the SEC, which would include some small U.S. firms and foreign firms.

both in terms of cross-sectional distributions and time trends.

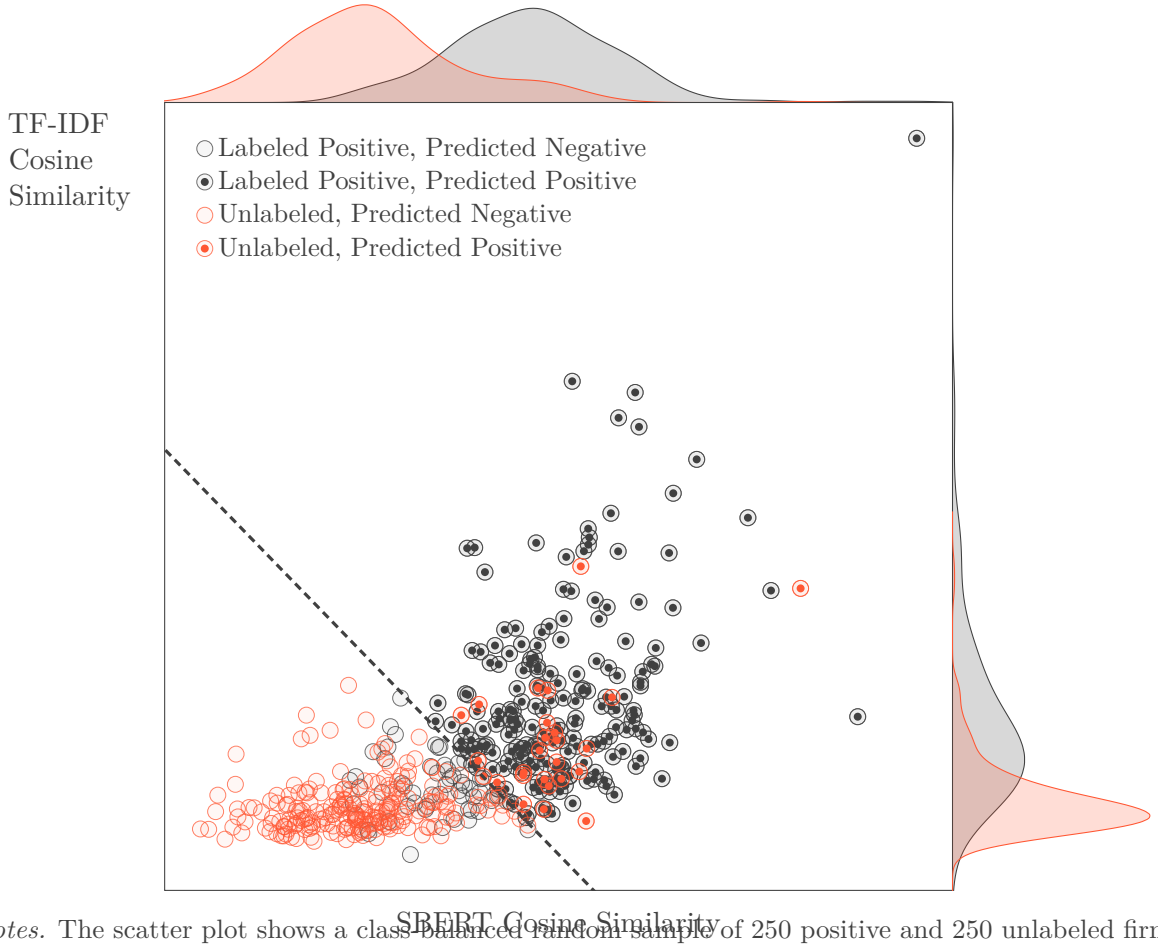
3.2 Positive and Unlabeled Learning

The machine learning literature has developed several approaches to positive and unlabeled learning, each suited to different assumptions about how positive instances are labeled. Bekker and Davis (2020) review these approaches and summarize the assumptions that underlie them. When all positive instances from a population are equally likely to be labeled positive, irrespective of the characteristics of each instance, the positive instances are said to be selected completely at random (SCAR). If the probability of a positive instance being labeled depends on observable characteristics, the instances are said to be selected at random (SAR). If the labeling mechanism depends on the probability of the instance truly being positive, even controlling for observable characteristics, the instances are said to be selected not at random (SNAR).

Many approaches to positive-unlabeled learning rely on the SCAR assumption, but our setting does not satisfy this assumption. Our positive labels arise from patent grants, and differences between patenting and non-patenting firms are well established. While these differences are less pronounced in our text-based features, they persist—suggesting SAR rather than SCAR conditions. Moreover, the labeling mechanism itself likely depends on the true positive status: firms that receive patent grants are more likely to find those patents genuinely useful, a characteristic of SNAR settings. Given these challenges, the SCAR assumption cannot be justified in our setting.

Instead, we employ a methodology that requires minimal assumptions about the labeling mechanism. Our approach belongs to a class of two-step methodologies that rely on assumptions of separability and smoothness (Bekker and Davis, 2020). Under separability, the classes are assumed to be separable such that a classifier exists that can map positive and negative instances to opposite sides of a decision threshold. Under the smoothness assumption, instances with similar features have similar label probabilities. Under these assumptions, a first-stage classifier can identify reliable negative instances based on their distance from labeled positive instances in feature space. A second-stage

Figure 4: Possible and Unlabeled Learning Classification Scatter Plot



Notes. The scatter plot shows a class-balanced random sample of 250 positive and 250 unlabeled firm-patent pairs, highlighting the separation between these groups. Firm-patent pairs where the firm owns the patent and our model predicts usefulness are marked as transparent black circles with solid black centers, while those predicted useless are marked as transparent black circles only. Pairs where the firm does not own the patent and our model predicts usefulness are marked as transparent red circles with solid red centers, while those predicted useless are marked as transparent red circles only. The horizontal axis represents SBERT cosine similarity scores, and the vertical axis represents TF-IDF cosine similarity scores. Kernel density plots above and to the right show the class-conditional distributions of each feature for the random sample. While the scatter plot demonstrates the predictive results of the simple two-feature Model 2a in Table 3 to illustrate class separation in two dimensions, our preferred Model 3c in Table 3, which incorporates additional features, achieves substantially better performance.

classifier can then be trained on the labeled positives and the reliable negatives that were identified in the first stage.

Figure 4 illustrates the class separation achievable using just two document similarity measures. The scatter plot shows that labeled positive and unlabeled firm-patent pairs exhibit substantial, though not perfect, separation in the feature space defined by SBERT and TF-IDF cosine similarities. The marginal distributions overlap but show distinct modes for each class–and importantly, since the unlabeled class contains both positive

and negative instances, the true separation between positive and negative classes would be even greater than what we observe here. This separation suffices for a two-stage methodology, as Bekker and Davis (2020) note that perfect separation is not required. Rather, there should exist regions where positive density significantly exceeds negative density, enabling the first stage to identify reliable negative examples with high confidence. The clear separation in the marginal distributions, particularly for SBERT, suggests our feature space meets this requirement. Additionally, the clustering of positive examples in the scatter plot, rather than random dispersion among unlabeled examples, supports the smoothness assumption that nearby points share similar class labels. While these two similarity measures alone achieve reasonable separation, our preferred model incorporates additional features and controls for improved performance.

Given this evidence for separability and smoothness, we implement a version of the two-stage spy methodology developed by Liu et al. (2002). In the first of the two stages, we create spies by removing the labels from a subset of positively labeled firm-patent pairs. We assign negative labels to the spies and to all unlabeled observations and train a first-stage classifier on this manipulated data. The first-stage classifier predicts probabilities that we use to identify reliable negatives. Specifically, we mark observations with first-stage probabilities below those of the spies as reliably negative.[10] This first stage filters out unlabeled firm-patent pairs that are likely positive. In the second stage, we train a classifier on the positive and reliably negative pairs.

3.3 Model Training

Our positive and unlabeled data is inherently class imbalanced: there are far fewer firm-patent pairs with positive labels than without labels. The imbalance arises because each patent has only one owner, and we rely on ownership for our positive labeling. This class imbalance can significantly affect our classifier's performance, producing a bias towards predicting the majority class, which in our case are the reliably-negative firm-patent

[10]The original spy methodology of Liu et al. (2002) uses the minimum positive label probability of the spies to establish the threshold. In our application, for robustness against possible noise in our positive labels, we take the tenth percentile rather than the minimum probability from the set of spies.

pairs. Because far fewer than one percent of our firm-patent pairs have positive labels, a classifier that never predicts positive labels would achieve near-perfect accuracy. But such a classifier would be worthless.

The machine learning literature has developed a range of sampling and ensemble methods for dealing with class imbalance (Galar et al., 2011). One such method, bootstrap resampling, is particularly well-suited to positive and unlabeled learning, especially in large datasets with few labeled observations (Mordelet and Vert, 2014). The method of bootstrap aggregation ("bagging"), as adapted to the positive-unlabeled setting by Mordelet and Vert, entails repeatedly under-sampling the unlabeled observations to produce a number of re-balanced training samples for an ensemble of classifiers. A simple average of the classifiers in the ensemble can then be used for prediction.

We adopt Mordelet and Vert's bootstrap aggregation procedure, using logistic regression as our core classification model. We prefer logistic regression over alternative models because it provides coefficients that indicate how each feature affects the model's decision, is efficient to train on large datasets, and is familiar to most economists. We form an ensemble of 10 logistic regression classifiers and class-balanced subsets of labeled and unlabeled observations in the bootstrapped training samples.

Table 3 reports log odds ratios and z statistics for three models, where each model is estimated without document controls, with document controls, and with document controls, SIC Division indicators, and CPC Section indicators. Document controls include the word count, corpus overlap, and lexical diversity of each document in a given firm-patent pair. Model 1, which uses only SBERT similarity scores, shows that semantic similarity is a strong predictor of patent usefulness, with odds ratios ranging from 18.36 to 53.83 depending on the specification. Model 2 adds TF-IDF similarity scores, which contribute additional predictive power while moderating the effect of SBERT similarity. In our preferred specification, Model 3c, which incorporates TF, TF-IDF, and SBERT similarity scores, both SBERT and TF-IDF maintain strong positive associations with patent usefulness (odds ratios of 20.17 and 6.56 respectively), while TF shows a weak negative association. The addition of document, industry, and patent category controls

Table 3: Logistic Regression on Positive and Unlabeled Data

Features	Average Odds Ratios								
	Model 1			Model 2			Model 3		
	(a)	(b)	(c)	(a)	(b)	(c)	(a)	(b)	(c)
SBERT	18.36	44.50	53.83	9.14	18.58	22.70	7.14	19.78	20.17
	(15.20)	(34.38)	(33.94)	(19.06)	(26.31)	(28.89)	(23.54)	(29.27)	(32.07)
TF-IDF				2.05	3.37	3.46	11.44	6.59	6.56
				(17.18)	(19.80)	(18.36)	(22.48)	(16.51)	(17.19)
TF							0.15	0.43	0.42
							(−22.45)	(−8.70)	(−9.05)
F_{1c}	0.48	0.61	0.61	0.51	0.65	0.65	0.56	0.65	0.67
Precision_c	0.54	0.59	0.60	0.55	0.62	0.62	0.57	0.62	0.63
Recall	0.73	0.80	0.80	0.74	0.80	0.81	0.76	0.80	0.81
Doc Ctrls	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Sic Ctrls	No	No	Yes	No	No	Yes	No	No	Yes
Cpc Ctrls	No	No	Yes	No	No	Yes	No	No	Yes
Penalty	L2	L2	L2	L2	L2	L2	L2	L2	L2
Obs/Reg	15489	15480	15470	15472	15485	15465	15486	15477	15515
Reg/Ens	10	10	10	10	10	10	10	10	10
RRS	2	2	2	2	2	2	2	2	2

Notes. The table shows alpha estimates for low and high decile technology momentum portfolios upper panel of the table shows odds ratio estimates from three logistic regression models trained on positive and unlabeled data, with z-statistics reported in parentheses under each estimate. Model 1 uses SBERT, Model 2 uses SBERT and TFIDF, and Model 3 uses SBERT, TFIDF, and TF similarity scores as features. We estimate each model without controls (columns a), with document controls (columns b), and with document, SIC, and CPC controls. All models are estimated with L2 ridge penalties applied.

improves model performance across all specifications, with model evaluations rising substantially from the baseline Model 1a to our preferred Model 3c.

3.4 Model Evaluation

Evaluating classifiers trained on positive and unlabeled data presents unique challenges that limit the informativeness of traditional performance metrics. In standard supervised learning, classifier performance is typically assessed using metrics derived from the confusion matrix—a matrix containing true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). Two fundamental metrics are precision and recall. For the positive class, precision is defined as $TP/(TP + FP)$, and measures the fraction of

positive predictions that are correct. Again for the positive class, recall is defined as $TP/(TP + FN)$, and measures the fraction of actual positive cases that are correctly identified. Both metrics also have a symmetric definition for the negative class. The F_1 score, given by $2 \times (\text{precision} \times \text{recall})/(\text{precision} + \text{recall})$, provides a single summary measure that balances these two objectives.

These traditional metrics must be interpreted with caution in the positive-unlabeled learning context. While recall can be reliably estimated using only positive examples, precision requires false positives—positive predictions on truly negative instances. In positive and unlabeled learning, we fundamentally cannot observe false positives since unlabeled examples may be either positive or negative. If traditional classifier performance metrics are applied to non-traditional classifiers, positively-predicted unlabeled instances are treated as false positives (the standard approach) and understate precision.[11]

This understatement is particularly severe in settings with significant class imbalance. Consider a classifier that correctly identifies a labeled positive instance but also predicts positive for three out of 1,000 unlabeled instances. Even if all three of these "false positives" are actually correct predictions of unlabeled positive cases, treating them as errors results in a precision of 25%. Thus, even strong classifier performance can appear poor when evaluated using traditional precision with positive and unlabeled data in a class-imbalanced setting.

To partially address the problem of class imbalance in our model evaluation, we follow Siblini et al. (2020) and "calibrate" our precision and F_1 scores to a reference class ratio of 1, making them more interpretable. This calibration does help with class imbalance, but does not correct for mistaken "false positives." We therefore rely primarily on a robust metric that, while less interpretable, is suitable for determining the optimal decision threshold for our classifier—that is, the threshold value above which predicted probabilities are classified as positive. Rather than choosing the threshold to maximize the calibrated F_{1c} score, which remains problematic in our setting, we maximize a modified

[11] Under the SCAR assumption, analytical corrections can be made (Elkan and Noto, 2008), but we do not make the SCAR assumption in our setting, as explained above.

Table 4: Positive and Unlabeled Classifier Performance

	Average Classifier Performance			
	Precision_c	Recall	F_{1c}	Support
Unlabeled Class	0.98	0.90	0.94	1,422,892
	(0.00)	(0.01)	(0.01)	(33)
Labeled Class	0.28	0.72	0.39	1,920
	(0.02)	(0.03)	(0.02)	(33)
Weighted Average	0.98	0.90	0.94	1,424,812
	(0.00)	(0.01)	(0.01)	(0)
Macro Average	0.63	0.81	0.67	1,424,812
	(0.01)	(0.01)	(0.01)	(0)

Notes. The tables shows traditional performance metrics for our classifier trained on positive and unlabeled data. In the upper panel of the table, we report average annual precision, recall, and F_1 performance metrics computed from individual classifiers trained on yearly patent and business description data. In the lower panel, we report weighted and unweighted (macro) averages across classes of the average annual class-specific performance metrics. To account for the substantial class imbalance in our positive and unlabeled data, we report calibrated precision and F_1 scores, indicated by the subscript $_c$ in the first and third columns.

version of the performance measure proposed by Lee and Liu (2003),

$$\text{Modified Lee-Liu Score:} \quad \lambda_\gamma = \frac{r^\gamma}{Pr(\hat{y} = 1)}\,, \tag{1}$$

where r is recall, $Pr(\hat{y} = 1)$ is the fraction of instances classified as positive, and the parameter γ can be adjusted to place greater emphasis on recall. The threshold λ_γ can be reliably estimated from positive and unlabeled data, making it suitable for positive-unlabeled learning. We set $\gamma = 3$ to reflect our preference for higher recall.[12]

We present traditional evaluation metrics (with the caveats noted above) in Table 4. These metrics are computed using repeated random sub-samples (RRS) of our training data. The RRS procedure involves repeatedly re-partitioning the data into training and test sets, training a new model on each training partition, and evaluating it on the corresponding test partition. This approach provides more robust performance estimates than a single train-test split. Within each year, we compute performance metrics and

[12]Lee and Liu (2003) show that their original metric with $\gamma = 2$ is proportional to the product of precision and recall ($r \times p$). Since the F_1 score is the harmonic mean of precision and recall, their metric captures similar information while being computable in the positive-unlabeled setting. Our choice of $\gamma = 3$ places more emphasis on recall, which is conceptually similar to using an F_β score with $\beta > 1$.

their standard errors across RRS iterations, then average both the metrics and standard errors across sub-samples and years. After the model evaluation stage, we use all available data to train a model for final predictions.

For the unlabeled class, our classifier achieves a calibrated precision of 0.98 and a recall of 0.90, yielding a calibrated F_{1c} score of 0.94. These results reflect the extreme class imbalance in our data—with such a low base rate of positive cases, even a naive classifier that predicts "unlabeled" for all observations would achieve near-perfect performance here.

The more challenging metrics are those for the labeled class, where our classifier achieves a calibrated precision of 0.28 and recall of 0.72. The low precision understates true performance—without true negative examples, many "false positives" are likely correct predictions of technological usefulness not captured in our labeled set. The higher recall indicates that our classifier identifies a substantial majority of labeled instances where a patent is owned by a particular firm and is therefore likely to be useful to the firm. Recall may also understate true performance, if our positive labels are noisy and contain instances of useless patents owned by firms.

The support column in Table 4 shows the scale of our evaluation, with metrics computed over more than 1.4 million firm-patent pairs per year. The standard errors, shown in parentheses, are computed across RRS iterations within each year and then averaged across years. These errors indicate that our performance estimates are stable across different random sub-samples. The substantial difference between weighted and macro averages (0.94 versus 0.67 for F_{1c}) shows how class imbalance affects the traditional performance metrics even after calibration.

The ultimate test of our estimates lies in their ability to predict or explain economic phenomena. Before turning to economic applications, however, we first examine the usefulness probabilities estimated by the model, and use the estimates to compare the technological associations of non-patenting and patenting firms. Despite the large number and economic importance of non-patenting firms, their technological profiles have remained largely unexplored by researchers due to data limitations. Our estimates remove these

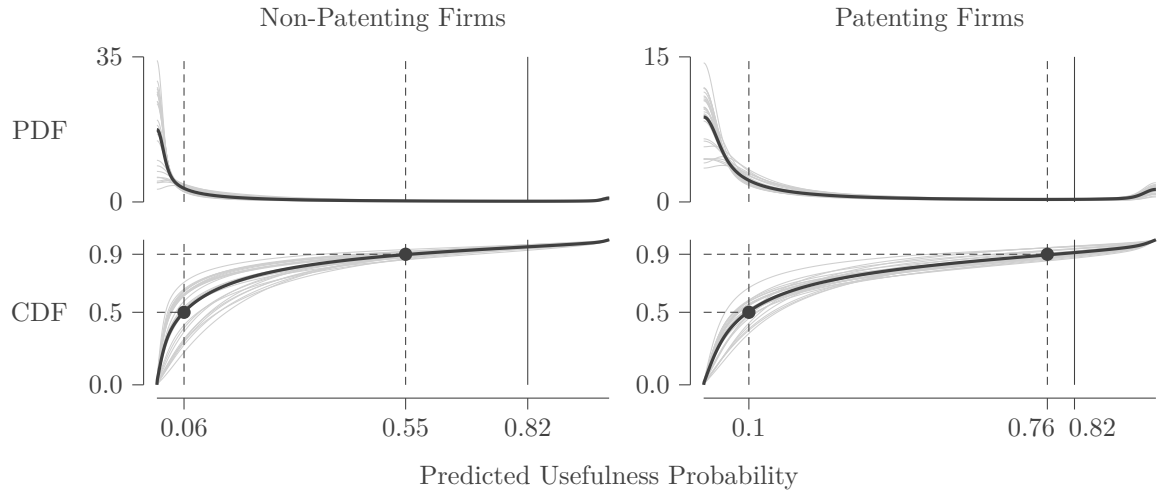
limitations, revealing surprising similarities between the technological associations of non-patenting and patenting firms.

3.5 Model Predictions

We use Model 3c from Table 3 to predict firm-patent associations, and plot the unconditional results of these predictions for all years in Figure 5. The figure displays the distribution of predicted usefulness probabilities for non-patenting firms (left) and patenting firms (right). The upper panels show probability density functions (PDFs) with annual distributions plotted as thin gray lines and their pointwise average as a thick black line. The lower panels show cumulative distribution functions (CDFs) with the same line conventions. Thin dashed lines mark the 50th and 90th percentiles, while the thin solid vertical line indicates the classification threshold that maximizes the modified Lee-Liu score λ_γ, which equals 0.82 when averaged across years.

Without conditioning on industry or firm size, we find that non-patenting firms have lower probabilities of associating with patents overall. The median predicted usefulness probability for non-patenting firms is 0.06, compared to 0.1 for patenting firms. Similarly, the 90th percentile values are 0.55 and 0.76 respectively, indicating that non-patenting firms have lower predicted associations than their patenting counterparts even at the upper end of the distribution. However, these raw differences primarily reflect the distinct industry and size compositions of the two firm types rather than fundamental differences in technological associations, as we emphasize in the following section.

Figure 5: Annual Distributions of Predicted Usefulness Probabilities



Notes. The figure shows the distribution of predicted usefulness probabilities for non-patenting firms (left) and patenting firms (right). Distributions are estimated from 1% samples of firm-patent usefulness probabilities drawn from each firm's set of usefulness probabilities each year, using kernel density functions with reflecting barriers at 0 and 1. In the upper panels, annual PDFs are plotted as thin gray lines, and the pointwise average of annual PDFs is plotted as a thick black line. In the lower panels, annual CDFs are plotted as thin gray lines, and the pointwise average of annual CDFs is plotted as a thick black line. The average of the annual 50th and 90th quantile values are indicated by thin dashed lines. The average of the annual classification threshold that maximizes the modified Lee-Liu score is indicated by a solid vertical line.

4 Characterizing Technological Associations

Having estimated usefulness probabilities for all firm-patent pairs, we now examine how technological associations vary between non-patenting and patenting firm types. We examine these associations at two levels. First, we examine aggregated patterns by studying the intensity of technological associations between industries and technology categories by firm type. Second, we examine firm-level patterns by studying the technological portfolios of individual firms. At both levels, we find surprising similarities between non-patenting and patenting firms. Cross-sectional distributions of several measures of technological association are similar across firm types, particularly in their central tendencies. Indeed, we find that within-type differences are much larger than between-type differences, after conditioning on industry and firm size.

4.1 Aggregated Technological Associations

We use Sankey diagrams to illustrate the intensity of technological associations between industries and technology categories. We define the aggregated intensity of technological association between SIC Division s and CPC Section c for firm type $\tau \in \{NP, P\}$ as:

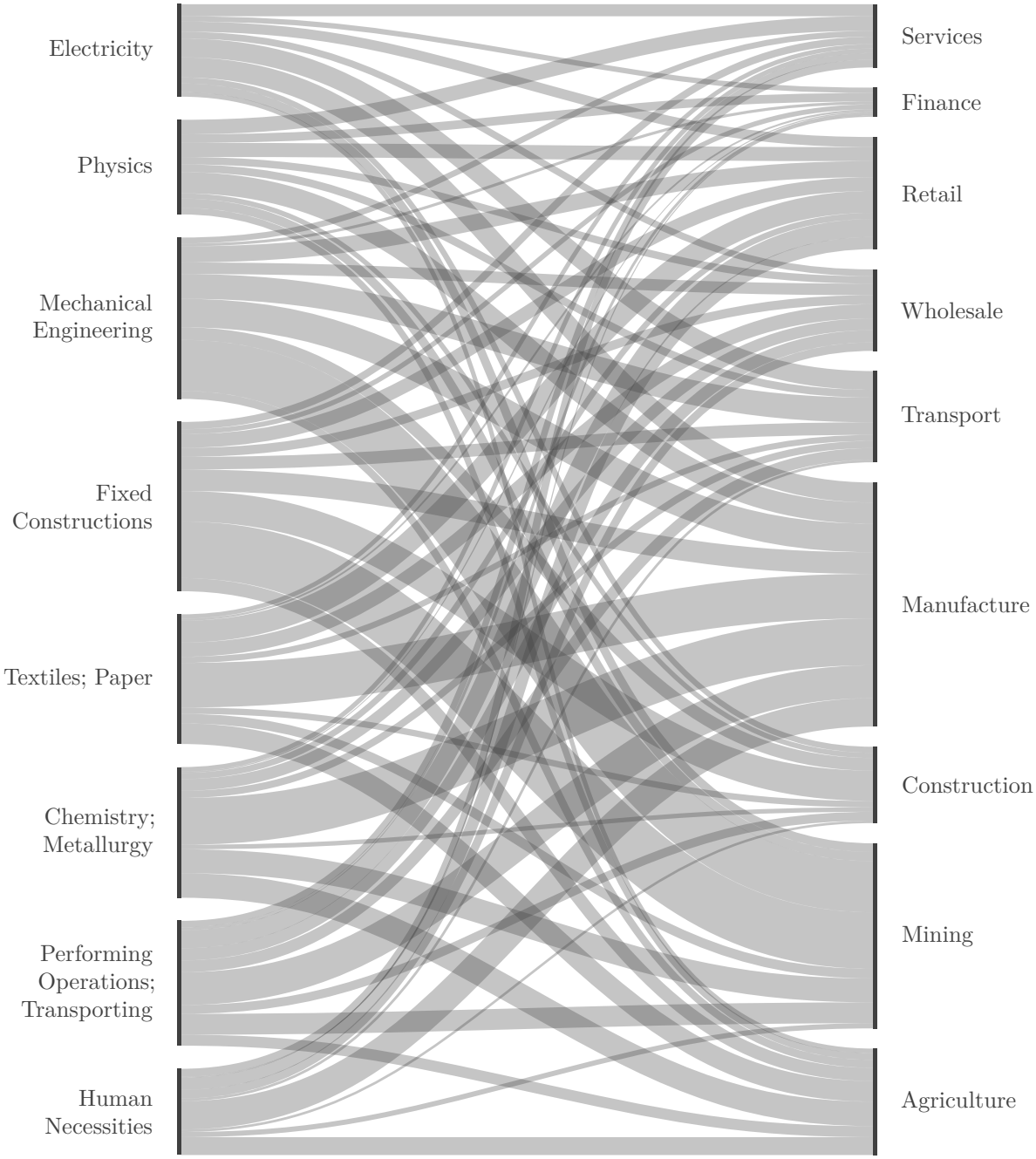
$$I_{sc}^{(\tau)} = \frac{A_{sc}^{(\tau)}}{F_s^{(\tau)} \times P_c} \tag{2}$$

where $A_{sc}^{(\tau)}$ is the number of firm-patent associations between Division s and Section c for firm type τ, $F_s^{(\tau)}$ is the number of firms of type τ in Division s, and P_c is the number of patents in Section c. This intensity measure normalizes raw association counts by the total number of possible firm-patent pairs in each industry-technology combination, providing a measure that is adjusted for compositional differences in firm and patent distributions across industries and technology categories. The Sankey diagrams in Figures 6 and 7 show the aggregated intensity-based associations for non-patenting and patenting firms, respectively. We provide Sankey diagrams based on raw association counts in Appendix B.1.

We draw two conclusions from the intensity-based Sankey diagrams. First, each SIC Division shows substantial associations with multiple CPC Sections, suggesting that technologies frequently cross industry boundaries. Second, the patterns for non-patenting firms in Figure 6 closely resemble those for patenting firms in Figure 7. While Sankey diagrams based on raw association counts in Appendix B.1 show greater differences, the intensity-based view reveals strong similarities after adjusting for compositional effects. That said, some differences between the intensity-based diagrams for non-patenting and patenting firms are apparent upon close inspection.

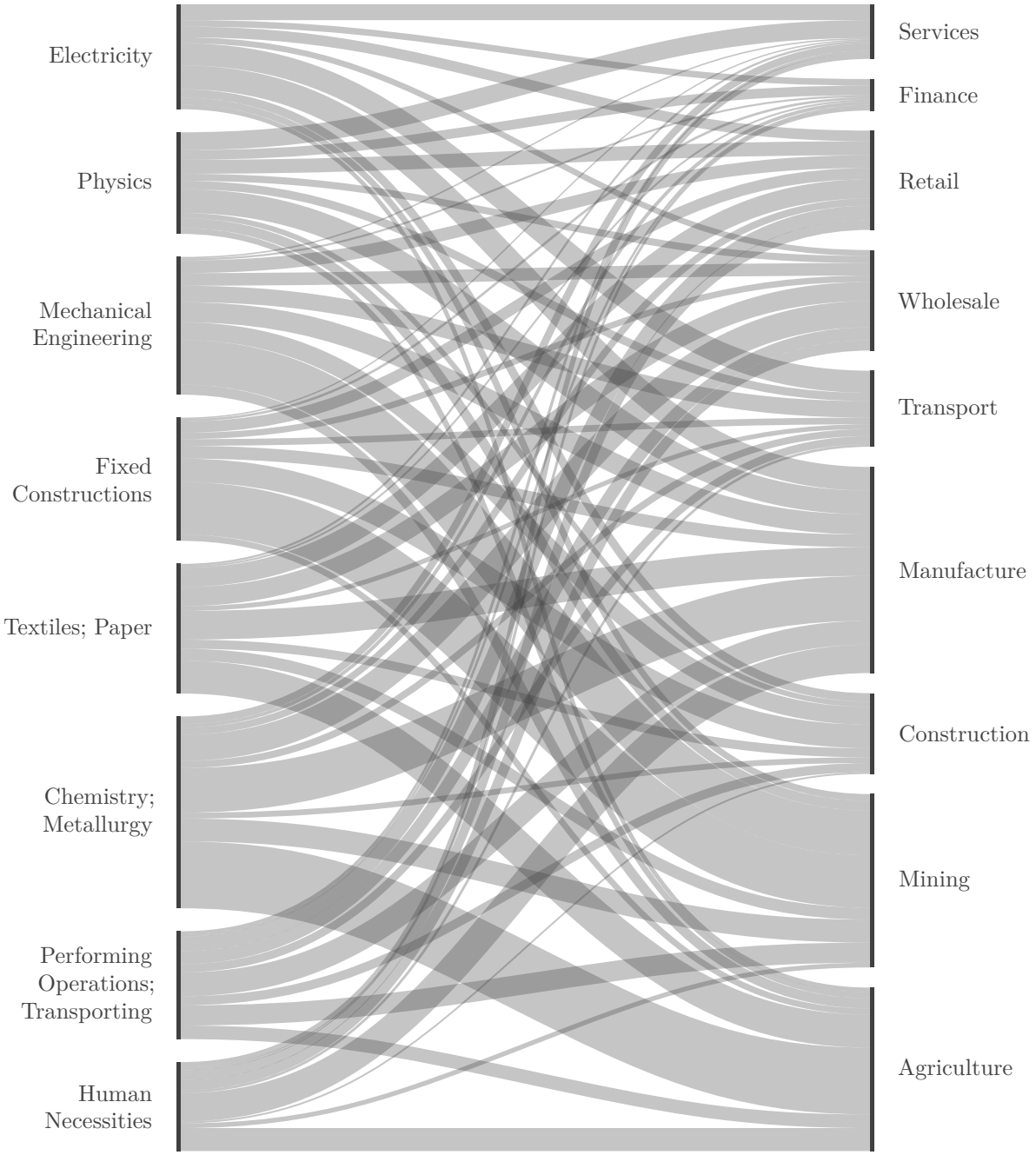
To facilitate the direct comparison of aggregated technological associations between non-patenting and patenting firms, we construct a heatmap representation in Figure 8 that quantifies differences between the Sankey diagrams in Figures 6 and 7 for each industry-technology combination. The heatmap displays a relative intensity metric R_{sc},

Figure 6: CPC-SIC Sankey Diagram: Intensity-Based, Non-Patenting Firms



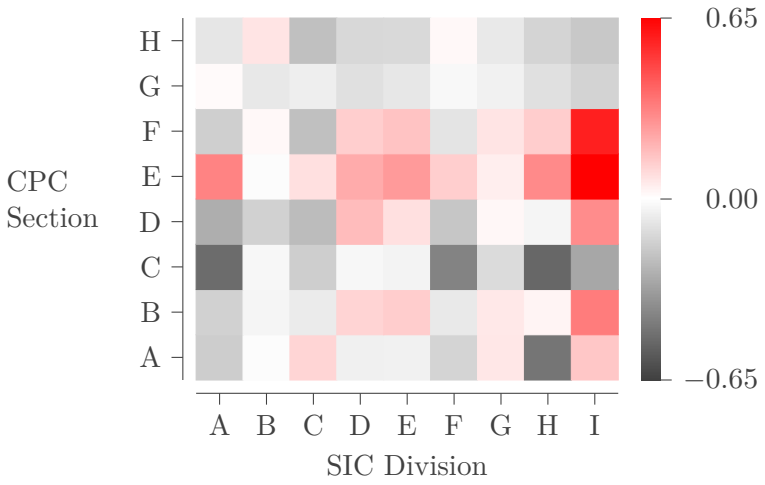
Notes. The figure shows a Sankey diagram of associations between CPC Sections (left) and SIC Divisions (right). Flows represent intensities of association, defined as the number of the number of associated firm-patent pairs for a given CPC-SIC combination, *relative* to the number of possible firm-patent pairs for that combination, aggregated over all sample years. CPC Sections: Human Necessities (A), Performing Operations; Transporting (B), Chemistry; Metallurgy (C), Textiles; Paper (D), Fixed Constructions (E), Mechanical Engineering; Lighting; Heating; Weapons; Blasting (F), Physics (G), and Electricity (H). SIC Divisions: Agriculture, Forestry, and Fishing (0100–0999), Mining (1000–1499), Construction (1500–1799), Manufacturing (2000–3999), Transportation, Communications, Electric, Gas and Sanitary Service (4000–4999), Wholesale Trade (5000–5199), Retail Trade (5200–5999), Finance and Insurance (6000–6799, excl 6500–6599 and 6700–6799), and Services (7000–8999). Some CPC and SIC names have been shortened for the figure.

Figure 7: CPC-SIC Sankey Diagram: Intensity-Based, Patenting Firms



Notes. The figure shows a Sankey diagram of associations between CPC Sections (left) and SIC Divisions (right). Flows represent intensities of association, defined as the number of the number of associated firm-patent pairs for a given CPC-SIC combination, *relative* to the number of possible firm-patent pairs for that combination, aggregated over all sample years. CPC Sections: Human Necessities (A), Performing Operations; Transporting (B), Chemistry; Metallurgy (C), Textiles; Paper (D), Fixed Constructions (E), Mechanical Engineering; Lighting; Heating; Weapons; Blasting (F), Physics (G), and Electricity (H). SIC Divisions: Agriculture, Forestry, and Fishing (0100–0999), Mining (1000–1499), Construction (1500–1799), Manufacturing (2000–3999), Transportation, Communications, Electric, Gas and Sanitary Service (4000–4999), Wholesale Trade (5000–5199), Retail Trade (5200–5999), Finance and Insurance (6000–6799, excl 6500–6599 and 6700–6799), and Services (7000–8999). Some CPC and SIC names have been shortened for the figure.

Figure 8: Relative Technological Associations: Non-Patenting vs. Patenting Firms



Notes. The figure shows the relative intensity metric defined in equation (3) for technological associations between CPC Sections and SIC Divisions. Red cells indicate stronger associations for non-patenting firms, gray cells indicate stronger associations for patenting firms, and white cells indicate equal associations. CPC Sections: Human Necessities (A), Performing Operations; Transporting (B), Chemistry; Metallurgy (C), Textiles; Paper (D), Fixed Constructions (E), Mechanical Engineering; Lighting; Heating; Weapons; Blasting (F), Physics (G), and Electricity (H). SIC Divisions: A (Agriculture, Forestry, and Fishing: 01000999), B (Mining: 10001499), C (Construction: 15001799), D (Manufacturing: 20003999), E (Transportation, Communications, Electric, Gas and Sanitary Services: 40004999), F (Wholesale Trade: 50005199), G (Retail Trade: 52005999), H (Finance, Insurance, and Real Estate: 6000–6799, excl 6500–6599 and 6700–6799), I (Services: 70008999).

which we define as

$$R_{sc} = \frac{I_{sc}^{(N)} - I_{sc}^{(P)}}{I_{sc}^{(N)} + I_{sc}^{(P)}}, \tag{3}$$

where superscripts N and P denote non-patenting and patenting firms, respectively. This bounded measure ranges from -1 to $+1$, with positive values (red shading) indicating stronger technological associations for non-patenting firms and negative values (grey shading) indicating stronger associations for patenting firms.

The heatmap reveals several systematic differences between the intensity of technological associations for non-patenting and patenting firms. We caution, however, that some of the strongest differences arise in industry-technology combinations with overall low intensities. For example, the services industry shows strongly positive relative intensity values in fixed constructions and mechanical engineering, indicating that non-patenting service firms associate more intensively with these technologies than patenting service firms. However, the service industry has low-intensity associations with these technologies

overall. Conversely, chemical and metallurgical technologies show consistently negative values across most industries, indicating that patenting firms associate more strongly with these technologies than non-patenting firms; this is especially true in agriculture, which shows a high-intensity association with chemistry.

Overall, a nuanced picture emerges from the aggregated results, where technologies frequently cross industry boundaries, and industry-technology associations on an intensity basis are surprisingly similar for non-patenting and patenting firms, with some important differences between firm types in isolated cases. Next, we turn to firm-level evidence.

4.2 Firm-Level Technological Associations

We consider three metrics that characterize firm-level technological associations. To assess technological breadth and depth, we count each firm's associated patents and CPC categories. To assess technological instability over time, we compute add and drop rates for each firm's associated CPC categories. To assess technological generality, we compute the average number of industries associated with each firm's associated patents and CPC categories. We compute metrics at the patent level and at the CPC Group, Subclass, Class levels. This approach allows us to assess firms associations with increasingly broad technology categories. For each metric at each level, we find that differences within non-patenting and patenting firm types exceed differences between firm types.

Categorical Associations. Our positive and unlabeled learning framework produces usefulness probabilities that associate individual firms with individual patents. We use a binomial testing framework to convert these firm-patent associations into probabilistic firm-category, patent-industry, and category-industry associations, which we require in order to compute technological breadth and depth, instability, and generality metrics at each level of the CPC system.

To associate firms with CPC categories, we compare the count of a firm's patent associations within the category, relative to the count of all patents in the category, with the count of all firms' patent associations within the category, relative to the count of all

firms multiplied by the count of all patents in the category. The binomial probability of A_{ic} associations between firm i and patents in CPC category c is given by

$$\pi_{ic} = \binom{P_c}{A_{ic}} \pi_{0c}^{A_{ic}} (1 - \pi_{0c})^{P_c - A_{ic}}, \quad \text{with} \quad \pi_{0c} \equiv \sum_i \frac{A_{ic}}{F \times P_c}, \tag{4}$$

where F is the number of firms, P_c is the number of patents in category c, and π_{0c} is the baseline probability of a firm-patent association in category c. Under the null hypothesis, $\pi_{ic} = \pi_{0c}$ and firm i is no more likely to associate with patents in category c than the average firm. We test the alternative hypothesis that $\pi_{ic} > \pi_{0c}$, and associate firms with categories when we fail to reject the alternative hypothesis at the 5% significance level.

This approach provides a statistical framework for identifying when a firm's relationship with a technology category is unlikely to have arisen by chance. It adjusts for CPC category size, requiring firms to have more patent associations in larger categories. And, because it uses category-specific baseline probabilities, it accounts for systematic differences in the number of associations across categories while maintaining a consistent threshold for statistical significance.

We adopt a similar procedure to associate patents with four-digit SIC industries. The binomial probability of A_{js} associations between patent j and firms in SIC industry s is given by

$$\pi_{js} = \binom{F_s}{A_{js}} \pi_{0s}^{A_{js}} (1 - \pi_{0s})^{F_s - A_{js}}, \quad \text{with} \quad \pi_{0s} \equiv \sum_j \frac{A_{js}}{F_s \times P}, \tag{5}$$

where F_s is the number of firms in industry s, P is the number of patents, and π_{0s} is the baseline probability of a firm-patent association in industry s. Under the null hypothesis, $\pi_{js} = \pi_{0s}$ and patent j is no more likely to associate with firms in SIC industry s than the average patent. We test the alternative hypothesis that $\pi_{js} > \pi_{0s}$, and associate patents with industries when we fail to reject the alternative hypothesis at the 5% significance level. We use the same procedure to associate CPC categories with four-digit SIC industries, replacing patent j with category c and number of patents P with number of categories C in the binomial probability calculation.

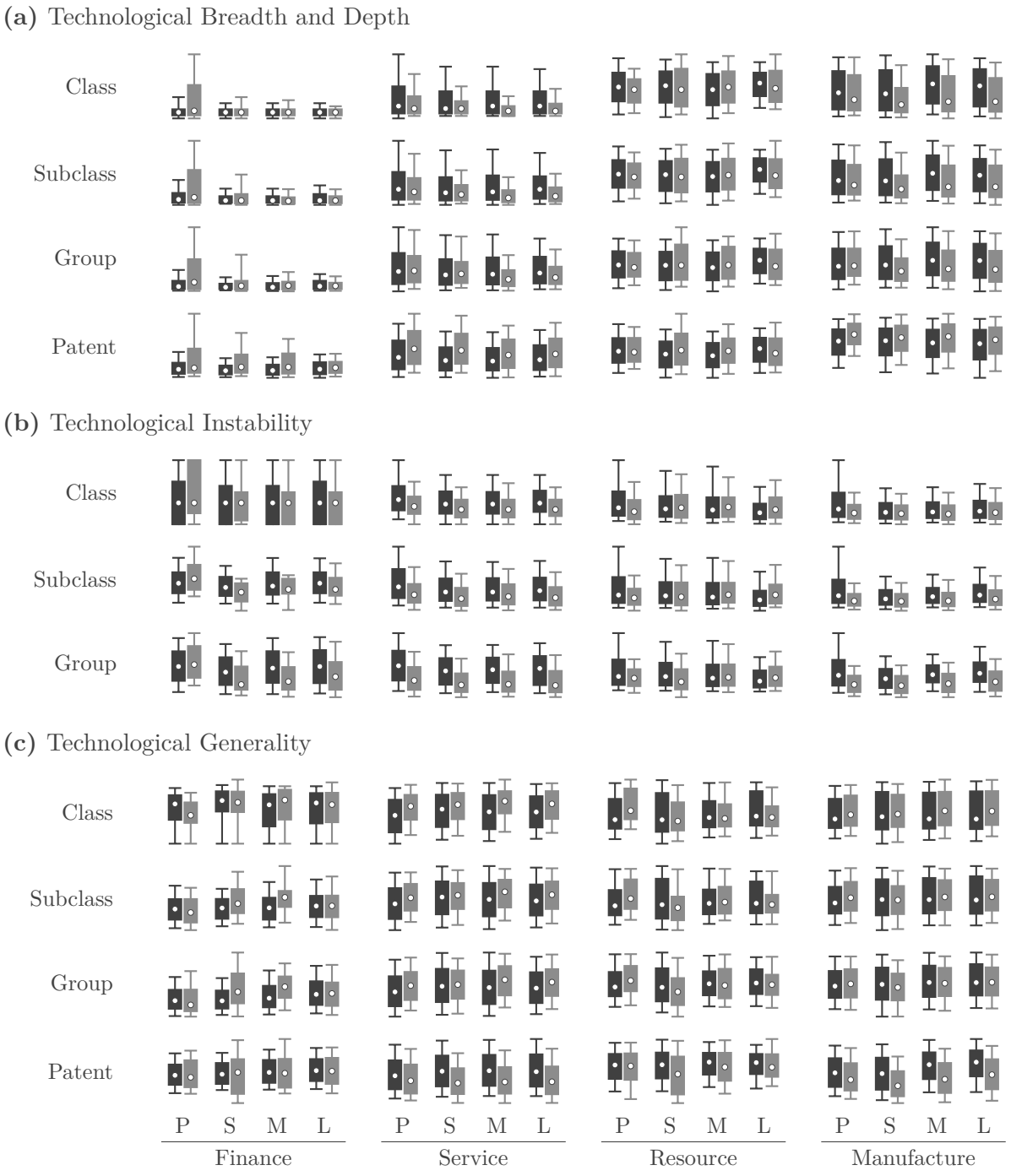
Non-Patenting and Patenting Firm Comparisons. With the firm-level category and industry associations in place, we can compute metrics for technological breadth and depth, instability, and generality at the patent level and at the level of CPC Classes, Subclasses, and Groups. We present the results in Figure 9, which plots the cross-sectional distribution of each metric, separately for non-patenting and patenting firms, conditional on broadly-defined industry group and size class. The industry groups (finance, service, resource, and manufacture) and size classes (private, small, medium, and large) are those used throughout the paper, and defined in Section 2.

Within industry group and size class, distributions for non-patenting and patenting firms are compared on a common scale, but the scales do differ across industry groups and size classes. We deliberately omit numerical values from the figure, to avoid distracting from the main insight: conditional on industry group and size class, the differences within non-patenting and patenting firm types exceed differences between firm types for each metric at each CPC level. We report extensive tabular results with statistical tests for differences in means and medians between firm types in Appendix B.

Panel 9a plots cross-sectional distributions of technological breadth and depth, measured at the firm level as the count of patents and CPC categories with which firms associate. We tabulate these results in Appendix B.2. While we find large and statistically significant differences in counts between industry groups, we find smaller differences between size classes within industry group, and yet smaller differences between non-patenting and patenting firms within industry group and size class. While differences between median non-patenting and patenting firms remain statistically significant after conditioning on industry and size, these differences are much smaller than the interquartile range of firm-level values found within the non-patenting and patenting firm types.

We interpret counts at the patent level as measures of technological depth, and do find consistently higher median counts for patenting firms at this level, in particular outside of finance. As we consider increasingly broad CPC categories, we interpret the counts as measures of technological breadth, and find that median counts for non-patenting firms frequently exceed median counts for patenting firms. Overall, the results suggest deeper

Figure 9: Firm-Level Technological Associations Within Industry Group and Size Class



Notes. The figure shows cross-sectional distributions of three metrics characterizing firms technological associations within industry group and size class. Panel 9a shows technological breadth and depth measured by counts of patent and CPC category associations. Panel 9b shows technological instability measured by CPC category churn rates (the average of add and drop rates). Panel 9c shows technological generality measured by cross-industry usage (the average number of industries associated with each firm's associated patents or CPC categories). Box plots show medians, interquartile ranges, and the 10th and 90th percentiles, separately for non-patenting firms (dark) and patenting firms (light). Distributions for non-patenting and patenting firms are compared on a common scale within industry group and size class, but scales do differ across industry groups and size classes. Industry groups and size classes are defined in Section 2. We report extensive tabular results with statistical tests for differences in means and medians between firm types in Appendix B.

technological associations for patenting firms, and broader technological associations for non-patenting firms outside of finance.

Panel 9b plots cross-sectional distributions of technological instability, measured at the firm level as the rate of churn in firms' associations with CPC categories over time. We tabulate these results in Appendix B.3. For each firm-year, we first calculate two measures, the add rate and the drop rate of CPC categories, defined as the percentage of CPC categories added to or dropped from a firm's associations in year $t+1$ relative to the average number of the firm's associated categories across years t and $t+1$. We then define the churn rate as the average of add and drop rates. As with the breadth and depth metrics, we find that differences in the median measure of technological instability between non-patenting and patenting firm types, conditional on industry group and size class, are substantially smaller than interquartile ranges of firm-level values found within either firm type.

That said, we do find statistically significant differences in median instability measures between non-patenting and patenting firm types, conditional on industry group and size class, with non-patenting firms generally showing higher degrees of technological instability. Differences are generally more pronounced at the level of narrowly-defined CPC Groups, and less pronounced for broadly-defined CPC Classes. We note, however, that exceptions can be found; see, for example, private financial firms or large resource firms. Overall, our results suggest that non-patenting firms have more flexible technological portfolios, potentially adapting more quickly to changing technological opportunities. This interpretation would align with our finding that non-patenting firms maintain broader but shallower technological portfolios.

Panel 9c plots cross-sectional distributions of technological generality, measured at the firm level as the average number of four-digit SIC industries associated with each firm's associated patents or CPC categories. We tabulate these results in Appendix B.4. For example, a firm with associated patents A and B would have a generality value of two at the patent level if patent A were associated with one industry and patent B were associated with three industries ($2 = (1+3)/2$). The generality metric quantifies the extent to which

firms' associated technologies are in general use across multiple industries.[13] Once again, differences in technological generality are larger within firm type than between firm types, conditional on industry group and firm size, though statistically significant differences across types do exist.

For CPC Classes and Subclasses, which are broadly-defined technology categories, we find that median patenting firms tend to associate with more general technologies than median non-patenting firms, particularly in services. At the CPC Group level, we find no clear pattern across industry groups and size classes. At the patent level, the pattern reserves, and we find that median non-patenting firms associate with more general technologies than median patenting firms. This result suggests that patenting firms associate with more specific patents, but technology categories with broader cross-industry appeal.

Taken together, these patterns in firm-level technological associations suggest that patenting and non-patenting firms differ in systematic ways: patenting firms exhibit deeper, more stable, and more specific technological focus than non-patenting firms, with some exceptions in particular industry groups and size classes. However, while these differences are statistically significant, they are small relative to the within-type variation across all three metrics, after conditioning on industry group and size class. Ultimately, we find modest differences and surprising similarities in the technological associations of non-patenting and patenting firms.

5 Technological Momentum

We now explore a first application of our patent usefulness probabilities to asset pricing, constructing technological momentum portfolios that include, for the first time, a large set of publicly-traded non-patenting firms whose technological profiles have previously been inaccessible to researchers. Our approach extends recent work by Lee, Sun, Wang, and

[13]This measure of technological generality relies on the assumption that four-digit SIC industries are defined with equal granularity in all parts of the economy, which may not hold in practice. However, by comparing non-patenting firms with patenting firms within broad industry groups and size classes, we mitigate problems arising from violations of this assumption.

Zhang (2019) and Bekkerman, Fich, and Khimich (2023) documenting the profitability of technological momentum strategies applied to patenting firms. These authors argue that technological momentum works because markets are slow to process technological information, especially for technologically intensive firms with limited investor attention. Non-patenting firms use technology as intensively as patenting firms but are typically smaller, with less analyst coverage and more opaque technological profiles. We therefore expect the technological momentum strategy to be particularly effective when extended to these firms.

Our methodology differs from prior approaches in two key ways. First, we expand coverage to include a majority of publicly-traded firms that do not patent and were previously excluded. Second, we measure technological similarity based on firms' exposure to useful technology rather than ownership of intellectual property. This distinction matters because firms often patent for strategic reasons unrelated to their core technological activities (Cohen, Nelson, and Walsh, 2000). Overall, these differences allow us to capture broader technological relationships for a wider set of firms.

5.1 Methodology

Technological momentum strategies exploit predictable patterns in how technology-related information affects stock prices. The strategy identifies firms with similar technological profiles, then takes long positions in firms whose technological peers recently performed well and short positions in firms whose peers performed poorly. Prior research demonstrates that technological peer performance predicts future returns, generating significant alpha.

In implementations of this strategy by Lee et al. (2019) and Bekkerman, Fich, and Khimich (2023), technological similarity is measured using the patent portfolios of patenting firms. Lee et al. (2019) computes the distribution of patents owned by patenting firms across USPTO classes, and correlates these distributions for pairs of patenting firms to assess firm-to-firm technological similarity, following Jaffe (1986). Bekkerman, Fich, and Khimich (2023) measure patent-to-patent textual similarity as the cosine similarity between TF-IDF vectors for each patent, and then average these over the patents owned

by patenting firms to assess firm-to-firm technological similarity. Both methods consider only patenting firms, excluding the majority of publicly-traded firms from the analysis.

To extend the technological momentum strategy to this excluded majority, we use our estimated usefulness probabilities to identify technological peers. These probabilities are available for all U.S. public firms and an annual sample of 50,000 patents, covering the period from 1997 to 2023. As noted in Section 2.4, only 15.64% of CRSP firms receive patent grants in an average year, accounting for 46.29% of CRSP market capitalization (or 25.07% of CRSP firms accounting for 54.08% of CRSP market capitalization, when counting patents in five-year rolling windows), so the expansion to non-patenting firms significantly improves coverage.

Because we estimate similarity scores between each firm and each patent each year, we can assess firm-to-firm technological similarity by computing the cosine similarity between vectors of patent-level usefulness probabilities directly. With 50,000 entries, each vector constitutes a highly granular technological profile of each firm. The usefulness probabilities in most of these entries lie below the classification threshold that maximizes the modified Lee-Liu score from the positive and unlabeled learning step.[14] We set usefulness probabilities below the threshold to zero before the similarity calculation.

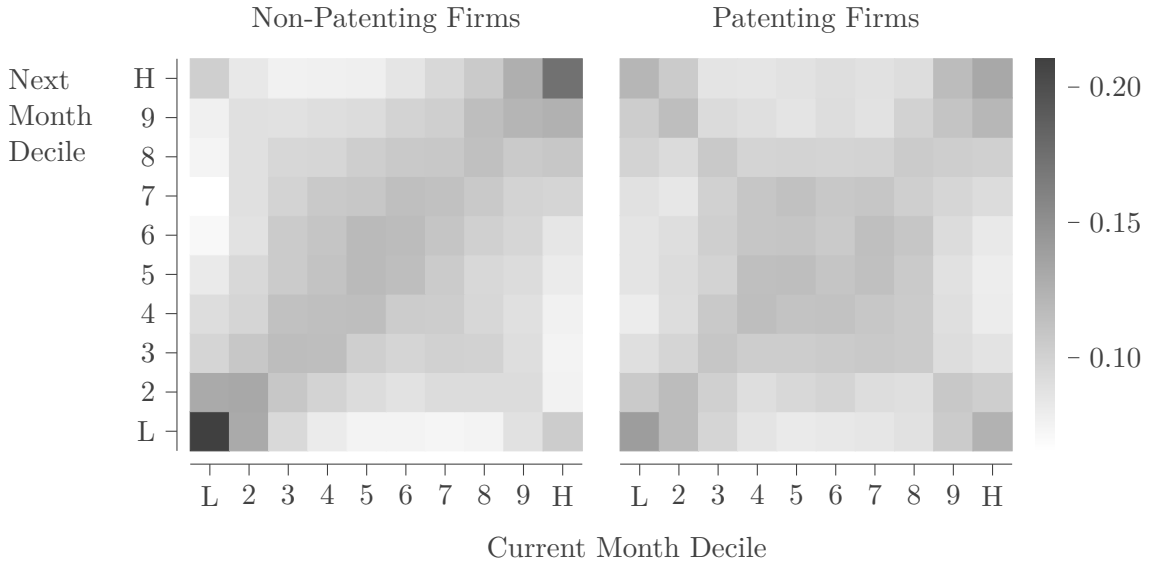
For focal firm i and month t, we calculate the weighted average return on a portfolio of the focal firm's top 100 technologically similar peers,

$$\text{TECHRET}_{it} = \frac{\sum_{j \neq i} w_{ijt} \times \text{RET}_{jt}}{\sum_{j \neq i} w_{ijt}} \tag{6}$$

where RET_{jt} is the return of firm j in month t and w_{ijt} is the lagged technological cosine similarity between firms i and j. We lag technological cosine similarity weights to avoid look-ahead bias. Specifically, the weight w_{ijt} used in month t is the firm-to-firm technological similarity computed from patent grants and annual reports in year $y_t - 1$, where y_t denotes the year in which month t occurs. Since annual reports are filed once per year and our similarity measures are calculated annually, all months within a given calendar year use the same set of lagged similarity weights from the previous year's annual

[14] The threshold varies from year to year, averaging 0.82 across all years, 1997 to 2023.

Figure 10: Technology Momentum Decile Transition Matrices



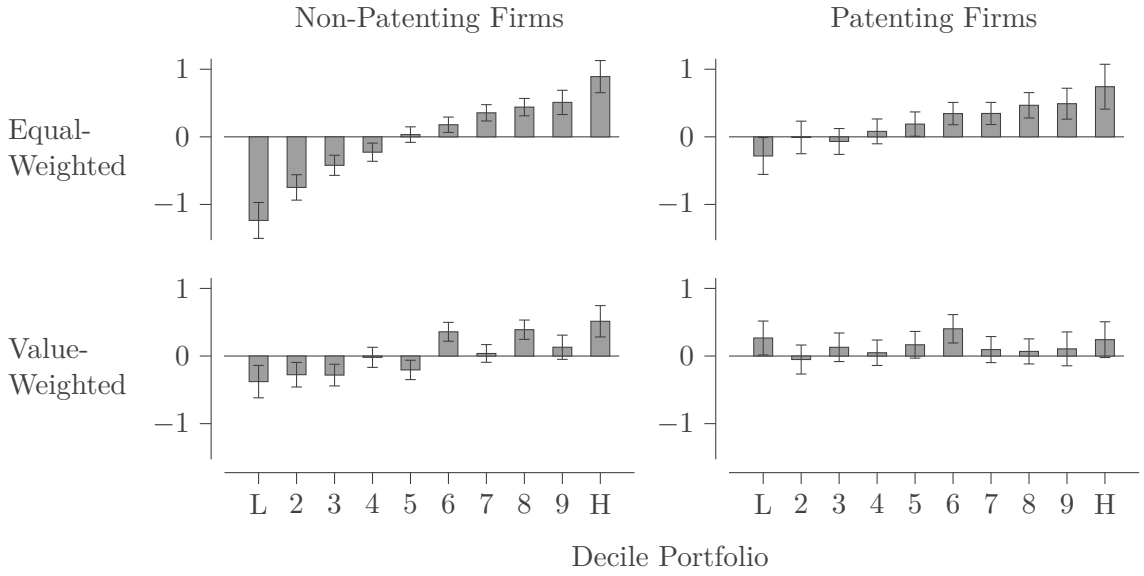
Notes. The figure shows monthly transition probabilities for firms sorted into deciles based on the prior month's stock market performance of technologically-similar peers, for non-patenting firms (left) and patenting firms (right). The transition probabilities are estimated using monthly returns data over the period 1998 to 2023. The portfolios are formed on the universe of CRSP firms with available annual reports, after a sample selection rule is applied. Technological similarity is measured using the cosine similarity between vectors of each firm's patent usefulness probabilities. Each firm's peer group is defined as the 100 most technologically-similar firms. The transition probabilities reflect the likelihood of a firm moving between decile portfolios from one month to the next, indicating the persistence of performance for decile portfolios formed on technological momentum.

reports and patent grants. For example, portfolio returns calculated in any month of 2020 use similarity weights computed from 2019 annual reports and patent grants.

We then sort firms into deciles based on lagged TECHRET values from the previous month, and form portfolios that go long the top decile (firms whose technological peers performed well in the previous month) and short the bottom decile (firms whose technological peers performed poorly). These portfolios are rebalanced monthly to maintain either equal or value weights. To be included in a decile portfolio, firms must have filed an annual report in the previous year and, to ensure tradability, must be above the 10th percentile of CRSP firms' one-month lagged prices, market values, and trading volumes.

Figure 10 shows the transition matrices for technological momentum deciles for non-patenting firms (left) and patenting firms (right). Each cell (m, n) represents the probability of a firm moving from decile m in the current month to decile n in the next month. The diagonal elements indicate performance persistence—firms remaining in the

Figure 11: Monthly Alpha of Technology Momentum Decile Portfolios



Notes. The figure shows monthly alpha estimates in percentage points for equal-weighted and value-weighted decile portfolios formed on the recent performance of technological peers, where L represents the lowest decile and H the highest. We estimate alpha using a four-factor model that includes market, size, value, and cross-sectional momentum factors. Lines extending from each bar show 95% confidence intervals. The portfolios are formed on the universe of CRSP firms, after a sample selection rule is applied. Portfolios are rebalanced monthly over the period January 1998 to December 2023 based on technology momentum decile rankings.

same decile from one month to the next. The transition matrices show that non-patenting firms are substantially more likely than patenting firms to remain in the highest and lowest deciles from one month to the next. This result suggests that technological information diffuses more slowly for non-patenting firms, creating more persistent return predictability. As we demonstrate next, this improved predictability translates into improved performance for the technological momentum strategy.

5.2 Empirical Results

Figure 11 presents the monthly alphas for technological momentum portfolio excess returns across deciles, reported separately for non-patenting and patenting firms and for equal and value-weighted portfolios. The alphas are estimated from a four-factor model that controls for exposure to market, size, value, and cross-sectional momentum factors.[15] For

[15]These four factors, commonly denoted MKT, SMB, HML, and MOM, were downloaded from the data library on Ken French's website (link) on 10 March 2024.

equal-weighted portfolios, the results reveal a striking monotonic pattern across deciles, with higher technological momentum deciles consistently earning higher risk-adjusted returns. This pattern is particularly pronounced for non-patenting firms, where the spread between high and low deciles is substantially larger than for patenting firms.

Table 5 reports the factor loadings for the technological momentum strategy. The high-minus-low portfolio loads negatively on the market factor and positively on the size factor for both non-patenting and patenting firms. Importantly, these factor exposures do not explain away the substantial alphas generated by the strategy, suggesting that the strategy represents a distinct anomaly not captured by standard risk factors, consistent with the findings of both Lee et al. (2019) and Bekkerman, Fich, and Khimich (2023). Table B43 in Appendix B.5 demonstrates that these results are robust to alternative asset pricing models, with monthly alphas for non-patenting firms ranging from 1.79% to 2.48% depending on the specification. The technological momentum results remains economically and statistically significant across all model specifications.

Table 5: Technology Momentum Four-Factor Model Loadings for Non-Patenting and Patenting Firms

		Non-Patenting Firms					Patenting Firms				
	Decile	Alpha	MKT	SMB	HML	MOM	Alpha	MKT	SMB	HML	MOM
Equal-Weighted	High	0.89	1.03	0.97	−0.04	−0.09	0.74	0.91	1.39	−0.55	−0.11
		(3.75)	(18.43)	(12.62)	(−0.59)	(−1.85)	(2.24)	(11.67)	(12.95)	(−5.68)	(−1.52)
	Low	−1.24	1.16	0.83	−0.10	−0.27	−0.28	1.23	0.70	−0.19	−0.22
		(−4.66)	(18.65)	(9.61)	(−1.26)	(−4.86)	(−1.04)	(19.27)	(7.96)	(−2.33)	(−3.80)
	High–Low	1.97	−0.13	0.15	0.06	0.18	0.87	−0.32	0.69	−0.37	0.11
		(4.74)	(−1.36)	(1.09)	(0.45)	(2.02)	(1.69)	(−2.66)	(4.14)	(−2.44)	(1.01)
Value-Weighted	High	0.51	0.97	0.07	0.08	0.08	0.24	0.90	0.17	−0.30	0.01
		(2.22)	(17.85)	(0.89)	(1.17)	(1.58)	(0.92)	(14.44)	(1.95)	(−3.89)	(0.17)
	Low	−0.38	1.12	0.22	−0.14	−0.14	0.27	1.13	−0.03	−0.17	−0.13
		(−1.58)	(19.95)	(2.79)	(−2.02)	(−2.78)	(1.06)	(19.24)	(−0.31)	(−2.34)	(−2.45)
	High–Low	0.74	−0.15	−0.15	0.22	0.21	−0.18	−0.23	0.19	−0.13	0.14
		(1.88)	(−1.63)	(−1.17)	(1.90)	(2.60)	(−0.42)	(−2.30)	(1.39)	(−1.04)	(1.51)

Notes. The table shows factor loadings from a four-factor model estimated using monthly excess returns from 1998 to 2023 for equal-weighted and value-weighted technology momentum portfolios, for non-patenting and patenting firms. Each sub-table shows factor loadings for high-decile and low-decile portfolios, as well as for a high-minus-low portfolio. T-statistics are reported in parentheses under each alpha estimate.

Our equal-weighted long-short portfolio generates a monthly alpha of 1.97% (t-statistic = 4.74) after controlling for the market, size, value, and momentum factors. This performance exceeds the 1.17% monthly alpha reported by Lee et al. (2019), though their alpha is estimated over a longer period 1963–2012. Bekkerman, Fich, and Khimich (2023) report a 1.29% monthly alpha for their text-based approach over the period 1977-2016. However, as Bekkerman, Fich, and Khimich (2023) document, the performance of technological momentum strategies has fallen over time, not risen, so the historical monthly alpha likely overstates performance in recent years. In that sense, our shorter and more recent sample period (1998–2023) works against finding strong results, reducing both statistical power and alpha. Despite this challenging sample period, the alpha we find for non-patenting firms remains statistically significant and exceeds that of previous studies, showing the effectiveness of our approach in identifying technological relationships that extend beyond traditional patent-based measures.

Figure 12 illustrates the cumulative performance of the high-decile technological momentum portfolio from 1998 to 2023, compared against standard technology benchmarks, on both natural and logarithmic scales. The shaded areas indicate NBER-dated recessions. For non-patenting firms, both equal and value-weighted high-decile portfolios substantially outperform the benchmarks over this period, with the equal-weighted portfolio performing particularly well. In contrast, a value-weighted technological momentum strategy for patenting firms performs no better than the technology benchmarks.

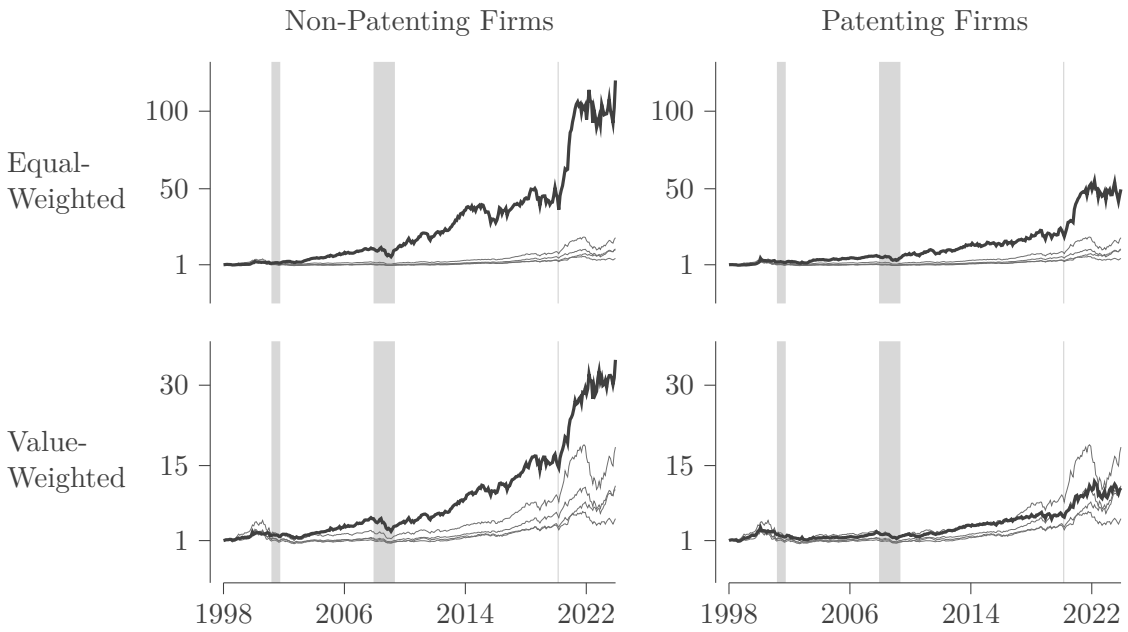
5.3 Economic Mechanism

What explains the superior performance of our strategy compared to previous approaches? We propose that the key factor is our ability to identify technological relationships for firms that do not directly own patents. These firms are substantially exposed to technology shocks through their use of technology in production, their dependence on complementary innovations, or their positions in broader technology networks, but their exposures have remained opaque to market participants.

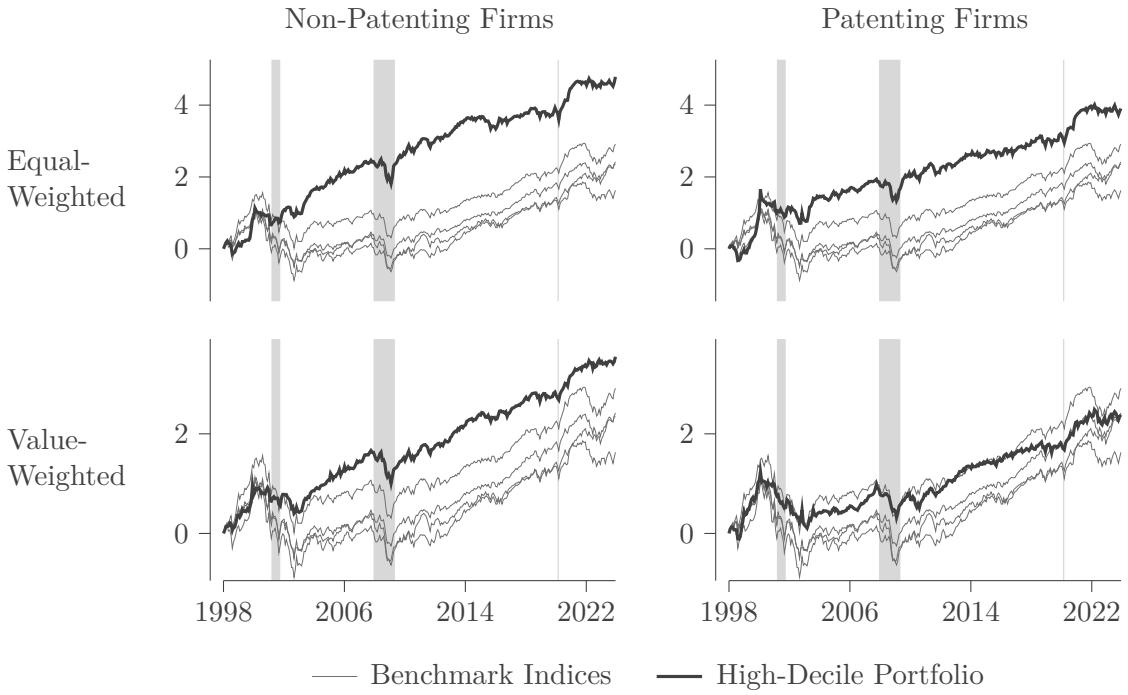
Consistent with Lee et al. (2019) and Bekkerman, Fich, and Khimich (2023), investors

Figure 12: Cumulative Growth of High-Decile Technology Momentum Portfolio

(a) Natural Scale



(b) Log Scale



Notes. The figure shows the cumulative growth of a $1 investment in equal-weighted and value-weighted high-decile technology momentum portfolios of stocks for non-patenting and patenting firms from 1998 to 2023. For comparison, the figure also shows the cumulative growth of four benchmark indices: NYSE Technology, S&P North American Technology Sector, FTSE All-World Technology, and Russell 2000 Technology. Panel 12a plots this growth on the natural scale, while Panel 12b plots it on the logarithmic scale. Technology momentum decile portfolios are formed on the universe of CRSP firms, after dropping the bottom 10% of firms by price, market capitalization, and trading volume. Portfolios are rebalanced monthly based on the recent performance of each firm's technological peers. Shading indicates U.S. recession dates.

appear to systematically underreact to technological information. This underreaction is more pronounced for non-patenting firms with less transparent technological profiles. The stronger persistence in decile membership for non-patenting firms, shown in Figure 10, supports this interpretation. Our results extend previous research by demonstrating that technological momentum effects are not limited to patenting firms but affect the broader universe of publicly-traded companies.

6 Conclusion

We develop a novel measure of firm-level technological usefulness by applying positive and unlabeled machine learning to patent and business descriptions. Our methodology addresses a fundamental limitation in the economics literature: existing datasets focus on the small minority of firms that own patents, while the technological profiles of most firms remain unobserved. By measuring the usefulness of patents to all firms rather than just their owners, we create a technology dataset that covers all U.S. public firms and all economically important technology categories.

Our analysis reveals three key findings. First, technological associations between non-patenting and patenting firms are remarkably similar after controlling for industry and size. Non-patenting firms maintain broader but shallower technological portfolios than patenting firms, with higher rates of technological instability, and greater technological generality. Second, we document substantial return predictability from technological momentum strategies, with monthly alphas of 1.97% for non-patenting firmssignificantly exceeding the performance of similar strategies applied only to patenting firms. Third, event study evidence shows that firms experience positive abnormal returns following announcements of patents we identify as useful to them, even when they do not own the patents. These results suggest that technological spillovers extend far beyond patent ownership.

Our approach has several limitations that future research should address. Our method assumes that firms' technological activities are reflected in their public disclosures, though

firms may use technologies they do not mention or strategically withhold information. Additionally, while our usefulness probabilities strongly predict patent ownership patterns, they remain estimates that may not fully capture the complexity of how firms use technology. Furthermore, our textual analysis depends on the quality and specificity of language in patent and business descriptions, though modern language models help mitigate concerns about ambiguity or technical jargon. Finally, our focus on U.S. firms and patents may limit generalizability to other institutional or international contexts.

Despite these limitations, our work makes important contributions. Methodologically, we introduce positive and unlabeled learning to economics, demonstrating how machine learning can address missing data problems that have constrained empirical research. Substantively, we provide the first comprehensive view of technology usage across all public firms, revealing that non-patenting firmswhich account for over half of market capitalizationhave rich technological profiles that matter for asset prices and firm performance. By moving beyond patent ownership to measure technological usefulness, we open new avenues for research on innovation, productivity, and technological change across the entire economy.

References

Aizawa, A. (2003). "An information-theoretic perspective of tf–idf measures." *Information Processing & Management* 39.1, pp. 45–65.

Akcigit, U. and W. R. Kerr (2018). "Growth through heterogeneous innovations." *Journal of Political Economy* 126.4, pp. 1374–1443.

Argente, D. et al. (2023). "Patents to products: Product innovation and firm dynamics."

Arora, A., S. Belenzon, and L. Sheer (2021). "Matching patents to Compustat firms, 1980–2015: Dynamic reassignment, name changes, and ownership structures." *Research Policy* 50.5, p. 104217.

Ash, E. and S. Hansen (2023). "Text algorithms in economics." *Annual Review of Economics* 15.

Autor, D. H., F. Levy, and R. J. Murnane (2003). "The skill content of recent technological change: An empirical exploration." *The Quarterly journal of economics* 118.4, pp. 1279–1333.

Bekker, J. and J. Davis (2020). "Learning from positive and unlabeled data: A survey." *Machine Learning* 109, pp. 719–760.

Bekkerman, R., E. M. Fich, and N. V. Khimich (2023). "The effect of innovation similarity on asset prices: Evidence from patents' big data." *The Review of Asset Pricing Studies* 13.1, pp. 99–145.

Bloom, N., M. Schankerman, and J. Van Reenen (2013). "Identifying technology spillovers and product market rivalry." *Econometrica* 81.4, pp. 1347–1393.
Bloom, N. et al. (2021). *The diffusion of disruptive technologies*. Tech. rep. National Bureau of Economic Research.
Bound, J. et al. (1982). "Who does R&D and who patents?"
Brown, S. V. and J. W. Tucker (2011). "Large-sample evidence on firms' year-over-year MD&A modifications." *Journal of Accounting Research* 49.2, pp. 309–346.
Cohen, L., C. Malloy, and Q. Nguyen (2020). "Lazy prices." *The Journal of Finance* 75.3, pp. 1371–1415.
Cohen, W. M., R. Nelson, and J. P. Walsh (2000). *Protecting their intellectual assets: Appropriability conditions and why US manufacturing firms patent (or not)*.
Dunne, T. (1994). "Plant Age and Technology use in U.S. Manufacturing Industries." *The RAND Journal of Economics* 25.3, p. 488.
Elkan, C. and K. Noto (2008). "Learning classifiers from only positive and unlabeled data." *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 213–220.
European Patent Office (EPO) and United States Patent and Trademark Office (USPTO) (2017). *Guide to the CPC (Cooperative Patent Classification)*. Version 1.0.
Fama, E. (2023). "Production of US Rm-Rf, SMB, and HML in the Fama-French Data Library." *Chicago Both Paper* 22-23.
Galar, M. et al. (2011). "A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches." *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42.4, pp. 463–484.
Gentzkow, M., B. Kelly, and M. Taddy (2019). "Text as Data." *Journal of Economic Literature* 57.3, pp. 535–574.
Graham, S. J., A. C. Marco, and R. Miller (2015). "The USPTO patent examination research dataset: A window on the process of patent examination." *Georgia Tech Scheller College of Business Research Paper No. WP* 43.
Graham, S. J., A. C. Marco, and A. F. Myers (2018). "Patent transactions in the marketplace: Lessons from the USPTO patent assignment dataset." *Journal of Economics & Management Strategy* 27.3, pp. 343–371.
Griliches, Z. (1987). *R&D, Patents and Productivity*. Conference report. University of Chicago Press.
— (1990). "Patent Statistics as Economic Indicators: A Survey." *Journal of Economic Literature* 28.4.
Hall, B. H., A. B. Jaffe, and M. Trajtenberg (2001). *The NBER patent citation data file: Lessons, insights and methodological tools*.
Hall, B. H., A. Jaffe, and M. Trajtenberg (2005). "Market Value and Patent Citations." *The RAND Journal of Economics* 36.1, pp. 16–38.
Hoberg, G. and G. Phillips (2016). "Text-based network industries and endogenous product differentiation." *Journal of Political Economy* 124.5, pp. 1423–1465.
— (2018). "Text-based industry momentum." *Journal of Financial and Quantitative Analysis* 53.6, pp. 2355–2388.
Jaffe, A. (1986). "Technological Opportunity and Spillovers of R&D: Evidence from Firms' Patents, Profits, and Market Value." *American Economic Review* 76.5, pp. 631–664.
Kakhbod, A. et al. (2024). "Measuring Creative Destruction." *Available at SSRN 5008685*.
Kelly, B. T. and D. Xiu (2023). "Financial machine learning." *Available at SSRN*.

Kogan, L. et al. (2017). "Technological innovation, resource allocation, and growth." *The Quarterly Journal of Economics* 132.2, pp. 665–712.

Kogan, L. et al. (2021). *Technology-skill complementarity and labor displacement: Evidence from linking two centuries of patents with occupations*. Tech. rep. National Bureau of Economic Research.

Kydland, F. E. and E. C. Prescott (1982). "Time to Build and Aggregate Fluctuations." *Econometrica* 50.6, p. 1345.

Lafond, F. and D. Kim (2019). "Long-run dynamics of the US patent classification system." *Journal of Evolutionary Economics* 29.2, pp. 631–664.

Lee, C. M. et al. (2019). "Technological links and predictable returns." *Journal of Financial Economics* 132.3, pp. 76–96.

Lee, W. S. and B. Liu (2003). "Learning with positive and unlabeled examples using weighted logistic regression." *ICML*. Vol. 3. 2003, pp. 448–455.

Lemley, M. A. and C. Shapiro (2005). "Probabilistic patents." *Journal of Economic Perspectives* 19.2, pp. 75–98.

Lerner, J. and A. Seru (2021). "The Use and Misuse of Patent Data: Issues for Finance and Beyond." *The Review of Financial Studies* 35.6, pp. 2667–2704.

Lerner, J. et al. (2021). *Financial innovation in the 21st century: Evidence from us patents*. Tech. rep. National Bureau of Economic Research.

Lerner, J. (1994). "The importance of patent scope: an empirical analysis." *The RAND Journal of Economics*, pp. 319–333.

Liu, B. et al. (2002). "Partially supervised classification of text documents." *ICML*. Vol. 2. 485. Sydney, NSW, pp. 387–394.

Lobo, J. and D. Strumsky (2019). "Sources of inventive novelty: two patent classification schemas, same story." *Scientometrics* 120.1, pp. 19–37.

Lopez-Lira, A. (2019). "Risk Factors That Matter: Textual Analysis of Risk Disclosures for the Cross-Section of Returns." *mimeo*.

Loughran, T. and B. McDonald (2020a). "Measuring firm complexity." *Journal of Financial and Quantitative Analysis*, pp. 1–55.

— (2020b). "Textual analysis in finance." *Annual Review of Financial Economics* 12, pp. 357–375.

McCarthy, P. M. and S. Jarvis (2010). "MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment." *Behavior research methods* 42.2, pp. 381–392.

Mezzanotti, F. and T. Simcoe (2023). *Innovation and appropriability: Revisiting the role of intellectual property*. Tech. rep. National Bureau of Economic Research.

Moore, K. A. (2005). "Worthless patents." *Berkeley Technology Law Journal* 20.4, pp. 1521–1552.

Mordelet, F. and J.-P. Vert (2014). "A bagging SVM to learn from positive and unlabeled examples." *Pattern Recognition Letters* 37, pp. 201–209.

Myers, K. R. and L. Lanahan (2022). "Estimating spillovers from publicly funded R&D: Evidence from the US Department of Energy." *American Economic Review* 112.7, pp. 2393–2423.

Reimers, N. and I. Gurevych (2019). "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks." *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Siblini, W. et al. (2020). "Master your metrics with calibration." *International Symposium on Intelligent Data Analysis*. Springer, pp. 457–469.

Simmons, H. J. (2014). "Categorizing the useful arts: Part, present, and future development of patent classification in the united states." *Law Libr. J.* 106, p. 563.

Solow, R. M. (1960). "Investment and Technical Progress." *Mathematical methods in the social sciences* 1, pp. 48–93.

Strumsky, D., J. Lobo, and S. Van der Leeuw (2012). "Using patent technology codes to study technological change." *Economics of Innovation and New technology* 21.3, pp. 267–286.

U.S. Congress (1836). *Act of July 4, 1836, ch. 357, 5 Stat. 117*. https://www.loc.gov/item/uscode18360704/.

— (1934). *Securities Exchange Act of 1934, 15 U.S.C. §§ 78l, 78m*. https://www.govinfo.gov/content/pkg/USCODE-2011-title15/pdf/USCODE-2011-title15-chap2B.pdf.

— (1952). *35 U.S.C. §§ 101-103 (Patent Act of 1952)*. https://www.govinfo.gov/content/pkg/USCODE-2012-title35/pdf/USCODE-2012-title35-partII-chap10-sec101.pdf.

— (2011). *35 U.S.C. § 8 (Leahy-Smith America Invents Act of 2011)*. https://www.law.cornell.edu/uscode/text/35/8.

U.S. Securities and Exchange Commission (2013). *17 CFR Part 229 (Regulation S-K)*. https://www.govinfo.gov/content/pkg/CFR-2013-title17-vol2/pdf/CFR-2013-title17-vol2-sec229-101.pdf.

United Nations (1971). *Profiles of Manufacturing Establishments*. Vol. I, II, and III. Industrial Planning and Programming Series 4. New York: United Nations.

United States Patent and Trademark Office (USPTO) (2013). *MPEP § 2107: Guidelines for Examination of Applications for Compliance with the Utility Requirement*.

World Intellectual Property Organization (2024). *Guide to the International Patent Classification (2024 Edition)*. World Intellectual Property Organization. Geneva, Switzerland.

ONLINE APPENDIX

A Methodological Appendix

A.1 Document Vectorization

In this section, we describe our methodology for measuring the textual similarity between business descriptions and patent descriptions. Specifically, we describe our procedures for vectorizing textual descriptions using traditional and modern techniques and our procedure for computing cosine similarity scores between the vector representations.

Word Frequency Vectorization. We begin by representing textual descriptions as numeric vectors using traditional term frequency (TF) and term frequency-inverse document frequency (TF-IDF) representations of text. These representations characterize texts by the frequency of words that appear within and across texts within collections of documents, called corpora. For our analysis, we study two corpora—namely, the collection of business descriptions that we extract from SEC annual report filings and the collection of patent descriptions provided by the USPTO.

For the traditional representations, we begin by tokenizing, lemmatizing, and removing standard English-language stop words from the descriptions. Tokenization is the process of splitting the text into individual words or tokens. Lemmatization involves reducing words to their base or dictionary form. Stop words, such as 'the', 'is', and 'and', are also removed. We use WordNet's lemmatizer and a standard English stopword list in this step.

Following the initial text preparation, we create a joint vocabulary from the processed business and patent descriptions, including only words that appear more than a minimum frequency and that appear in both corpora. In our baseline model, we set the minimum frequency to one, meaning that all words appearing only once in either corpus are dropped from the vocabulary. This approach reduces noise caused by infrequent terms and ensures the vocabulary used to create the TF and TF-IDF representations consists of words that are relevant to both businesses and patents.

We then transform the text into TF and TF-IDF forms suitable for mathematical analysis. TF measures how frequently a term occurs in a document. TF-IDF, on the other hand, additionally downscales words that appear frequently across many documents, capturing both the term's frequency within a document and its rarity across documents.

The TF and TF-IDF techniques were created in the mid-20th century, with TF predating TF-IDF. TF, which counts a term's occurrence in a document, stems from the early years of information retrieval. The approach assumes that a term's importance corresponds directly to its frequency. However, TF's drawback lies in its failure to consider a term's relevance in the broader document collection, thus often overemphasizing common, less informative words. To address this, IDF was introduced, measuring the weight of terms appearing frequently across all documents. By the 1970s, the two were combined into TF-IDF, which multiplies term frequencies in a document by inverse document frequencies, as the name suggest. Aizawa (2003) provides a brief history of these developments and a theoretical discussion of TF-IDF.

While TF and TF-IDF have proven effective for various information retrieval and text mining tasks, and still find use in modern text analysis applications in economics and finance, they primarily operate on word frequency without capturing semantic and contextual information, and are therefore outperformed in some tasks by more advanced models capable of better semantic understanding.

For the TF and TF-IDF representations, the vectorized text of each document is stored in a sparse matrix, one for each corpus, where rows correspond to documents and columns correspond to vocabulary words. Each element of the matrix holds the TF or TF-IDF value of the corresponding word for the corresponding document.

We repeat this process for each year in our sample, producing vectorized descriptions stored in annual sparse matrices for each corpus. This process is computationally intensive and the runtime depends on the size of the data and the computational resources available. However, using cloud computing, and splitting the computation across multiple cloud computing machines, the vectorization can be completed in reasonable time. The results provide the groundwork for one part of our subsequent analysis of the relationship between

business descriptions and patent ownership.

SBERT Vctorization. In addition to the traditional TF and TF-IDF vectorization techniques, which emphasize word frequencies, we also transform business and patent descriptions into mathematical representations known as embeddings. Embeddings are high-dimensional numerical representations of text that capture various aspects of text meanings and usages. The basic idea behind embeddings is to map discrete, categorical language data into continuous, high-dimensional vector spaces.

Our study uses an approach based on sentence embeddings, specifically the Sentence-BERT (SBERT) model. See Reimers and Gurevych (2019) for a detailed description of the model. SBERT is a modified version of the well-known BERT model that is designed to produce sentence-level embeddings, simplifying the task of identifying semantic similarities between different text documents. SBERT is an example of a transformer model. In transformer models, embeddings typically form the input layer of the model and are typically initialized with pre-trained word embeddings such as Word2Vec or GloVe. These initial embeddings are then further refined and updated during training through the transformer's attention mechanism. This allows the model to learn context-dependent representations, wherein the same word can have different embeddings based on its context, reflecting different potential meanings.

We use SBERT, specifically the MiniLM model (all-MiniLM-L6-v2), to capture semantic meaning in business and patent descriptions. This model can understand sentence-level semantics even when documents use different terminology to describe similar concepts—an important capability given the diverse language used across patent technical descriptions and business disclosures. MiniLM provides an efficient implementation of SBERT that balances computational speed with semantic accuracy.

Transformer models, including SBERT, have dramatically improved the quality of natural language processing, thanks to their ability to capture the complex context and semantics of text. However, they do come with one notable limitation: a maximum token limit, typically set at 512 tokens. This limit is primarily a result of the quadratic time complexity of the self-attention mechanism of transformer models. The self-attention

mechanism computes a score for each pair of tokens, leading to a time complexity of $O(n^2)$ for n tokens. In practice, this means that processing longer sequences requires significantly more computational resources, both in terms of time and memory. Given these constraints, these models often set a practical limit of 512 tokens to keep computational requirements manageable.

In our study, we confront the token limitation in SBERT while working with extended business descriptions that typically contain thousands or tens of thousands of words. To manage the token limit while preserving document-level semantic information, we employ a chunking and sampling strategy. We divide each document into overlapping chunks of text, with chunk size dynamically calculated based on the model's token limits and the average token-to-character ratio in our corpus. We set an overlap of 25 characters between consecutive chunks to maintain continuity.

From these chunks, we filter out those that are overly numerical (more than 20% digits) or too short (fewer than 25 characters). We then sample 25% of the chunks for processing: we always include the first chunk to capture introductory content, and randomly sample additional chunks to reach our 25% target. This sampling approach allows us to process documents that exceed token limits while capturing information from throughout the text.

For each selected chunk, we compute embeddings using SBERT and then aggregate these chunk-level embeddings to create document-level representations. We employ two pooling strategies: average pooling across all chunk embeddings and max pooling followed by averaging. Both resulting embeddings are normalized to unit length. This approach ensures consistent representation regardless of document length while preserving semantic information from multiple sections of lengthy documents.

We compute embeddings for all business descriptions and patent descriptions in each year, storing the embeddings in annual matrices with documents in rows and embedding dimensions in columns. While computationally intensive, this process can be efficiently parallelized across CPU cores or GPUs, allowing us to process large corpora in reasonable time using cloud computing resources.

A.2 Cosine Similarity

Having vectorized our document corpora using TF, TF-IDF, and SBERT embeddings, we next compute similarity scores between business and patent descriptions. The cosine similarity between business description b and patent description p, denoted $s(\boldsymbol{v}_b, \boldsymbol{v}_p)$, is a standard measure of similarity commonly used in text analysis and defined as the normalized dot product of two vector representations of documents,

$$s(\boldsymbol{v}_b, \boldsymbol{v}_p) = \frac{\boldsymbol{v}_b \cdot \boldsymbol{v}_p}{\|\boldsymbol{v}_b\| \|\boldsymbol{v}_p\|} \tag{A1}$$

where $\boldsymbol{v}_b$ and $\boldsymbol{v}_p$ denote document vectors for documents b and p and where $\|\cdot\|$ denotes the Euclidean norm. The cosine similarity measures the angle between two vectors and takes higher values when the angle between vectors is smaller. Because cosine similarity is based on vector angles rather than vector lengths, cosine similarity is, in theory, robust to length differences across documents.

We compute cosine similarity scores between every possible pair of business description and patent description in our sample each year, using each of our vectorization methods (TF, TF-IDF, and SBERT embeddings). These cosine similarity scores provide a meaningful and quantifiable measure of the textual relatedness of business and patent descriptions. These scores form the basis for our subsequent investigation into the predictive power of document similarity with respect to patent ownership, and form the basis for our characterization of firm-level technology profiles.

The number of possible pairs between business descriptions and patent descriptions grows based on the product of their counts. If both the number of business descriptions and patent descriptions increase, the total pairs grow geometrically. Given the large number of firms and larger number of patents in our sample each year, the possible pairs number in the billions. The cosine similarity computations are fast and perform well at this scale.

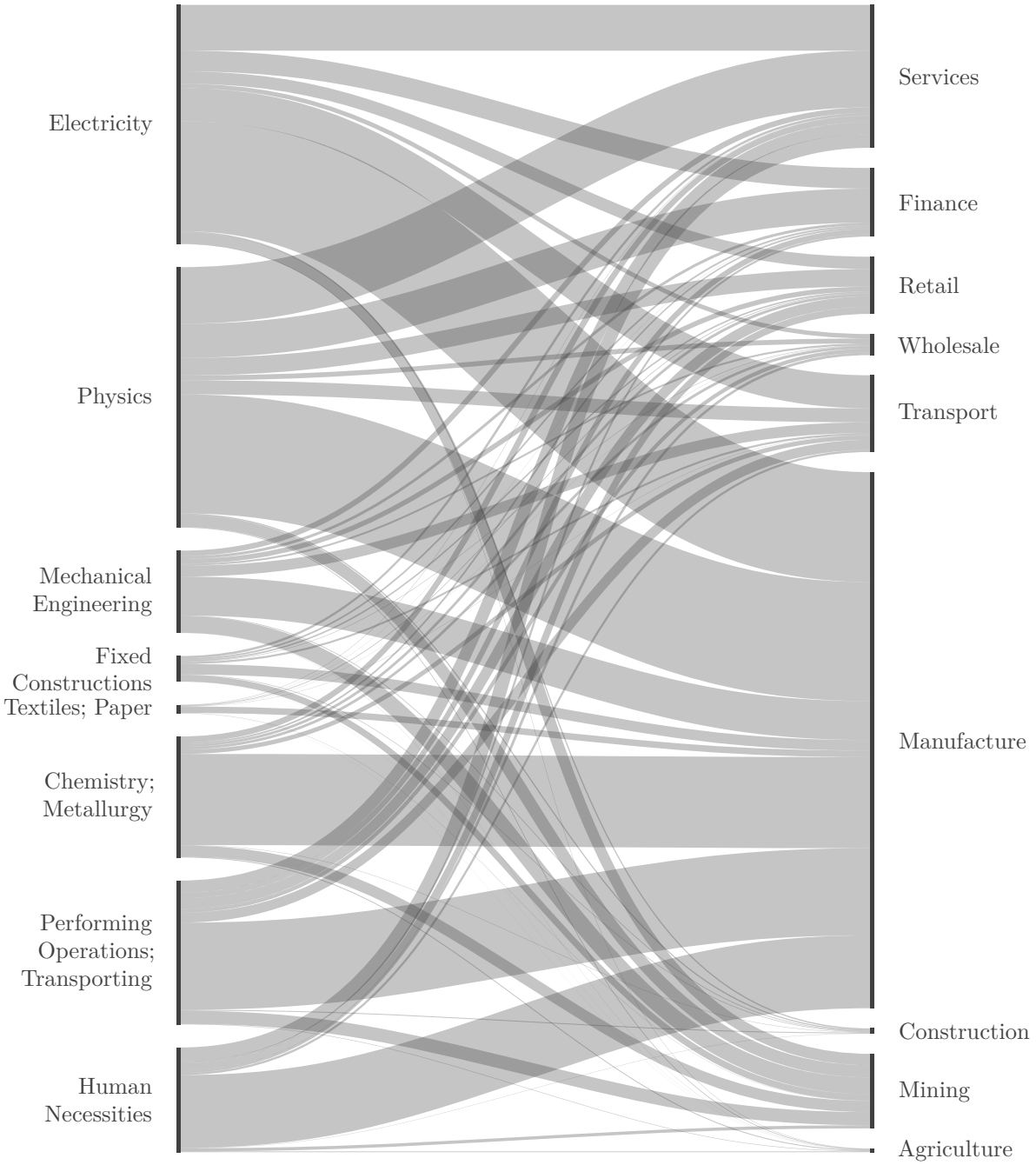
While cosine similarity provides a useful input to our analysis, raw cosine similarity scores should not be used directly to characterize technological associations. First, cosine

similarities depend heavily on the choice of document vectorization method. Different approaches to representing documents as vectors—whether based on term frequencies, TF-IDF weightings, or embeddings—can produce substantially different similarity scores for the same document pair. Second, despite the theoretical invariance of cosine similarity to document length, in practice longer documents tend to receive higher similarity scores due to increased opportunities for vocabulary overlap (Brown and Tucker, 2011). Third, and most fundamentally, cosine similarities measure textual similarity rather than technological usefulness. While greater textual similarity may predict greater technological usefulness, the relationship may be complex and non-linear. This leads to a fourth limitation: cosine similarity scores lack economically meaningful units. That is, a similarity score of 0.8 between two documents does not have an obvious economic or technological meaning, whereas a predicted probability of 0.8 that a patent is useful to a firm provides an interpretable measure that can be used in economic analysis. For these reasons, we use cosine similarities as features in our classifier rather than as direct measures of technological relationships.

B Statistical Appendix

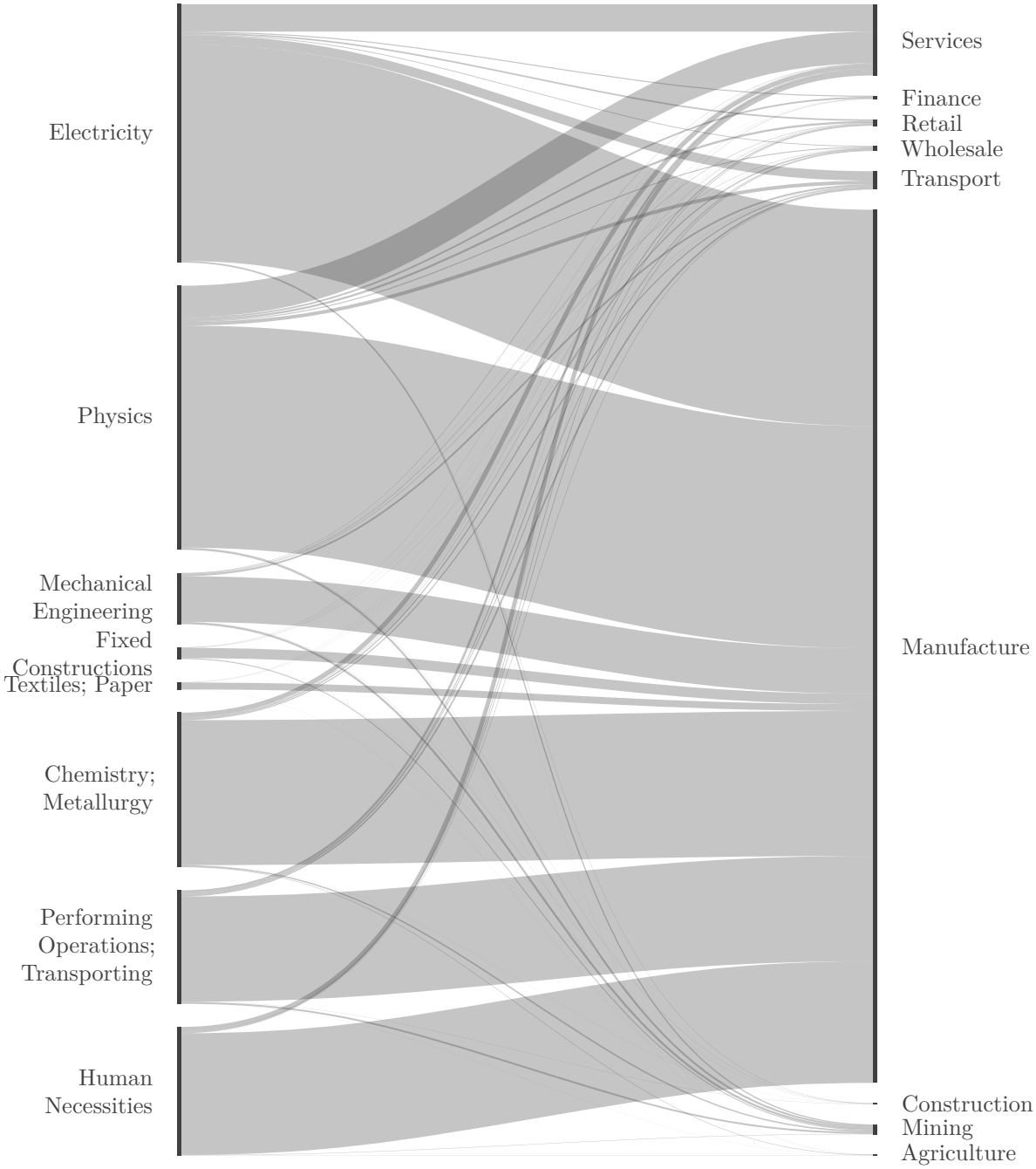
B.1 Sankey Diagrams

Figure B1: CPC-SIC Sankey Diagram: Frequency-Based, Non-Patenting Firms



Notes. The figure shows a Sankey diagram of associations between CPC Sections (left) and SIC Divisions (right). Flows represent frequencies of association, defined as the number of associated firm-patent pairs for a given CPC-SIC combination, aggregated over all sample years. CPC Sections: Human Necessities (A), Performing Operations; Transporting (B), Chemistry; Metallurgy (C), Textiles; Paper (D), Fixed Constructions (E), Mechanical Engineering; Lighting; Heating; Weapons; Blasting (F), Physics (G), and Electricity (H). SIC Divisions: Agriculture, Forestry, and Fishing (0100–0999), Mining (1000–1499), Construction (1500–1799), Manufacturing (2000–3999), Transportation, Communications, Electric, Gas and Sanitary Service (4000–4999), Wholesale Trade (5000–5199), Retail Trade (5200–5999), Finance and Insurance (6000–6799, excl 6500–6599 and 6700–6799), and Services (7000–8999). Some CPC and SIC names have been shortened for the figure.

Figure B2: CPC-SIC Sankey Diagram: Frequency-Based, Patenting Firms



Notes. The figure shows a Sankey diagram of associations between CPC Sections (left) and SIC Divisions (right). Flows represent frequencies of association, defined as the number of associated firm-patent pairs for a given CPC-SIC combination, aggregated over all sample years. CPC Sections: Human Necessities (A), Performing Operations; Transporting (B), Chemistry; Metallurgy (C), Textiles; Paper (D), Fixed Constructions (E), Mechanical Engineering; Lighting; Heating; Weapons; Blasting (F), Physics (G), and Electricity (H). SIC Divisions: Agriculture, Forestry, and Fishing (0100–0999), Mining (1000–1499), Construction (1500–1799), Manufacturing (2000–3999), Transportation, Communications, Electric, Gas and Sanitary Service (4000–4999), Wholesale Trade (5000–5199), Retail Trade (5200–5999), Finance and Insurance (6000–6799, excl 6500–6599 and 6700–6799), and Services (7000–8999). Some CPC and SIC names have been shortened for the figure.

B.2 Technological Breadth and Depth

Table B1: Count Metrics by Industry: CPC Section, in Counts and Fractions

(a) Count Statistics for Non-Patenting (NP) and Patenting (P) Firms

Industry Group	Mean Count			Median Count		
	NP	P	Diff	NP	P	Diff
Finance	0.22	0.50	−0.28	0.00	0.00	0.00
	(0.61)	(1.03)	(0.00)	(0.00)	(1.00)	(1.00)
Service	1.34	1.48	−0.14	1.00	2.00	−1.00
	(1.34)	(1.03)	(0.00)	(2.00)	(1.00)	(0.00)
Resource	2.69	2.67	0.02	3.00	3.00	0.00
	(1.59)	(1.54)	(0.66)	(3.00)	(2.00)	(1.00)
Manufacture	4.06	3.57	0.48	4.00	3.00	1.00
	(1.96)	(1.82)	(0.00)	(4.00)	(3.00)	(0.00)

(b) Fraction Statistics for Non-Patenting (NP) and Patenting (P) Firms

Industry Group	Mean Fraction			Median Fraction		
	NP	P	Diff	NP	P	Diff
Finance	0.03	0.06	−0.03	0.00	0.00	0.00
	(0.08)	(0.13)	(0.00)	(0.00)	(0.12)	(1.00)
Service	0.17	0.18	−0.02	0.12	0.25	−0.12
	(0.17)	(0.13)	(0.00)	(0.25)	(0.12)	(0.00)
Resource	0.34	0.33	0.00	0.38	0.38	0.00
	(0.20)	(0.19)	(0.68)	(0.38)	(0.25)	(1.00)
Manufacture	0.51	0.45	0.06	0.50	0.38	0.12
	(0.24)	(0.23)	(0.00)	(0.50)	(0.38)	(0.00)

Notes. The table summarizes count metrics, reported as counts and fractions by industry group at the CPC Section level. NP and P columns show statistics for non-patenting and patenting firms, respectively, and Diff columns show the difference between them. Means and medians are computed from firm-year observations for the period 1997 to 2023. Panels B1a and B1b report count and fraction statistics, respectively. Standard deviations and interquartile ranges are reported below means and medians, respectively. P-values, reported below differences, are based on 1,000 bootstrap iterations under the null that values for both firm types are drawn from the same distribution. In each iteration, we sample with replacement from the pooled data, compute the difference in means and medians, and report the p-value as $(c+1)/(n+1)$, where c is the count of iterations where the absolute resampled difference exceeds or equals the observed value, and n is the total number of valid bootstrap samples.

Table B2: Count Metrics by Industry: CPC Class, in Counts and Fractions

(a) Count Statistics for Non-Patenting (NP) and Patenting (P) Firms

Industry Group	Mean Count			Median Count		
	NP	P	Diff	NP	P	Diff
Finance	2.66	3.70	−1.05	2.00	2.00	0.00
	(4.05)	(7.54)	(0.00)	(2.00)	(2.00)	(1.00)
Service	10.27	7.16	3.11	6.00	5.00	1.00
	(12.67)	(8.73)	(0.00)	(9.00)	(5.00)	(0.01)
Resource	28.49	27.89	0.60	28.00	26.00	2.00
	(16.15)	(16.25)	(0.25)	(22.00)	(24.00)	(0.02)
Manufacture	37.86	30.77	7.09	35.00	23.00	12.00
	(23.51)	(21.41)	(0.00)	(40.00)	(31.00)	(0.00)

(b) Fraction Statistics for Non-Patenting (NP) and Patenting (P) Firms

Industry Group	Mean Fraction			Median Fraction		
	NP	P	Diff	NP	P	Diff
Finance	0.02	0.03	−0.01	0.02	0.02	−0.00
	(0.03)	(0.06)	(0.00)	(0.02)	(0.02)	(0.24)
Service	0.08	0.06	0.03	0.05	0.04	0.01
	(0.10)	(0.07)	(0.00)	(0.08)	(0.04)	(0.00)
Resource	0.24	0.23	0.01	0.23	0.21	0.02
	(0.13)	(0.13)	(0.22)	(0.18)	(0.20)	(0.01)
Manufacture	0.31	0.25	0.06	0.29	0.19	0.10
	(0.19)	(0.18)	(0.00)	(0.33)	(0.25)	(0.00)

Notes. The table summarizes count metrics, reported as counts and fractions by industry group at the CPC Class level. NP and P columns show statistics for non-patenting and patenting firms, respectively, and Diff columns show the difference between them. Means and medians are computed from firm-year observations for the period 1997 to 2023. Panels B2a and B2b report count and fraction statistics, respectively. Standard deviations and interquartile ranges are reported below means and medians, respectively. P-values, reported below differences, are based on 1,000 bootstrap iterations under the null that values for both firm types are drawn from the same distribution. In each iteration, we sample with replacement from the pooled data, compute the difference in means and medians, and report the p-value as $(c+1)/(n+1)$, where c is the count of iterations where the absolute resampled difference exceeds or equals the observed value, and n is the total number of valid bootstrap samples.

Table B3: Count Metrics by Industry: CPC Subclass, in Counts and Fractions

(a) Count Statistics for Non-Patenting (NP) and Patenting (P) Firms

Industry Group	Mean Count			Median Count		
	NP	P	Diff	NP	P	Diff
Finance	8.99	11.63	−2.64	5.00	5.00	0.00
	(12.65)	(23.37)	(0.00)	(7.00)	(8.00)	(1.00)
Service	32.82	23.46	9.36	21.00	16.00	5.00
	(39.09)	(26.82)	(0.00)	(29.00)	(18.00)	(0.00)
Resource	80.20	79.14	1.05	79.00	74.00	5.00
	(44.57)	(44.57)	(0.45)	(59.00)	(64.00)	(0.01)
Manufacture	114.32	96.15	18.17	102.00	78.00	24.00
	(68.98)	(62.61)	(0.00)	(106.00)	(81.00)	(0.00)

(b) Fraction Statistics for Non-Patenting (NP) and Patenting (P) Firms

Industry Group	Mean Fraction			Median Fraction		
	NP	P	Diff	NP	P	Diff
Finance	0.02	0.02	−0.00	0.01	0.01	−0.00
	(0.02)	(0.04)	(0.00)	(0.01)	(0.01)	(0.39)
Service	0.06	0.04	0.02	0.04	0.03	0.01
	(0.07)	(0.05)	(0.00)	(0.05)	(0.03)	(0.00)
Resource	0.14	0.14	0.00	0.14	0.13	0.01
	(0.08)	(0.08)	(0.42)	(0.10)	(0.11)	(0.01)
Manufacture	0.20	0.17	0.03	0.18	0.13	0.04
	(0.12)	(0.11)	(0.00)	(0.18)	(0.14)	(0.00)

Notes. The table summarizes count metrics, reported as counts and fractions by industry group at the CPC Subclass level. NP and P columns show statistics for non-patenting and patenting firms, respectively, and Diff columns show the difference between them. Means and medians are computed from firm-year observations for the period 1997 to 2023. Panels B3a and B3b report count and fraction statistics, respectively. Standard deviations and interquartile ranges are reported below means and medians, respectively. P-values, reported below differences, are based on 1,000 bootstrap iterations under the null that values for both firm types are drawn from the same distribution. In each iteration, we sample with replacement from the pooled data, compute the difference in means and medians, and report the p-value as $(c+1)/(n+1)$, where c is the count of iterations where the absolute resampled difference exceeds or equals the observed value, and n is the total number of valid bootstrap samples.

Table B4: Count Metrics by Industry: CPC Group, in Counts and Fractions

(a) Count Statistics for Non-Patenting (NP) and Patenting (P) Firms

Industry Group	Mean Count			Median Count		
	NP	P	Diff	NP	P	Diff
Finance	27.04	38.70	−11.66	17.00	21.00	−4.00
	(35.04)	(66.44)	(0.00)	(21.00)	(25.00)	(0.00)
Service	98.14	84.36	13.78	71.00	64.00	7.00
	(110.68)	(83.58)	(0.00)	(83.00)	(61.00)	(0.00)
Resource	204.45	212.85	−8.40	197.00	188.50	8.50
	(119.60)	(128.25)	(0.02)	(165.00)	(186.25)	(0.08)
Manufacture	353.45	319.24	34.21	323.00	281.00	42.00
	(200.09)	(183.20)	(0.00)	(288.00)	(235.00)	(0.00)

(b) Fraction Statistics for Non-Patenting (NP) and Patenting (P) Firms

Industry Group	Mean Fraction			Median Fraction		
	NP	P	Diff	NP	P	Diff
Finance	0.01	0.01	−0.00	0.00	0.01	−0.00
	(0.01)	(0.02)	(0.00)	(0.01)	(0.01)	(0.00)
Service	0.03	0.02	0.00	0.02	0.02	0.00
	(0.03)	(0.02)	(0.00)	(0.02)	(0.02)	(0.00)
Resource	0.05	0.05	−0.00	0.05	0.05	0.00
	(0.03)	(0.03)	(0.04)	(0.04)	(0.05)	(0.43)
Manufacture	0.09	0.08	0.01	0.08	0.07	0.01
	(0.05)	(0.04)	(0.00)	(0.07)	(0.06)	(0.00)

Notes. The table summarizes count metrics, reported as counts and fractions by industry group at the CPC Group level. NP and P columns show statistics for non-patenting and patenting firms, respectively, and Diff columns show the difference between them. Means and medians are computed from firm-year observations for the period 1997 to 2023. Panels B4a and B4b report count and fraction statistics, respectively. Standard deviations and interquartile ranges are reported below means and medians, respectively. P-values, reported below differences, are based on 1,000 bootstrap iterations under the null that values for both firm types are drawn from the same distribution. In each iteration, we sample with replacement from the pooled data, compute the difference in means and medians, and report the p-value as $(c+1)/(n+1)$, where c is the count of iterations where the absolute resampled difference exceeds or equals the observed value, and n is the total number of valid bootstrap samples.

Table B5: Count Metrics by Industry: CPC Patent, in Counts and Fractions

(a) Count Statistics for Non-Patenting (NP) and Patenting (P) Firms

Industry Group	Mean Count			Median Count		
	NP	P	Diff	NP	P	Diff
Finance	845.85	1114.37	−268.53	725.00	851.00	−126.00
	(522.19)	(913.81)	(0.00)	(493.00)	(677.00)	(0.00)
Service	1902.85	2229.30	−326.45	1615.00	2028.00	−413.00
	(1488.97)	(1364.19)	(0.00)	(1422.00)	(1669.00)	(0.00)
Resource	2452.30	2577.32	−125.02	2349.00	2414.50	−65.50
	(1257.68)	(1312.03)	(0.00)	(1580.00)	(1721.75)	(0.21)
Manufacture	4888.13	5342.60	−454.47	5050.00	5497.50	−447.50
	(2122.68)	(2111.09)	(0.00)	(2927.00)	(2774.00)	(0.00)

(b) Fraction Statistics for Non-Patenting (NP) and Patenting (P) Firms

Industry Group	Mean Fraction			Median Fraction		
	NP	P	Diff	NP	P	Diff
Finance	0.02	0.02	−0.01	0.01	0.02	−0.00
	(0.01)	(0.02)	(0.00)	(0.01)	(0.01)	(0.00)
Service	0.04	0.05	−0.01	0.03	0.04	−0.01
	(0.03)	(0.03)	(0.00)	(0.03)	(0.03)	(0.00)
Resource	0.05	0.05	−0.00	0.05	0.05	−0.00
	(0.03)	(0.03)	(0.00)	(0.03)	(0.03)	(0.46)
Manufacture	0.10	0.11	−0.01	0.10	0.11	−0.01
	(0.04)	(0.04)	(0.00)	(0.06)	(0.05)	(0.00)

Notes. The table summarizes count metrics, reported as counts and fractions by industry group at the CPC Patent level. NP and P columns show statistics for non-patenting and patenting firms, respectively, and Diff columns show the difference between them. Means and medians are computed from firm-year observations for the period 1997 to 2023. Panels B5a and B5b report count and fraction statistics, respectively. Standard deviations and interquartile ranges are reported below means and medians, respectively. P-values, reported below differences, are based on 1,000 bootstrap iterations under the null that values for both firm types are drawn from the same distribution. In each iteration, we sample with replacement from the pooled data, compute the difference in means and medians, and report the p-value as $(c+1)/(n+1)$, where c is the count of iterations where the absolute resampled difference exceeds or equals the observed value, and n is the total number of valid bootstrap samples.

Table B6: Count Metrics by Industry and Size: CPC Section, in Counts

Industry Group	Size Class	Mean Rate			Median Rate		
		NP	P	Diff	NP	P	Diff
Finance	Large	0.24	0.35	−0.12	0.00	0.00	0.00
		(0.67)	(0.94)	(0.02)	(0.00)	(0.00)	(1.00)
	Mid	0.22	0.39	−0.17	0.00	0.00	0.00
		(0.58)	(0.59)	(0.03)	(0.00)	(1.00)	(1.00)
	Small	0.18	0.59	−0.40	0.00	0.00	0.00
		(0.52)	(0.90)	(0.00)	(0.00)	(1.00)	(1.00)
	Private	0.27	1.00	−0.73	0.00	0.00	0.00
		(0.71)	(1.53)	(0.00)	(0.00)	(2.00)	(1.00)
Service	Large	1.30	1.45	−0.15	1.00	2.00	−1.00
		(1.34)	(1.04)	(0.00)	(2.00)	(1.00)	(0.00)
	Mid	1.28	1.34	−0.06	1.00	1.00	0.00
		(1.29)	(0.91)	(0.18)	(2.00)	(1.00)	(1.00)
	Small	1.28	1.54	−0.25	1.00	2.00	−1.00
		(1.29)	(1.01)	(0.00)	(2.00)	(1.00)	(0.00)
	Private	1.41	1.51	−0.09	1.00	1.00	0.00
		(1.40)	(1.16)	(0.04)	(2.00)	(1.00)	(1.00)
Resource	Large	2.80	2.70	0.11	3.00	3.00	0.00
		(1.47)	(1.44)	(0.37)	(2.00)	(2.00)	(1.00)
	Mid	2.41	2.71	−0.30	2.00	3.00	−1.00
		(1.51)	(1.40)	(0.01)	(3.00)	(1.00)	(0.44)
	Small	2.72	2.84	−0.13	3.00	3.00	0.00
		(1.62)	(1.64)	(0.17)	(3.00)	(2.00)	(1.00)
	Private	2.71	2.38	0.33	3.00	2.00	1.00
		(1.59)	(1.53)	(0.00)	(2.00)	(2.00)	(0.19)
Manufacture	Large	4.14	3.54	0.60	4.00	3.00	1.00
		(2.01)	(1.85)	(0.00)	(3.00)	(3.00)	(0.06)
	Mid	4.33	3.64	0.69	4.00	3.00	1.00
		(1.94)	(1.91)	(0.00)	(3.00)	(3.00)	(0.00)
	Small	3.99	3.45	0.54	4.00	3.00	1.00
		(1.94)	(1.76)	(0.00)	(3.00)	(3.00)	(0.17)
	Private	4.07	3.99	0.08	4.00	4.00	0.00
		(1.97)	(1.85)	(0.04)	(4.00)	(3.00)	(1.00)

Notes. The table summarizes count metrics, reported as counts by industry group and size class at the CPC Section level. NP and P columns show statistics for non-patenting and patenting firms, respectively, and Diff columns show the difference between them. Means and medians are computed from firm-year observations for the period 1997 to 2023. Standard deviations and interquartile ranges are reported below means and medians, respectively. P-values, reported below differences, are based on 1,000 bootstrap iterations under the null that both groups are drawn from the same distribution. In each iteration, we sample with replacement from the pooled data, compute the difference in means and medians, and report the p-value as the proportion of iterations where the absolute resampled difference exceeds or equals the observed value.

Table B7: Count Metrics by Industry and Size: CPC Class, in Counts

Industry Group	Size Class	Mean Rate			Median Rate		
		NP	P	Diff	NP	P	Diff
Finance	Large	2.70	3.26	−0.56	2.00	2.00	0.00
		(5.07)	(8.33)	(0.15)	(2.00)	(2.00)	(1.00)
	Mid	2.27	2.46	−0.20	2.00	2.00	0.00
		(3.65)	(2.96)	(0.63)	(2.00)	(2.00)	(1.00)
	Small	2.39	3.04	−0.65	2.00	2.00	0.00
		(3.16)	(3.96)	(0.05)	(2.00)	(2.00)	(1.00)
	Private	3.11	7.23	−4.12	2.00	2.50	−0.50
		(4.85)	(9.74)	(0.00)	(2.00)	(10.00)	(0.30)
Service	Large	9.85	6.73	3.12	6.00	4.00	2.00
		(12.62)	(9.45)	(0.00)	(9.00)	(4.00)	(0.00)
	Mid	9.76	5.44	4.32	6.00	4.00	2.00
		(11.93)	(6.56)	(0.00)	(9.00)	(4.00)	(0.00)
	Small	9.65	7.36	2.29	5.00	5.00	0.00
		(12.08)	(8.41)	(0.00)	(9.00)	(5.00)	(1.00)
	Private	11.04	8.98	2.06	6.00	5.00	1.00
		(13.38)	(10.34)	(0.00)	(11.00)	(7.00)	(0.25)
Resource	Large	30.44	29.06	1.39	31.00	27.00	4.00
		(13.91)	(16.16)	(0.21)	(18.00)	(24.00)	(0.02)
	Mid	26.32	28.64	−2.32	26.00	28.00	−2.00
		(16.02)	(16.12)	(0.06)	(24.00)	(24.00)	(0.31)
	Small	28.38	28.14	0.25	29.00	26.00	3.00
		(16.53)	(17.54)	(0.80)	(24.00)	(29.00)	(0.05)
	Private	28.73	26.23	2.50	28.00	26.00	2.00
		(16.15)	(14.52)	(0.01)	(22.00)	(18.00)	(0.27)
Manufacture	Large	40.18	32.01	8.17	41.00	25.00	16.00
		(22.53)	(21.81)	(0.00)	(38.00)	(34.25)	(0.00)
	Mid	42.76	32.68	10.08	43.00	25.00	18.00
		(23.08)	(22.72)	(0.00)	(38.00)	(36.00)	(0.00)
	Small	37.13	28.74	8.39	33.00	22.00	11.00
		(23.81)	(20.22)	(0.00)	(41.00)	(25.00)	(0.00)
	Private	37.35	34.49	2.86	34.00	27.00	7.00
		(23.21)	(22.62)	(0.00)	(39.00)	(36.00)	(0.00)

Notes. The table summarizes count metrics, reported as counts by industry group and size class at the CPC Class level. NP and P columns show statistics for non-patenting and patenting firms, respectively, and Diff columns show the difference between them. Means and medians are computed from firm-year observations for the period 1997 to 2023. Standard deviations and interquartile ranges are reported below means and medians, respectively. P-values, reported below differences, are based on 1,000 bootstrap iterations under the null that both groups are drawn from the same distribution. In each iteration, we sample with replacement from the pooled data, compute the difference in means and medians, and report the p-value as the proportion of iterations where the absolute resampled difference exceeds or equals the observed value.

Table B8: Count Metrics by Industry and Size: CPC Subclass, in Counts

Industry Group	Size Class	Mean Rate			Median Rate		
		NP	P	Diff	NP	P	Diff
Finance	Large	9.08	10.18	−1.10	5.00	5.00	0.00
		(15.63)	(26.19)	(0.35)	(8.00)	(6.50)	(1.00)
	Mid	7.81	7.07	0.74	5.00	4.50	0.50
		(11.63)	(7.76)	(0.59)	(6.00)	(6.00)	(0.58)
	Small	7.94	10.85	−2.91	5.00	5.00	0.00
		(10.10)	(15.75)	(0.02)	(6.00)	(8.75)	(1.00)
	Private	10.64	21.78	−11.14	6.00	8.00	−2.00
		(14.93)	(27.34)	(0.00)	(9.00)	(30.50)	(0.07)
Service	Large	32.93	22.47	10.46	23.00	15.00	8.00
		(41.38)	(30.80)	(0.00)	(26.00)	(16.25)	(0.00)
	Mid	31.93	18.15	13.78	20.00	13.00	7.00
		(38.41)	(20.94)	(0.00)	(28.25)	(14.00)	(0.00)
	Small	30.99	23.91	7.08	19.00	17.00	2.00
		(38.29)	(25.54)	(0.00)	(27.00)	(18.00)	(0.00)
	Private	34.81	29.13	5.68	23.00	20.00	3.00
		(39.69)	(30.05)	(0.00)	(32.00)	(25.00)	(0.01)
Resource	Large	87.48	82.18	5.30	89.00	76.00	13.00
		(38.94)	(45.45)	(0.10)	(50.00)	(62.00)	(0.00)
	Mid	75.49	79.90	−4.40	74.00	77.00	−3.00
		(44.63)	(45.36)	(0.21)	(63.00)	(66.00)	(0.54)
	Small	78.20	78.75	−0.56	78.00	73.00	5.00
		(45.17)	(46.47)	(0.83)	(64.00)	(72.00)	(0.19)
	Private	81.19	77.04	4.15	79.00	73.00	6.00
		(44.62)	(40.82)	(0.12)	(59.00)	(52.00)	(0.10)
Manufacture	Large	118.17	98.68	19.50	116.00	81.00	35.00
		(65.48)	(64.29)	(0.00)	(99.00)	(97.00)	(0.00)
	Mid	126.42	101.33	25.10	122.00	81.00	41.00
		(68.47)	(66.92)	(0.00)	(105.75)	(95.00)	(0.00)
	Small	113.46	90.80	22.66	99.00	74.00	25.00
		(70.22)	(59.45)	(0.00)	(110.00)	(68.00)	(0.00)
	Private	112.15	106.49	5.66	100.00	86.00	14.00
		(67.67)	(64.41)	(0.00)	(103.00)	(91.00)	(0.00)

Notes. The table summarizes count metrics, reported as counts by industry group and size class at the CPC Subclass level. NP and P columns show statistics for non-patenting and patenting firms, respectively, and Diff columns show the difference between them. Means and medians are computed from firm-year observations for the period 1997 to 2023. Standard deviations and interquartile ranges are reported below means and medians, respectively. P-values, reported below differences, are based on 1,000 bootstrap iterations under the null that both groups are drawn from the same distribution. In each iteration, we sample with replacement from the pooled data, compute the difference in means and medians, and report the p-value as the proportion of iterations where the absolute resampled difference exceeds or equals the observed value.

Table B9: Count Metrics by Industry and Size: CPC Group, in Counts

Industry Group	Size Class	Mean Rate			Median Rate		
		NP	P	Diff	NP	P	Diff
Finance	Large	28.65	32.53	−3.88	19.00	20.00	−1.00
		(43.11)	(71.42)	(0.24)	(26.00)	(20.00)	(0.71)
	Mid	24.06	29.59	−5.53	16.00	21.00	−5.00
		(32.13)	(34.20)	(0.14)	(20.00)	(24.25)	(0.03)
	Small	23.69	40.26	−16.57	16.00	20.00	−4.00
		(28.35)	(55.80)	(0.00)	(17.00)	(28.00)	(0.02)
	Private	31.87	66.53	−34.67	18.00	31.50	−13.50
		(41.01)	(75.25)	(0.00)	(26.00)	(94.00)	(0.00)
Service	Large	100.12	78.11	22.01	71.50	58.50	13.00
		(124.46)	(95.02)	(0.00)	(78.00)	(51.00)	(0.00)
	Mid	95.73	66.48	29.25	68.00	53.00	15.00
		(111.60)	(66.06)	(0.00)	(79.00)	(47.00)	(0.00)
	Small	93.93	88.39	5.54	66.00	69.00	−3.00
		(110.70)	(81.24)	(0.02)	(75.00)	(62.00)	(0.09)
	Private	102.62	100.01	2.61	76.00	78.00	−2.00
		(108.70)	(90.17)	(0.45)	(91.00)	(77.50)	(0.57)
Resource	Large	224.97	210.72	14.25	226.00	189.00	37.00
		(108.36)	(126.36)	(0.11)	(151.00)	(181.50)	(0.00)
	Mid	193.36	211.12	−17.76	181.00	196.50	−15.50
		(121.27)	(124.84)	(0.05)	(171.00)	(196.75)	(0.24)
	Small	200.11	223.18	−23.06	195.00	195.00	0.00
		(121.62)	(140.21)	(0.00)	(175.00)	(216.00)	(1.00)
	Private	206.41	201.96	4.45	197.00	182.50	14.50
		(119.10)	(114.34)	(0.53)	(161.00)	(145.75)	(0.13)
Manufacture	Large	363.42	320.57	42.85	364.00	284.50	79.50
		(198.46)	(191.98)	(0.00)	(307.00)	(278.25)	(0.00)
	Mid	382.24	330.74	51.50	366.00	289.00	77.00
		(202.80)	(198.70)	(0.00)	(295.00)	(273.00)	(0.00)
	Small	356.50	304.08	52.43	324.00	269.00	55.00
		(202.85)	(172.65)	(0.00)	(296.00)	(204.00)	(0.00)
	Private	342.64	358.21	−15.57	313.00	317.00	−4.00
		(195.66)	(184.11)	(0.00)	(275.00)	(246.00)	(0.43)

Notes. The table summarizes count metrics, reported as counts by industry group and size class at the CPC Group level. NP and P columns show statistics for non-patenting and patenting firms, respectively, and Diff columns show the difference between them. Means and medians are computed from firm-year observations for the period 1997 to 2023. Standard deviations and interquartile ranges are reported below means and medians, respectively. P-values, reported below differences, are based on 1,000 bootstrap iterations under the null that both groups are drawn from the same distribution. In each iteration, we sample with replacement from the pooled data, compute the difference in means and medians, and report the p-value as the proportion of iterations where the absolute resampled difference exceeds or equals the observed value.

Table B10: Count Metrics by Industry and Size: CPC Patent, in Counts

Industry Group	Size Class	Mean Rate			Median Rate		
		NP	P	Diff	NP	P	Diff
Finance	Large	885.47	986.96	−101.49	771.00	842.50	−71.50
		(584.69)	(769.41)	(0.02)	(594.00)	(552.50)	(0.05)
	Mid	812.42	1104.97	−292.55	704.00	876.50	−172.50
		(500.51)	(775.96)	(0.00)	(503.00)	(1000.25)	(0.01)
	Small	796.16	1239.93	−443.77	697.00	879.00	−182.00
		(457.35)	(1032.54)	(0.00)	(442.00)	(909.00)	(0.00)
	Private	912.91	1416.53	−503.62	763.00	832.50	−69.50
		(584.23)	(1218.82)	(0.00)	(551.25)	(1202.75)	(0.24)
Service	Large	1931.83	2121.27	−189.44	1593.00	1910.50	−317.50
		(1688.86)	(1539.23)	(0.00)	(1384.25)	(1601.50)	(0.00)
	Mid	1793.12	1992.75	−199.64	1517.00	1853.00	−336.00
		(1476.25)	(1215.96)	(0.00)	(1255.00)	(1589.25)	(0.00)
	Small	1821.14	2312.02	−490.88	1550.00	2112.00	−562.00
		(1460.78)	(1332.06)	(0.00)	(1297.00)	(1673.75)	(0.00)
	Private	2010.15	2400.20	−390.04	1724.00	2198.00	−474.00
		(1489.94)	(1366.85)	(0.00)	(1604.00)	(1845.00)	(0.00)
Resource	Large	2537.09	2474.19	62.90	2581.50	2282.00	299.50
		(1125.00)	(1233.00)	(0.49)	(1527.25)	(1698.25)	(0.03)
	Mid	2219.40	2470.21	−250.81	2127.50	2431.00	−303.50
		(1190.28)	(1198.42)	(0.01)	(1550.50)	(1554.25)	(0.01)
	Small	2300.86	2705.93	−405.07	2230.00	2479.00	−249.00
		(1245.55)	(1492.05)	(0.00)	(1655.50)	(1976.50)	(0.00)
	Private	2543.56	2554.20	−10.64	2413.50	2355.00	58.50
		(1272.10)	(1172.25)	(0.89)	(1570.00)	(1517.00)	(0.56)
Manufacture	Large	4588.20	5041.97	−453.77	4777.00	5205.00	−428.00
		(2315.14)	(2194.99)	(0.00)	(3286.50)	(2935.00)	(0.00)
	Mid	4746.07	5338.81	−592.74	4866.00	5567.00	−701.00
		(2185.16)	(2232.43)	(0.00)	(3034.00)	(3035.75)	(0.00)
	Small	4913.42	5309.39	−395.97	5103.50	5447.00	−343.50
		(2164.54)	(2112.04)	(0.00)	(2993.00)	(2803.50)	(0.00)
	Private	4916.14	5761.54	−845.40	5060.00	5799.00	−739.00
		(2040.19)	(1767.13)	(0.00)	(2814.00)	(2298.00)	(0.00)

Notes. The table summarizes count metrics, reported as counts by industry group and size class at the CPC Patent level. NP and P columns show statistics for non-patenting and patenting firms, respectively, and Diff columns show the difference between them. Means and medians are computed from firm-year observations for the period 1997 to 2023. Standard deviations and interquartile ranges are reported below means and medians, respectively. P-values, reported below differences, are based on 1,000 bootstrap iterations under the null that both groups are drawn from the same distribution. In each iteration, we sample with replacement from the pooled data, compute the difference in means and medians, and report the p-value as the proportion of iterations where the absolute resampled difference exceeds or equals the observed value.

Table B11: Count Metrics by Industry and Size: CPC Section, in Fractions

Industry Group	Size Class	Mean Rate			Median Rate		
		NP	P	Diff	NP	P	Diff
Finance	Large	0.03	0.04	−0.01	0.00	0.00	0.00
		(0.08)	(0.12)	(0.02)	(0.00)	(0.00)	(1.00)
	Mid	0.03	0.05	−0.02	0.00	0.00	0.00
		(0.07)	(0.07)	(0.02)	(0.00)	(0.12)	(1.00)
	Small	0.02	0.07	−0.05	0.00	0.00	0.00
		(0.07)	(0.11)	(0.00)	(0.00)	(0.12)	(1.00)
	Private	0.03	0.12	−0.09	0.00	0.00	0.00
		(0.09)	(0.19)	(0.00)	(0.00)	(0.25)	(1.00)
Service	Large	0.16	0.18	−0.02	0.12	0.25	−0.12
		(0.17)	(0.13)	(0.00)	(0.25)	(0.12)	(0.00)
	Mid	0.16	0.17	−0.01	0.12	0.12	0.00
		(0.16)	(0.11)	(0.14)	(0.25)	(0.12)	(1.00)
	Small	0.16	0.19	−0.03	0.12	0.25	−0.12
		(0.16)	(0.13)	(0.00)	(0.25)	(0.12)	(0.00)
	Private	0.18	0.19	−0.01	0.12	0.12	0.00
		(0.18)	(0.14)	(0.04)	(0.25)	(0.12)	(1.00)
Resource	Large	0.35	0.34	0.01	0.38	0.38	0.00
		(0.18)	(0.18)	(0.35)	(0.25)	(0.25)	(1.00)
	Mid	0.30	0.34	−0.04	0.25	0.38	−0.12
		(0.19)	(0.18)	(0.01)	(0.38)	(0.12)	(0.48)
	Small	0.34	0.36	−0.02	0.38	0.38	0.00
		(0.20)	(0.20)	(0.15)	(0.38)	(0.25)	(1.00)
	Private	0.34	0.30	0.04	0.38	0.25	0.12
		(0.20)	(0.19)	(0.00)	(0.25)	(0.25)	(0.18)
Manufacture	Large	0.52	0.44	0.07	0.50	0.38	0.12
		(0.25)	(0.23)	(0.00)	(0.38)	(0.38)	(0.07)
	Mid	0.54	0.45	0.09	0.50	0.38	0.12
		(0.24)	(0.24)	(0.00)	(0.38)	(0.38)	(0.00)
	Small	0.50	0.43	0.07	0.50	0.38	0.12
		(0.24)	(0.22)	(0.00)	(0.38)	(0.38)	(0.17)
	Private	0.51	0.50	0.01	0.50	0.50	0.00
		(0.25)	(0.23)	(0.04)	(0.50)	(0.38)	(1.00)

Notes. The table summarizes count metrics, reported as fractions by industry group and size class at the CPC Section level. NP and P columns show statistics for non-patenting and patenting firms, respectively, and Diff columns show the difference between them. Means and medians are computed from firm-year observations for the period 1997 to 2023. Standard deviations and interquartile ranges are reported below means and medians, respectively. P-values, reported below differences, are based on 1,000 bootstrap iterations under the null that both groups are drawn from the same distribution. In each iteration, we sample with replacement from the pooled data, compute the difference in means and medians, and report the p-value as the proportion of iterations where the absolute resampled difference exceeds or equals the observed value.

Table B12: Count Metrics by Industry and Size: CPC Class, in Fractions

Industry Group	Size Class	Mean Rate			Median Rate		
		NP	P	Diff	NP	P	Diff
Finance	Large	0.02	0.03	−0.00	0.02	0.02	−0.00
		(0.04)	(0.07)	(0.14)	(0.02)	(0.02)	(0.47)
	Mid	0.02	0.02	−0.00	0.02	0.02	−0.00
		(0.03)	(0.02)	(0.64)	(0.02)	(0.02)	(0.61)
	Small	0.02	0.03	−0.01	0.02	0.02	−0.00
		(0.03)	(0.03)	(0.05)	(0.02)	(0.02)	(0.55)
	Private	0.03	0.06	−0.03	0.02	0.02	−0.00
		(0.04)	(0.08)	(0.00)	(0.02)	(0.08)	(0.20)
Service	Large	0.08	0.06	0.03	0.05	0.03	0.02
		(0.10)	(0.08)	(0.00)	(0.07)	(0.03)	(0.00)
	Mid	0.08	0.04	0.04	0.05	0.03	0.02
		(0.10)	(0.05)	(0.00)	(0.07)	(0.03)	(0.00)
	Small	0.08	0.06	0.02	0.04	0.04	0.00
		(0.10)	(0.07)	(0.00)	(0.07)	(0.04)	(0.23)
	Private	0.09	0.07	0.02	0.05	0.04	0.01
		(0.11)	(0.09)	(0.00)	(0.09)	(0.06)	(0.00)
Resource	Large	0.25	0.24	0.01	0.26	0.22	0.03
		(0.11)	(0.13)	(0.23)	(0.15)	(0.20)	(0.01)
	Mid	0.22	0.24	−0.02	0.21	0.23	−0.02
		(0.13)	(0.13)	(0.07)	(0.20)	(0.20)	(0.14)
	Small	0.23	0.23	0.00	0.24	0.21	0.02
		(0.14)	(0.14)	(0.78)	(0.20)	(0.24)	(0.02)
	Private	0.24	0.22	0.02	0.23	0.21	0.02
		(0.13)	(0.12)	(0.01)	(0.18)	(0.15)	(0.10)
Manufacture	Large	0.33	0.26	0.07	0.34	0.21	0.14
		(0.19)	(0.18)	(0.00)	(0.31)	(0.28)	(0.00)
	Mid	0.35	0.27	0.08	0.36	0.20	0.15
		(0.19)	(0.19)	(0.00)	(0.31)	(0.30)	(0.00)
	Small	0.31	0.24	0.07	0.27	0.18	0.09
		(0.20)	(0.17)	(0.00)	(0.34)	(0.21)	(0.00)
	Private	0.31	0.28	0.02	0.28	0.22	0.06
		(0.19)	(0.19)	(0.00)	(0.32)	(0.30)	(0.00)

Notes. The table summarizes count metrics, reported as fractions by industry group and size class at the CPC Class level. NP and P columns show statistics for non-patenting and patenting firms, respectively, and Diff columns show the difference between them. Means and medians are computed from firm-year observations for the period 1997 to 2023. Standard deviations and interquartile ranges are reported below means and medians, respectively. P-values, reported below differences, are based on 1,000 bootstrap iterations under the null that both groups are drawn from the same distribution. In each iteration, we sample with replacement from the pooled data, compute the difference in means and medians, and report the p-value as the proportion of iterations where the absolute resampled difference exceeds or equals the observed value.

Table B13: Count Metrics by Industry and Size: CPC Subclass, in Fractions

Industry Group	Size Class	Mean Rate			Median Rate		
		NP	P	Diff	NP	P	Diff
Finance	Large	0.02	0.02	−0.00	0.01	0.01	0.00
		(0.03)	(0.05)	(0.29)	(0.01)	(0.01)	(0.66)
	Mid	0.01	0.01	0.00	0.01	0.01	0.00
		(0.02)	(0.01)	(0.57)	(0.01)	(0.01)	(0.56)
	Small	0.01	0.02	−0.01	0.01	0.01	−0.00
		(0.02)	(0.03)	(0.02)	(0.01)	(0.02)	(0.84)
	Private	0.02	0.04	−0.02	0.01	0.01	−0.00
		(0.03)	(0.05)	(0.00)	(0.01)	(0.05)	(0.01)
Service	Large	0.06	0.04	0.02	0.04	0.03	0.01
		(0.07)	(0.05)	(0.00)	(0.04)	(0.03)	(0.00)
	Mid	0.06	0.03	0.02	0.04	0.02	0.01
		(0.07)	(0.04)	(0.00)	(0.05)	(0.02)	(0.00)
	Small	0.05	0.04	0.01	0.03	0.03	0.00
		(0.07)	(0.04)	(0.00)	(0.05)	(0.03)	(0.00)
	Private	0.06	0.05	0.01	0.04	0.03	0.01
		(0.07)	(0.05)	(0.00)	(0.05)	(0.04)	(0.00)
Resource	Large	0.15	0.14	0.01	0.15	0.13	0.02
		(0.07)	(0.08)	(0.12)	(0.09)	(0.11)	(0.00)
	Mid	0.13	0.14	−0.01	0.13	0.13	−0.00
		(0.08)	(0.08)	(0.23)	(0.11)	(0.11)	(0.80)
	Small	0.14	0.14	−0.00	0.14	0.13	0.01
		(0.08)	(0.08)	(0.72)	(0.11)	(0.12)	(0.15)
	Private	0.14	0.13	0.01	0.14	0.13	0.01
		(0.08)	(0.07)	(0.09)	(0.10)	(0.09)	(0.07)
Manufacture	Large	0.20	0.17	0.03	0.20	0.14	0.06
		(0.11)	(0.11)	(0.00)	(0.17)	(0.17)	(0.00)
	Mid	0.22	0.17	0.04	0.21	0.14	0.07
		(0.12)	(0.11)	(0.00)	(0.18)	(0.16)	(0.00)
	Small	0.19	0.16	0.04	0.17	0.13	0.04
		(0.12)	(0.10)	(0.00)	(0.19)	(0.12)	(0.00)
	Private	0.19	0.18	0.01	0.17	0.15	0.02
		(0.12)	(0.11)	(0.00)	(0.18)	(0.15)	(0.00)

Notes. The table summarizes count metrics, reported as fractions by industry group and size class at the CPC Subclass level. NP and P columns show statistics for non-patenting and patenting firms, respectively, and Diff columns show the difference between them. Means and medians are computed from firm-year observations for the period 1997 to 2023. Standard deviations and interquartile ranges are reported below means and medians, respectively. P-values, reported below differences, are based on 1,000 bootstrap iterations under the null that both groups are drawn from the same distribution. In each iteration, we sample with replacement from the pooled data, compute the difference in means and medians, and report the p-value as the proportion of iterations where the absolute resampled difference exceeds or equals the observed value.

Table B14: Count Metrics by Industry and Size: CPC Group, in Fractions

Industry Group	Size Class	Mean Rate			Median Rate		
		NP	P	Diff	NP	P	Diff
Finance	Large	0.01	0.01	−0.00	0.00	0.01	−0.00
		(0.01)	(0.02)	(0.09)	(0.01)	(0.01)	(0.19)
	Mid	0.01	0.01	−0.00	0.00	0.01	−0.00
		(0.01)	(0.01)	(0.17)	(0.01)	(0.01)	(0.02)
	Small	0.01	0.01	−0.00	0.00	0.01	−0.00
		(0.01)	(0.01)	(0.00)	(0.00)	(0.01)	(0.00)
	Private	0.01	0.02	−0.01	0.00	0.01	−0.00
		(0.01)	(0.02)	(0.00)	(0.01)	(0.02)	(0.00)
Service	Large	0.03	0.02	0.01	0.02	0.02	0.00
		(0.03)	(0.03)	(0.00)	(0.02)	(0.01)	(0.00)
	Mid	0.02	0.02	0.01	0.02	0.01	0.00
		(0.03)	(0.02)	(0.00)	(0.02)	(0.01)	(0.00)
	Small	0.02	0.02	0.00	0.02	0.02	−0.00
		(0.03)	(0.02)	(0.05)	(0.02)	(0.02)	(0.02)
	Private	0.03	0.03	0.00	0.02	0.02	−0.00
		(0.03)	(0.02)	(0.27)	(0.02)	(0.02)	(0.54)
Resource	Large	0.06	0.05	0.00	0.06	0.05	0.01
		(0.03)	(0.03)	(0.22)	(0.04)	(0.05)	(0.00)
	Mid	0.05	0.05	−0.00	0.05	0.05	−0.00
		(0.03)	(0.03)	(0.08)	(0.04)	(0.05)	(0.40)
	Small	0.05	0.06	−0.01	0.05	0.05	−0.00
		(0.03)	(0.04)	(0.00)	(0.04)	(0.06)	(0.52)
	Private	0.05	0.05	0.00	0.05	0.05	0.00
		(0.03)	(0.03)	(0.23)	(0.04)	(0.04)	(0.14)
Manufacture	Large	0.09	0.08	0.01	0.09	0.07	0.02
		(0.05)	(0.05)	(0.00)	(0.07)	(0.07)	(0.00)
	Mid	0.10	0.08	0.01	0.09	0.07	0.02
		(0.05)	(0.05)	(0.00)	(0.07)	(0.07)	(0.00)
	Small	0.09	0.08	0.01	0.08	0.07	0.01
		(0.05)	(0.04)	(0.00)	(0.07)	(0.05)	(0.00)
	Private	0.09	0.09	−0.00	0.08	0.08	0.00
		(0.05)	(0.04)	(0.00)	(0.07)	(0.06)	(0.47)

Notes. The table summarizes count metrics, reported as fractions by industry group and size class at the CPC Group level. NP and P columns show statistics for non-patenting and patenting firms, respectively, and Diff columns show the difference between them. Means and medians are computed from firm-year observations for the period 1997 to 2023. Standard deviations and interquartile ranges are reported below means and medians, respectively. P-values, reported below differences, are based on 1,000 bootstrap iterations under the null that both groups are drawn from the same distribution. In each iteration, we sample with replacement from the pooled data, compute the difference in means and medians, and report the p-value as the proportion of iterations where the absolute resampled difference exceeds or equals the observed value.

Table B15: Count Metrics by Industry and Size: CPC Patent, in Fractions

Industry Group	Size Class	Mean Rate			Median Rate		
		NP	P	Diff	NP	P	Diff
Finance	Large	0.02	0.02	−0.00	0.02	0.02	−0.00
		(0.01)	(0.02)	(0.01)	(0.01)	(0.01)	(0.02)
	Mid	0.02	0.02	−0.01	0.01	0.02	−0.00
		(0.01)	(0.02)	(0.00)	(0.01)	(0.02)	(0.01)
	Small	0.02	0.03	−0.01	0.01	0.02	−0.00
		(0.01)	(0.02)	(0.00)	(0.01)	(0.02)	(0.00)
	Private	0.02	0.03	−0.01	0.02	0.02	−0.00
		(0.01)	(0.02)	(0.00)	(0.01)	(0.02)	(0.11)
Service	Large	0.04	0.04	−0.00	0.03	0.04	−0.01
		(0.03)	(0.03)	(0.00)	(0.03)	(0.03)	(0.00)
	Mid	0.04	0.04	−0.00	0.03	0.04	−0.01
		(0.03)	(0.02)	(0.00)	(0.02)	(0.03)	(0.00)
	Small	0.04	0.05	−0.01	0.03	0.04	−0.01
		(0.03)	(0.03)	(0.00)	(0.03)	(0.03)	(0.00)
	Private	0.04	0.05	−0.01	0.04	0.04	−0.01
		(0.03)	(0.03)	(0.00)	(0.03)	(0.04)	(0.00)
Resource	Large	0.05	0.05	0.00	0.05	0.05	0.01
		(0.02)	(0.02)	(0.31)	(0.03)	(0.03)	(0.01)
	Mid	0.05	0.05	−0.00	0.04	0.05	−0.01
		(0.02)	(0.02)	(0.01)	(0.03)	(0.03)	(0.02)
	Small	0.05	0.05	−0.01	0.05	0.05	−0.00
		(0.03)	(0.03)	(0.00)	(0.03)	(0.04)	(0.01)
	Private	0.05	0.05	−0.00	0.05	0.05	0.00
		(0.03)	(0.03)	(0.74)	(0.03)	(0.03)	(0.58)
Manufacture	Large	0.09	0.10	−0.01	0.10	0.10	−0.01
		(0.04)	(0.04)	(0.00)	(0.06)	(0.06)	(0.00)
	Mid	0.10	0.11	−0.01	0.10	0.11	−0.01
		(0.04)	(0.04)	(0.00)	(0.06)	(0.06)	(0.00)
	Small	0.10	0.11	−0.01	0.10	0.11	−0.01
		(0.04)	(0.04)	(0.00)	(0.06)	(0.05)	(0.00)
	Private	0.10	0.12	−0.02	0.10	0.12	−0.01
		(0.04)	(0.03)	(0.00)	(0.06)	(0.05)	(0.00)

Notes. The table summarizes count metrics, reported as fractions by industry group and size class at the CPC Patent level. NP and P columns show statistics for non-patenting and patenting firms, respectively, and Diff columns show the difference between them. Means and medians are computed from firm-year observations for the period 1997 to 2023. Standard deviations and interquartile ranges are reported below means and medians, respectively. P-values, reported below differences, are based on 1,000 bootstrap iterations under the null that both groups are drawn from the same distribution. In each iteration, we sample with replacement from the pooled data, compute the difference in means and medians, and report the p-value as the proportion of iterations where the absolute resampled difference exceeds or equals the observed value.

B.3 Technological Instability

Table B16: Churn Metrics by Industry: CPC Section, in Add Rates and Drop Rates

(a) Add Rate Statistics for Non-Patenting (NP) and Patenting (P) Firms

Industry Group	Mean Add Rate			Median Add Rate		
	NP	P	Diff	NP	P	Diff
Finance	12.38	16.36	−3.98	0.00	0.00	0.00
	(33.02)	(37.59)	(0.02)	(0.00)	(0.00)	(1.00)
Service	29.52	20.63	8.89	0.00	0.00	0.00
	(45.04)	(38.73)	(0.00)	(66.67)	(0.00)	(1.00)
Resource	25.29	21.87	3.42	0.00	0.00	0.00
	(40.03)	(37.61)	(0.01)	(40.00)	(40.00)	(1.00)
Manufacture	14.04	9.23	4.81	0.00	0.00	0.00
	(28.54)	(20.53)	(0.00)	(18.18)	(0.00)	(1.00)

(b) Drop Rate Statistics for Non-Patenting (NP) and Patenting (P) Firms

Industry Group	Mean Drop Rate			Median Drop Rate		
	NP	P	Diff	NP	P	Diff
Finance	11.63	15.75	−4.13	0.00	0.00	0.00
	(32.09)	(36.26)	(0.01)	(0.00)	(0.00)	(1.00)
Service	28.97	21.28	7.69	0.00	0.00	0.00
	(44.75)	(39.71)	(0.00)	(66.67)	(0.00)	(1.00)
Resource	21.31	21.16	0.15	0.00	0.00	0.00
	(37.23)	(36.48)	(0.90)	(28.57)	(40.00)	(1.00)
Manufacture	13.29	10.30	2.99	0.00	0.00	0.00
	(26.44)	(22.39)	(0.00)	(18.18)	(13.33)	(1.00)

Notes. The table summarizes churn metrics, reported as add rates and drop rates by industry group at the CPC Section level. NP and P columns show statistics for non-patenting and patenting firms, respectively, and Diff columns show the difference between them. Means and medians are computed from firm-year observations for the period 1997 to 2023. Panels B16a and B16b report add rate and drop rate statistics, respectively. Standard deviations and interquartile ranges are reported below means and medians, respectively. P-values, reported below differences, are based on 1,000 bootstrap iterations under the null that values for both firm types are drawn from the same distribution. In each iteration, we sample with replacement from the pooled data, compute the difference in means and medians, and report the p-value as $(c + 1)/(n + 1)$, where c is the count of iterations where the absolute resampled difference exceeds or equals the observed value, and n is the total number of valid bootstrap samples.

Table B17: Churn Metrics by Industry: CPC Class, in Add Rates and Drop Rates

(a) Add Rate Statistics for Non-Patenting (NP) and Patenting (P) Firms

Industry Group	Mean Add Rate			Median Add Rate		
	NP	P	Diff	NP	P	Diff
Finance	41.62	38.27	3.35	25.00	14.29	10.71
	(47.06)	(45.93)	(0.13)	(80.00)	(66.67)	(0.15)
Service	34.34	23.70	10.65	20.00	10.53	9.47
	(41.69)	(33.09)	(0.00)	(50.00)	(38.25)	(0.00)
Resource	32.35	27.59	4.76	19.67	16.13	3.54
	(37.71)	(32.43)	(0.00)	(33.07)	(28.56)	(0.00)
Manufacture	23.76	18.09	5.67	15.38	13.64	1.75
	(28.70)	(19.06)	(0.00)	(19.66)	(16.75)	(0.00)

(b) Drop Rate Statistics for Non-Patenting (NP) and Patenting (P) Firms

Industry Group	Mean Drop Rate			Median Drop Rate		
	NP	P	Diff	NP	P	Diff
Finance	38.25	36.63	1.62	5.88	9.55	−3.66
	(46.86)	(46.63)	(0.49)	(66.67)	(66.67)	(0.74)
Service	35.05	26.58	8.48	20.69	14.29	6.40
	(42.01)	(35.25)	(0.00)	(52.83)	(40.00)	(0.00)
Resource	29.02	27.38	1.64	17.24	17.45	−0.20
	(34.21)	(30.44)	(0.15)	(31.15)	(29.45)	(0.80)
Manufacture	23.50	20.47	3.02	15.38	14.75	0.64
	(26.61)	(21.63)	(0.00)	(21.12)	(18.82)	(0.00)

Notes. The table summarizes churn metrics, reported as add rates and drop rates by industry group at the CPC Class level. NP and P columns show statistics for non-patenting and patenting firms, respectively, and Diff columns show the difference between them. Means and medians are computed from firm-year observations for the period 1997 to 2023. Panels B17a and B17b report add rate and drop rate statistics, respectively. Standard deviations and interquartile ranges are reported below means and medians, respectively. P-values, reported below differences, are based on 1,000 bootstrap iterations under the null that values for both firm types are drawn from the same distribution. In each iteration, we sample with replacement from the pooled data, compute the difference in means and medians, and report the p-value as $(c + 1)/(n + 1)$, where c is the count of iterations where the absolute resampled difference exceeds or equals the observed value, and n is the total number of valid bootstrap samples.

Table B18: Churn Metrics by Industry: CPC Subclass, in Add Rates and Drop Rates

(a) Add Rate Statistics for Non-Patenting (NP) and Patenting (P) Firms

Industry Group	Mean Add Rate			Median Add Rate		
	NP	P	Diff	NP	P	Diff
Finance	43.31	38.41	4.90	33.33	28.08	5.25
	(41.00)	(41.37)	(0.02)	(61.57)	(63.87)	(0.11)
Service	40.03	29.73	10.31	27.78	21.05	6.73
	(39.14)	(30.63)	(0.00)	(43.03)	(31.30)	(0.00)
Resource	39.37	34.89	4.47	28.57	26.09	2.48
	(35.88)	(30.83)	(0.00)	(30.78)	(27.75)	(0.01)
Manufacture	30.20	23.67	6.53	22.52	19.80	2.71
	(27.80)	(17.86)	(0.00)	(19.32)	(15.76)	(0.00)

(b) Drop Rate Statistics for Non-Patenting (NP) and Patenting (P) Firms

Industry Group	Mean Drop Rate			Median Drop Rate		
	NP	P	Diff	NP	P	Diff
Finance	40.69	38.84	1.84	28.57	26.67	1.90
	(42.30)	(42.24)	(0.39)	(66.67)	(65.02)	(0.43)
Service	41.31	33.58	7.73	28.89	24.00	4.89
	(39.57)	(33.03)	(0.00)	(44.03)	(35.85)	(0.00)
Resource	36.09	35.09	0.99	26.04	26.39	−0.35
	(32.77)	(29.66)	(0.35)	(30.26)	(29.17)	(0.73)
Manufacture	30.19	26.37	3.82	22.83	21.28	1.56
	(25.65)	(20.19)	(0.00)	(20.43)	(17.48)	(0.00)

Notes. The table summarizes churn metrics, reported as add rates and drop rates by industry group at the CPC Subclass level. NP and P columns show statistics for non-patenting and patenting firms, respectively, and Diff columns show the difference between them. Means and medians are computed from firm-year observations for the period 1997 to 2023. Panels B18a and B18b report add rate and drop rate statistics, respectively. Standard deviations and interquartile ranges are reported below means and medians, respectively. P-values, reported below differences, are based on 1,000 bootstrap iterations under the null that values for both firm types are drawn from the same distribution. In each iteration, we sample with replacement from the pooled data, compute the difference in means and medians, and report the p-value as $(c + 1)/(n + 1)$, where c is the count of iterations where the absolute resampled difference exceeds or equals the observed value, and n is the total number of valid bootstrap samples.

Table B19: Churn Metrics by Industry: CPC Group, in Add Rates and Drop Rates

(a) Add Rate Statistics for Non-Patenting (NP) and Patenting (P) Firms

Industry Group	Mean Add Rate			Median Add Rate		
	NP	P	Diff	NP	P	Diff
Finance	53.67	47.03	6.65	47.62	38.87	8.75
	(36.67)	(37.55)	(0.00)	(50.68)	(47.19)	(0.00)
Service	48.86	35.19	13.66	38.33	26.75	11.58
	(37.81)	(29.30)	(0.00)	(43.75)	(29.27)	(0.00)
Resource	54.73	49.11	5.62	46.67	42.48	4.19
	(34.44)	(29.56)	(0.00)	(33.27)	(30.08)	(0.00)
Manufacture	42.75	34.40	8.34	36.07	30.69	5.38
	(27.39)	(18.30)	(0.00)	(22.20)	(17.98)	(0.00)

(b) Drop Rate Statistics for Non-Patenting (NP) and Patenting (P) Firms

Industry Group	Mean Drop Rate			Median Drop Rate		
	NP	P	Diff	NP	P	Diff
Finance	51.07	48.32	2.75	41.75	39.35	2.39
	(38.37)	(38.79)	(0.15)	(49.21)	(43.92)	(0.28)
Service	50.23	38.89	11.34	39.93	29.98	9.95
	(38.15)	(31.17)	(0.00)	(44.41)	(32.11)	(0.00)
Resource	51.51	50.19	1.32	44.38	44.78	−0.40
	(31.99)	(29.03)	(0.21)	(31.96)	(32.09)	(0.66)
Manufacture	43.34	37.97	5.38	37.17	33.41	3.76
	(25.03)	(20.00)	(0.00)	(22.23)	(18.23)	(0.00)

Notes. The table summarizes churn metrics, reported as add rates and drop rates by industry group at the CPC Group level. NP and P columns show statistics for non-patenting and patenting firms, respectively, and Diff columns show the difference between them. Means and medians are computed from firm-year observations for the period 1997 to 2023. Panels B19a and B19b report add rate and drop rate statistics, respectively. Standard deviations and interquartile ranges are reported below means and medians, respectively. P-values, reported below differences, are based on 1,000 bootstrap iterations under the null that values for both firm types are drawn from the same distribution. In each iteration, we sample with replacement from the pooled data, compute the difference in means and medians, and report the p-value as $(c + 1)/(n + 1)$, where c is the count of iterations where the absolute resampled difference exceeds or equals the observed value, and n is the total number of valid bootstrap samples.

Table B20: Churn Metrics by Industry and Size: CPC Section, in Add Rates

Industry Group	Size Class	Mean Rate			Median Rate		
		NP	P	Diff	NP	P	Diff
Finance	Large	15.21	13.90	1.31	0.00	0.00	0.00
		(36.59)	(33.95)	(0.62)	(0.00)	(0.00)	(1.00)
	Mid	12.91	17.95	−5.03	0.00	0.00	0.00
		(33.61)	(38.18)	(0.29)	(0.00)	(0.00)	(1.00)
	Small	11.01	10.65	0.36	0.00	0.00	0.00
		(31.34)	(30.94)	(0.94)	(0.00)	(0.00)	(1.00)
	Private	13.60	32.14	−18.54	0.00	0.00	0.00
		(34.35)	(52.59)	(0.00)	(0.00)	(66.67)	(1.00)
Service	Large	29.54	18.84	10.70	0.00	0.00	0.00
		(45.47)	(37.63)	(0.00)	(66.67)	(0.00)	(1.00)
	Mid	28.54	23.03	5.51	0.00	0.00	0.00
		(44.42)	(40.39)	(0.00)	(66.67)	(66.67)	(1.00)
	Small	28.43	20.02	8.41	0.00	0.00	0.00
		(44.34)	(38.20)	(0.00)	(66.67)	(0.00)	(1.00)
	Private	30.93	21.33	9.60	0.00	0.00	0.00
		(45.85)	(39.17)	(0.00)	(66.67)	(28.57)	(1.00)
Resource	Large	21.95	20.98	0.97	0.00	0.00	0.00
		(37.27)	(36.64)	(0.74)	(33.33)	(40.00)	(1.00)
	Mid	21.25	17.81	3.44	0.00	0.00	0.00
		(37.22)	(33.65)	(0.25)	(28.57)	(24.31)	(1.00)
	Small	23.14	22.92	0.22	0.00	0.00	0.00
		(37.58)	(39.10)	(0.91)	(40.00)	(34.29)	(1.00)
	Private	27.11	23.92	3.19	0.00	0.00	0.00
		(41.52)	(38.78)	(0.26)	(40.00)	(40.00)	(1.00)
Manufacture	Large	13.27	10.81	2.47	0.00	0.00	0.00
		(25.98)	(23.52)	(0.00)	(18.18)	(13.33)	(1.00)
	Mid	12.01	9.57	2.44	0.00	0.00	0.00
		(24.04)	(20.49)	(0.00)	(18.18)	(13.33)	(1.00)
	Small	11.12	8.64	2.49	0.00	0.00	0.00
		(22.98)	(19.58)	(0.00)	(16.67)	(0.00)	(1.00)
	Private	17.98	9.51	8.48	0.00	0.00	0.00
		(34.47)	(20.89)	(0.00)	(22.22)	(13.33)	(1.00)

Notes. The table summarizes churn metrics, reported as add rates by industry group and size class at the CPC Section level. NP and P columns show statistics for non-patenting and patenting firms, respectively, and Diff columns show the difference between them. Means and medians are computed from firm-year observations for the period 1997 to 2023. Standard deviations and interquartile ranges are reported below means and medians, respectively. P-values, reported below differences, are based on 1,000 bootstrap iterations under the null that both groups are drawn from the same distribution. In each iteration, we sample with replacement from the pooled data, compute the difference in means and medians, and report the p-value as the proportion of iterations where the absolute resampled difference exceeds or equals the observed value.

Table B21: Churn Metrics by Industry and Size: CPC Class, in Add Rates

Industry Group	Size Class	Mean Rate			Median Rate		
		NP	P	Diff	NP	P	Diff
Finance	Large	41.26	36.24	5.02	22.22	0.00	22.22
		(48.51)	(44.85)	(0.16)	(80.00)	(66.67)	(0.11)
	Mid	39.94	37.43	2.51	16.67	18.01	−1.35
		(47.26)	(44.12)	(0.71)	(75.00)	(66.67)	(0.94)
	Small	40.90	36.05	4.85	28.57	21.05	7.52
		(46.03)	(39.01)	(0.36)	(66.67)	(66.67)	(0.64)
	Private	43.19	49.93	−6.73	27.27	20.02	7.25
		(48.08)	(58.35)	(0.33)	(84.59)	(100.00)	(0.70)
Service	Large	32.13	23.62	8.51	17.14	6.67	10.48
		(41.02)	(34.30)	(0.00)	(50.00)	(36.36)	(0.00)
	Mid	31.70	24.45	7.24	17.98	4.26	13.72
		(39.03)	(34.41)	(0.00)	(50.00)	(40.00)	(0.00)
	Small	31.07	21.72	9.35	16.67	8.70	7.97
		(39.08)	(31.12)	(0.00)	(48.24)	(33.33)	(0.00)
	Private	38.75	27.92	10.83	23.53	16.67	6.86
		(44.62)	(34.79)	(0.00)	(59.87)	(40.00)	(0.00)
Resource	Large	23.50	27.55	−4.05	14.87	16.67	−1.80
		(28.17)	(32.25)	(0.10)	(24.17)	(29.74)	(0.30)
	Mid	29.25	27.37	1.88	16.67	14.34	2.32
		(36.10)	(33.07)	(0.53)	(29.62)	(34.32)	(0.25)
	Small	30.27	28.81	1.46	19.05	17.70	1.35
		(34.93)	(32.28)	(0.46)	(32.73)	(31.45)	(0.31)
	Private	34.41	26.10	8.30	20.90	15.27	5.63
		(39.47)	(32.42)	(0.00)	(35.12)	(23.20)	(0.00)
Manufacture	Large	21.74	18.72	3.01	15.38	13.81	1.57
		(23.46)	(19.37)	(0.00)	(17.95)	(17.56)	(0.01)
	Mid	20.43	17.70	2.73	14.63	13.55	1.09
		(22.05)	(18.75)	(0.00)	(17.13)	(16.95)	(0.00)
	Small	20.35	17.68	2.68	14.81	13.33	1.48
		(22.20)	(18.43)	(0.00)	(17.86)	(16.41)	(0.00)
	Private	28.69	19.58	9.11	16.67	14.33	2.34
		(35.64)	(21.24)	(0.00)	(23.55)	(16.98)	(0.00)

Notes. The table summarizes churn metrics, reported as add rates by industry group and size class at the CPC Class level. NP and P columns show statistics for non-patenting and patenting firms, respectively, and Diff columns show the difference between them. Means and medians are computed from firm-year observations for the period 1997 to 2023. Standard deviations and interquartile ranges are reported below means and medians, respectively. P-values, reported below differences, are based on 1,000 bootstrap iterations under the null that both groups are drawn from the same distribution. In each iteration, we sample with replacement from the pooled data, compute the difference in means and medians, and report the p-value as the proportion of iterations where the absolute resampled difference exceeds or equals the observed value.

Table B22: Churn Metrics by Industry and Size: CPC Subclass, in Add Rates

Industry Group	Size Class	Mean Rate			Median Rate		
		NP	P	Diff	NP	P	Diff
Finance	Large	45.80	39.14	6.66	37.04	30.00	7.04
		(42.66)	(40.49)	(0.04)	(65.52)	(66.67)	(0.13)
	Mid	44.42	33.06	11.36	35.29	28.57	6.72
		(41.81)	(31.02)	(0.05)	(66.67)	(50.00)	(0.37)
	Small	40.54	27.57	12.98	32.65	12.50	20.15
		(38.49)	(32.37)	(0.01)	(66.67)	(50.00)	(0.00)
	Private	46.29	55.45	−9.16	35.90	47.21	−11.32
		(43.43)	(56.61)	(0.13)	(62.91)	(81.35)	(0.16)
Service	Large	36.80	29.41	7.40	24.62	20.00	4.62
		(37.26)	(32.00)	(0.00)	(39.19)	(32.59)	(0.00)
	Mid	37.04	30.72	6.32	25.86	22.22	3.64
		(35.96)	(32.15)	(0.00)	(38.87)	(32.97)	(0.00)
	Small	36.35	27.95	8.40	25.00	20.00	5.00
		(36.12)	(28.58)	(0.00)	(39.02)	(28.98)	(0.00)
	Private	45.08	33.44	11.63	31.25	24.12	7.13
		(42.54)	(32.16)	(0.00)	(48.57)	(34.92)	(0.00)
Resource	Large	31.12	36.69	−5.57	23.50	26.32	−2.82
		(27.34)	(30.70)	(0.02)	(23.58)	(30.50)	(0.11)
	Mid	35.72	35.55	0.17	25.35	24.50	0.85
		(34.44)	(33.27)	(0.94)	(26.79)	(34.86)	(0.65)
	Small	37.42	34.94	2.49	28.40	26.57	1.83
		(33.28)	(30.68)	(0.18)	(29.61)	(27.88)	(0.19)
	Private	41.41	33.12	8.29	29.58	26.27	3.31
		(37.49)	(29.39)	(0.00)	(33.21)	(23.98)	(0.05)
Manufacture	Large	29.83	24.91	4.92	24.35	20.85	3.49
		(23.45)	(18.39)	(0.00)	(19.59)	(16.46)	(0.00)
	Mid	27.46	23.76	3.70	22.37	19.75	2.62
		(21.15)	(18.20)	(0.00)	(18.36)	(16.49)	(0.00)
	Small	26.56	23.02	3.54	21.54	19.35	2.18
		(21.51)	(17.04)	(0.00)	(17.46)	(15.56)	(0.00)
	Private	35.13	24.81	10.32	23.69	20.25	3.44
		(34.46)	(19.67)	(0.00)	(23.25)	(15.14)	(0.00)

Notes. The table summarizes churn metrics, reported as add rates by industry group and size class at the CPC Subclass level. NP and P columns show statistics for non-patenting and patenting firms, respectively, and Diff columns show the difference between them. Means and medians are computed from firm-year observations for the period 1997 to 2023. Standard deviations and interquartile ranges are reported below means and medians, respectively. P-values, reported below differences, are based on 1,000 bootstrap iterations under the null that both groups are drawn from the same distribution. In each iteration, we sample with replacement from the pooled data, compute the difference in means and medians, and report the p-value as the proportion of iterations where the absolute resampled difference exceeds or equals the observed value.

Table B23: Churn Metrics by Industry and Size: CPC Group, in Add Rates

Industry Group	Size Class	Mean Rate			Median Rate		
		NP	P	Diff	NP	P	Diff
Finance	Large	55.55	46.17	9.38	51.06	36.84	14.22
		(39.21)	(35.60)	(0.00)	(56.83)	(46.12)	(0.00)
	Mid	55.42	44.13	11.29	50.00	44.64	5.36
		(37.89)	(28.90)	(0.04)	(54.14)	(37.08)	(0.43)
	Small	51.45	41.78	9.67	45.90	36.67	9.23
		(34.64)	(34.06)	(0.01)	(46.66)	(41.59)	(0.08)
	Private	55.83	60.22	−4.39	49.08	56.07	−6.99
		(38.31)	(51.91)	(0.42)	(53.91)	(63.44)	(0.32)
Service	Large	46.88	34.77	12.11	36.43	25.45	10.97
		(37.35)	(31.36)	(0.00)	(43.62)	(30.54)	(0.00)
	Mid	46.58	35.31	11.27	37.01	26.32	10.69
		(35.27)	(30.38)	(0.00)	(41.50)	(30.65)	(0.00)
	Small	45.33	33.74	11.59	35.76	26.22	9.54
		(35.23)	(27.48)	(0.00)	(40.04)	(26.59)	(0.00)
	Private	53.38	39.16	14.22	42.11	30.39	11.71
		(40.58)	(30.10)	(0.00)	(48.42)	(32.19)	(0.00)
Resource	Large	47.57	50.49	−2.92	42.01	43.56	−1.55
		(26.53)	(28.21)	(0.21)	(25.45)	(31.13)	(0.46)
	Mid	51.17	50.40	0.77	43.18	42.92	0.26
		(33.54)	(34.14)	(0.76)	(30.28)	(33.86)	(0.90)
	Small	53.01	48.40	4.61	45.99	41.79	4.20
		(32.66)	(29.11)	(0.02)	(32.06)	(31.99)	(0.01)
	Private	56.59	48.21	8.37	47.95	41.97	5.98
		(35.65)	(27.68)	(0.00)	(34.46)	(24.88)	(0.00)
Manufacture	Large	43.11	35.96	7.15	38.10	32.29	5.81
		(23.89)	(19.22)	(0.00)	(21.27)	(18.88)	(0.00)
	Mid	41.39	34.83	6.56	36.97	31.32	5.64
		(21.96)	(18.61)	(0.00)	(21.03)	(18.58)	(0.00)
	Small	39.24	33.60	5.63	34.98	29.97	5.01
		(22.00)	(17.64)	(0.00)	(20.40)	(17.61)	(0.00)
	Private	47.11	35.34	11.77	37.01	31.14	5.87
		(33.24)	(19.23)	(0.00)	(25.83)	(17.10)	(0.00)

Notes. The table summarizes churn metrics, reported as add rates by industry group and size class at the CPC Group level. NP and P columns show statistics for non-patenting and patenting firms, respectively, and Diff columns show the difference between them. Means and medians are computed from firm-year observations for the period 1997 to 2023. Standard deviations and interquartile ranges are reported below means and medians, respectively. P-values, reported below differences, are based on 1,000 bootstrap iterations under the null that both groups are drawn from the same distribution. In each iteration, we sample with replacement from the pooled data, compute the difference in means and medians, and report the p-value as the proportion of iterations where the absolute resampled difference exceeds or equals the observed value.

Table B24: Churn Metrics by Industry and Size: CPC Section, in Drop Rates

Industry Group	Size Class	Mean Rate			Median Rate		
		NP	P	Diff	NP	P	Diff
Finance	Large	14.28	14.27	0.01	0.00	0.00	0.00
		(35.35)	(35.18)	(1.00)	(0.00)	(0.00)	(1.00)
	Mid	12.12	17.95	−5.83	0.00	0.00	0.00
		(32.72)	(38.18)	(0.21)	(0.00)	(0.00)	(1.00)
	Small	10.45	8.74	1.71	0.00	0.00	0.00
		(30.49)	(27.87)	(0.64)	(0.00)	(0.00)	(1.00)
	Private	12.64	29.03	−16.39	0.00	0.00	0.00
		(33.39)	(45.08)	(0.00)	(0.00)	(54.17)	(1.00)
Service	Large	29.81	20.26	9.55	0.00	0.00	0.00
		(45.35)	(38.97)	(0.00)	(66.67)	(0.00)	(1.00)
	Mid	28.50	23.30	5.20	0.00	0.00	0.00
		(44.43)	(41.80)	(0.00)	(66.67)	(40.00)	(1.00)
	Small	28.53	20.89	7.64	0.00	0.00	0.00
		(44.77)	(39.20)	(0.00)	(66.67)	(0.00)	(1.00)
	Private	29.47	21.07	8.40	0.00	0.00	0.00
		(44.76)	(39.31)	(0.00)	(66.67)	(25.00)	(1.00)
Resource	Large	19.05	18.23	0.82	0.00	0.00	0.00
		(33.99)	(36.29)	(0.80)	(28.57)	(22.22)	(1.00)
	Mid	20.23	19.72	0.51	0.00	0.00	0.00
		(36.49)	(35.28)	(0.87)	(28.57)	(38.33)	(1.00)
	Small	20.51	21.89	−1.38	0.00	0.00	0.00
		(36.70)	(38.03)	(0.54)	(28.57)	(28.57)	(1.00)
	Private	22.00	23.23	−1.24	0.00	0.00	0.00
		(37.79)	(35.29)	(0.63)	(34.09)	(40.00)	(1.00)
Manufacture	Large	13.96	11.98	1.98	0.00	0.00	0.00
		(27.93)	(25.53)	(0.04)	(18.18)	(15.38)	(1.00)
	Mid	13.46	10.57	2.89	0.00	0.00	0.00
		(26.02)	(23.14)	(0.00)	(18.18)	(13.33)	(1.00)
	Small	11.35	10.05	1.31	0.00	0.00	0.00
		(24.03)	(22.04)	(0.00)	(15.38)	(0.00)	(1.00)
	Private	15.43	9.29	6.14	0.00	0.00	0.00
		(28.82)	(19.11)	(0.00)	(22.22)	(14.29)	(1.00)

Notes. The table summarizes churn metrics, reported as drop rates by industry group and size class at the CPC Section level. NP and P columns show statistics for non-patenting and patenting firms, respectively, and Diff columns show the difference between them. Means and medians are computed from firm-year observations for the period 1997 to 2023. Standard deviations and interquartile ranges are reported below means and medians, respectively. P-values, reported below differences, are based on 1,000 bootstrap iterations under the null that both groups are drawn from the same distribution. In each iteration, we sample with replacement from the pooled data, compute the difference in means and medians, and report the p-value as the proportion of iterations where the absolute resampled difference exceeds or equals the observed value.

Table B25: Churn Metrics by Industry and Size: CPC Class, in Drop Rates

Industry Group	Size Class	Mean Rate			Median Rate		
		NP	P	Diff	NP	P	Diff
Finance	Large	38.01	35.38	2.64	2.41	0.00	2.41
		(47.85)	(47.53)	(0.47)	(66.67)	(66.67)	(0.65)
	Mid	36.00	34.39	1.62	0.00	0.00	0.00
		(46.35)	(44.20)	(0.81)	(66.67)	(66.67)	(1.00)
	Small	37.52	37.09	0.43	0.00	13.33	−13.33
		(46.08)	(44.51)	(0.94)	(66.67)	(66.67)	(0.40)
	Private	40.01	42.89	−2.88	16.00	22.02	−6.02
		(47.86)	(48.83)	(0.67)	(66.67)	(75.00)	(0.76)
Service	Large	33.92	26.24	7.68	20.00	15.38	4.62
		(41.34)	(34.77)	(0.00)	(50.00)	(40.00)	(0.01)
	Mid	32.92	26.62	6.30	19.05	10.53	8.52
		(40.39)	(36.77)	(0.00)	(50.00)	(40.00)	(0.00)
	Small	33.22	25.77	7.45	20.00	13.33	6.67
		(40.43)	(34.46)	(0.00)	(50.00)	(40.00)	(0.00)
	Private	37.70	28.90	8.79	22.22	16.67	5.56
		(43.95)	(35.91)	(0.00)	(57.14)	(43.55)	(0.00)
Resource	Large	24.75	27.10	−2.35	15.38	16.36	−0.98
		(29.21)	(31.86)	(0.37)	(27.02)	(28.03)	(0.61)
	Mid	29.78	29.65	0.13	17.93	21.05	−3.12
		(34.85)	(32.33)	(0.96)	(32.05)	(30.66)	(0.17)
	Small	28.82	27.50	1.32	17.14	14.81	2.33
		(33.76)	(31.71)	(0.49)	(30.60)	(30.52)	(0.11)
	Private	29.29	25.79	3.50	17.39	17.93	−0.54
		(34.61)	(26.02)	(0.11)	(31.82)	(27.26)	(0.70)
Manufacture	Large	23.95	21.86	2.09	15.58	15.38	0.20
		(26.89)	(23.02)	(0.02)	(20.64)	(20.83)	(0.73)
	Mid	22.38	20.61	1.77	15.09	14.29	0.81
		(23.93)	(23.07)	(0.00)	(19.66)	(18.82)	(0.02)
	Small	21.44	20.25	1.19	14.55	14.63	−0.09
		(23.74)	(21.36)	(0.00)	(19.31)	(18.68)	(0.73)
	Private	26.11	19.79	6.32	16.44	15.00	1.44
		(29.93)	(19.07)	(0.00)	(23.70)	(17.95)	(0.00)

Notes. The table summarizes churn metrics, reported as drop rates by industry group and size class at the CPC Class level. NP and P columns show statistics for non-patenting and patenting firms, respectively, and Diff columns show the difference between them. Means and medians are computed from firm-year observations for the period 1997 to 2023. Standard deviations and interquartile ranges are reported below means and medians, respectively. P-values, reported below differences, are based on 1,000 bootstrap iterations under the null that both groups are drawn from the same distribution. In each iteration, we sample with replacement from the pooled data, compute the difference in means and medians, and report the p-value as the proportion of iterations where the absolute resampled difference exceeds or equals the observed value.

Table B26: Churn Metrics by Industry and Size: CPC Subclass, in Drop Rates

Industry Group	Size Class	Mean Rate			Median Rate		
		NP	P	Diff	NP	P	Diff
Finance	Large	43.41	40.43	2.98	31.58	28.57	3.01
		(44.12)	(43.42)	(0.37)	(66.67)	(64.43)	(0.49)
	Mid	40.72	34.23	6.48	28.57	24.26	4.31
		(43.24)	(35.46)	(0.28)	(66.67)	(59.52)	(0.57)
	Small	38.14	29.97	8.17	28.57	14.46	14.11
		(40.22)	(34.72)	(0.07)	(60.00)	(57.78)	(0.01)
	Private	43.68	49.25	−5.57	30.77	31.97	−1.20
		(44.17)	(50.26)	(0.35)	(60.54)	(84.49)	(0.89)
Service	Large	39.61	33.44	6.17	28.57	23.53	5.04
		(38.41)	(33.85)	(0.00)	(43.26)	(37.09)	(0.00)
	Mid	39.10	34.95	4.15	27.62	25.00	2.62
		(37.93)	(34.99)	(0.00)	(41.21)	(40.00)	(0.04)
	Small	39.48	32.43	7.04	28.57	23.53	5.04
		(37.69)	(31.87)	(0.00)	(42.05)	(33.63)	(0.00)
	Private	44.05	35.11	8.93	30.77	25.00	5.77
		(41.84)	(32.76)	(0.00)	(46.95)	(36.71)	(0.00)
Resource	Large	33.11	36.04	−2.92	25.12	26.95	−1.83
		(29.16)	(31.13)	(0.25)	(27.70)	(28.93)	(0.37)
	Mid	36.74	37.90	−1.17	27.05	28.42	−1.38
		(33.46)	(32.95)	(0.66)	(28.72)	(30.31)	(0.45)
	Small	36.09	34.34	1.75	26.34	23.64	2.70
		(32.50)	(30.01)	(0.34)	(29.80)	(30.57)	(0.08)
	Private	36.20	33.48	2.72	25.81	28.77	−2.96
		(33.01)	(25.40)	(0.19)	(31.00)	(24.24)	(0.08)
Manufacture	Large	32.18	28.39	3.79	23.84	22.64	1.20
		(26.70)	(21.97)	(0.00)	(22.47)	(19.39)	(0.05)
	Mid	29.82	26.86	2.95	23.64	21.18	2.46
		(22.79)	(21.50)	(0.00)	(19.85)	(17.60)	(0.00)
	Small	27.94	25.98	1.95	21.92	21.05	0.87
		(22.56)	(19.92)	(0.00)	(18.33)	(17.29)	(0.00)
	Private	32.73	25.22	7.51	23.93	21.28	2.66
		(29.10)	(17.31)	(0.00)	(23.25)	(16.05)	(0.00)

Notes. The table summarizes churn metrics, reported as drop rates by industry group and size class at the CPC Subclass level. NP and P columns show statistics for non-patenting and patenting firms, respectively, and Diff columns show the difference between them. Means and medians are computed from firm-year observations for the period 1997 to 2023. Standard deviations and interquartile ranges are reported below means and medians, respectively. P-values, reported below differences, are based on 1,000 bootstrap iterations under the null that both groups are drawn from the same distribution. In each iteration, we sample with replacement from the pooled data, compute the difference in means and medians, and report the p-value as the proportion of iterations where the absolute resampled difference exceeds or equals the observed value.

Table B27: Churn Metrics by Industry and Size: CPC Group, in Drop Rates

Industry Group	Size Class	Mean Rate			Median Rate		
		NP	P	Diff	NP	P	Diff
Finance	Large	53.10	48.39	4.71	43.90	40.00	3.90
		(40.37)	(39.92)	(0.12)	(51.10)	(38.71)	(0.33)
	Mid	51.71	45.31	6.40	42.11	36.50	5.61
		(39.67)	(36.02)	(0.22)	(50.51)	(51.72)	(0.38)
	Small	49.04	44.96	4.08	40.00	38.10	1.90
		(36.41)	(34.65)	(0.31)	(46.40)	(46.11)	(0.70)
	Private	53.27	55.50	−2.22	43.48	44.44	−0.97
		(40.02)	(42.15)	(0.69)	(51.06)	(67.28)	(0.88)
Service	Large	49.57	38.60	10.98	39.20	30.86	8.34
		(38.25)	(31.59)	(0.00)	(44.69)	(32.69)	(0.00)
	Mid	48.88	39.39	9.50	39.43	29.37	10.06
		(36.85)	(33.07)	(0.00)	(42.48)	(34.80)	(0.00)
	Small	48.49	37.98	10.52	38.67	29.24	9.43
		(36.50)	(30.29)	(0.00)	(42.79)	(29.94)	(0.00)
	Private	52.49	40.92	11.57	41.38	32.61	8.77
		(40.02)	(30.74)	(0.00)	(46.78)	(33.55)	(0.00)
Resource	Large	50.45	51.33	−0.88	45.21	45.42	−0.21
		(27.98)	(29.38)	(0.70)	(27.46)	(31.69)	(0.90)
	Mid	53.09	52.97	0.12	45.47	45.92	−0.45
		(32.91)	(33.25)	(0.96)	(30.93)	(36.07)	(0.83)
	Small	52.03	48.06	3.97	44.63	39.94	4.69
		(31.92)	(29.24)	(0.03)	(32.20)	(31.76)	(0.01)
	Private	51.11	50.36	0.75	44.04	47.45	−3.40
		(32.13)	(24.99)	(0.72)	(32.46)	(29.89)	(0.05)
Manufacture	Large	46.61	40.33	6.28	38.95	35.34	3.61
		(26.64)	(21.64)	(0.00)	(23.49)	(20.11)	(0.00)
	Mid	44.81	38.64	6.18	39.13	33.55	5.58
		(23.20)	(21.24)	(0.00)	(22.42)	(18.58)	(0.00)
	Small	41.59	37.43	4.16	36.21	33.00	3.21
		(22.51)	(19.69)	(0.00)	(20.21)	(18.07)	(0.00)
	Private	44.74	36.81	7.94	37.76	33.21	4.56
		(27.81)	(17.48)	(0.00)	(24.53)	(16.87)	(0.00)

Notes. The table summarizes churn metrics, reported as drop rates by industry group and size class at the CPC Group level. NP and P columns show statistics for non-patenting and patenting firms, respectively, and Diff columns show the difference between them. Means and medians are computed from firm-year observations for the period 1997 to 2023. Standard deviations and interquartile ranges are reported below means and medians, respectively. P-values, reported below differences, are based on 1,000 bootstrap iterations under the null that both groups are drawn from the same distribution. In each iteration, we sample with replacement from the pooled data, compute the difference in means and medians, and report the p-value as the proportion of iterations where the absolute resampled difference exceeds or equals the observed value.

B.4 Technological Generality

Table B28: Cross Metrics by Industry: CPC Section, in Counts and Fractions

(a) Count Statistics for Non-Patenting (NP) and Patenting (P) Firms

Industry Group	Mean Count			Median Count		
	NP	P	Diff	NP	P	Diff
Finance	9.68	15.28	−5.60	0.00	0.00	0.00
	(23.10)	(25.75)	(0.00)	(0.00)	(32.50)	(1.00)
Service	41.56	48.70	−7.14	52.00	58.67	−6.67
	(29.81)	(24.55)	(0.00)	(65.00)	(24.67)	(0.00)
Resource	46.49	46.59	−0.10	47.00	47.33	−0.33
	(18.73)	(18.12)	(0.85)	(18.00)	(18.50)	(0.52)
Manufacture	50.76	51.01	−0.25	50.00	52.20	−2.20
	(18.44)	(16.89)	(0.08)	(24.00)	(23.89)	(0.00)

(b) Fraction Statistics for Non-Patenting (NP) and Patenting (P) Firms

Industry Group	Mean Fraction			Median Fraction		
	NP	P	Diff	NP	P	Diff
Finance	2.32	3.66	−1.34	0.00	0.00	0.00
	(5.52)	(6.17)	(0.00)	(0.00)	(7.58)	(1.00)
Service	9.70	11.45	−1.75	12.24	13.69	−1.45
	(6.90)	(5.74)	(0.00)	(15.16)	(5.81)	(0.00)
Resource	10.85	10.89	−0.03	11.07	11.08	−0.01
	(4.23)	(4.13)	(0.81)	(3.98)	(4.01)	(0.88)
Manufacture	11.82	11.87	−0.06	11.73	12.16	−0.44
	(4.10)	(3.77)	(0.10)	(5.32)	(5.40)	(0.00)

Notes. The table summarizes cross metrics, reported as counts and fractions by industry group at the CPC Section level. NP and P columns show statistics for non-patenting and patenting firms, respectively, and Diff columns show the difference between them. Means and medians are computed from firm-year observations for the period 1997 to 2023. Panels B28a and B28b report count and fraction statistics, respectively. Standard deviations and interquartile ranges are reported below means and medians, respectively. P-values, reported below differences, are based on 1,000 bootstrap iterations under the null that values for both firm types are drawn from the same distribution. In each iteration, we sample with replacement from the pooled data, compute the difference in means and medians, and report the p-value as $(c+1)/(n+1)$, where c is the count of iterations where the absolute resampled difference exceeds or equals the observed value, and n is the total number of valid bootstrap samples.

Table B29: Cross Metrics by Industry: CPC Class, in Counts and Fractions

(a) Count Statistics for Non-Patenting (NP) and Patenting (P) Firms

Industry Group	Mean Count			Median Count		
	NP	P	Diff	NP	P	Diff
Finance	60.75	58.74	2.01	70.00	65.94	4.06
	(33.53)	(35.80)	(0.21)	(39.94)	(47.67)	(0.03)
Service	60.61	66.94	−6.32	62.00	68.67	−6.67
	(20.97)	(18.12)	(0.00)	(28.00)	(22.38)	(0.00)
Resource	46.93	47.71	−0.78	44.80	46.20	−1.40
	(13.01)	(11.66)	(0.05)	(16.25)	(13.91)	(0.00)
Manufacture	50.74	52.44	−1.69	47.89	51.06	−3.16
	(14.93)	(14.36)	(0.00)	(22.89)	(21.68)	(0.00)

(b) Fraction Statistics for Non-Patenting (NP) and Patenting (P) Firms

Industry Group	Mean Fraction			Median Fraction		
	NP	P	Diff	NP	P	Diff
Finance	14.14	13.88	0.26	16.38	15.47	0.91
	(7.70)	(8.42)	(0.45)	(8.35)	(11.41)	(0.01)
Service	14.10	15.69	−1.59	14.37	16.28	−1.91
	(4.74)	(4.16)	(0.00)	(6.16)	(4.93)	(0.00)
Resource	10.94	11.13	−0.19	10.46	10.77	−0.31
	(2.81)	(2.49)	(0.03)	(3.29)	(2.95)	(0.00)
Manufacture	11.81	12.20	−0.39	11.17	11.85	−0.68
	(3.19)	(3.08)	(0.00)	(4.80)	(4.63)	(0.00)

Notes. The table summarizes cross metrics, reported as counts and fractions by industry group at the CPC Class level. NP and P columns show statistics for non-patenting and patenting firms, respectively, and Diff columns show the difference between them. Means and medians are computed from firm-year observations for the period 1997 to 2023. Panels B29a and B29b report count and fraction statistics, respectively. Standard deviations and interquartile ranges are reported below means and medians, respectively. P-values, reported below differences, are based on 1,000 bootstrap iterations under the null that values for both firm types are drawn from the same distribution. In each iteration, we sample with replacement from the pooled data, compute the difference in means and medians, and report the p-value as $(c+1)/(n+1)$, where c is the count of iterations where the absolute resampled difference exceeds or equals the observed value, and n is the total number of valid bootstrap samples.

Table B30: Cross Metrics by Industry: CPC Subclass, in Counts and Fractions

(a) Count Statistics for Non-Patenting (NP) and Patenting (P) Firms

Industry Group	Mean Count			Median Count		
	NP	P	Diff	NP	P	Diff
Finance	67.31	72.35	−5.04	65.83	69.98	−4.14
	(26.44)	(30.57)	(0.00)	(28.54)	(32.00)	(0.00)
Service	58.31	61.11	−2.80	57.60	61.75	−4.15
	(18.23)	(15.57)	(0.00)	(23.63)	(19.42)	(0.00)
Resource	48.08	48.11	−0.03	45.63	46.25	−0.62
	(13.41)	(12.42)	(0.94)	(16.39)	(12.84)	(0.12)
Manufacture	49.89	50.44	−0.55	48.64	50.25	−1.61
	(14.37)	(13.63)	(0.00)	(20.95)	(19.03)	(0.00)

(b) Fraction Statistics for Non-Patenting (NP) and Patenting (P) Firms

Industry Group	Mean Fraction			Median Fraction		
	NP	P	Diff	NP	P	Diff
Finance	15.67	17.06	−1.39	15.27	16.37	−1.10
	(6.06)	(7.11)	(0.00)	(6.15)	(7.40)	(0.00)
Service	13.56	14.32	−0.76	13.34	14.54	−1.20
	(4.05)	(3.50)	(0.00)	(5.21)	(4.47)	(0.00)
Resource	11.20	11.20	−0.00	10.63	10.77	−0.14
	(2.85)	(2.60)	(0.96)	(3.35)	(2.67)	(0.11)
Manufacture	11.61	11.72	−0.12	11.32	11.65	−0.33
	(3.03)	(2.87)	(0.00)	(4.34)	(4.02)	(0.00)

Notes. The table summarizes cross metrics, reported as counts and fractions by industry group at the CPC Subclass level. NP and P columns show statistics for non-patenting and patenting firms, respectively, and Diff columns show the difference between them. Means and medians are computed from firm-year observations for the period 1997 to 2023. Panels B30a and B30b report count and fraction statistics, respectively. Standard deviations and interquartile ranges are reported below means and medians, respectively. P-values, reported below differences, are based on 1,000 bootstrap iterations under the null that values for both firm types are drawn from the same distribution. In each iteration, we sample with replacement from the pooled data, compute the difference in means and medians, and report the p-value as $(c+1)/(n+1)$, where c is the count of iterations where the absolute resampled difference exceeds or equals the observed value, and n is the total number of valid bootstrap samples.

Table B31: Cross Metrics by Industry: CPC Group, in Counts and Fractions

(a) Count Statistics for Non-Patenting (NP) and Patenting (P) Firms

Industry Group	Mean Count			Median Count		
	NP	P	Diff	NP	P	Diff
Finance	54.76	60.15	−5.40	51.90	58.30	−6.40
	(16.03)	(20.14)	(0.00)	(17.74)	(25.03)	(0.00)
Service	50.59	52.78	−2.20	49.19	52.74	−3.55
	(15.40)	(13.74)	(0.00)	(20.29)	(18.07)	(0.00)
Resource	42.82	42.35	0.47	41.27	41.61	−0.33
	(10.77)	(11.11)	(0.17)	(13.11)	(13.31)	(0.40)
Manufacture	41.78	41.12	0.66	41.27	41.03	0.24
	(11.70)	(11.31)	(0.00)	(16.42)	(15.24)	(0.02)

(b) Fraction Statistics for Non-Patenting (NP) and Patenting (P) Firms

Industry Group	Mean Fraction			Median Fraction		
	NP	P	Diff	NP	P	Diff
Finance	12.75	14.20	−1.45	12.05	13.60	−1.56
	(3.64)	(4.68)	(0.00)	(3.85)	(5.78)	(0.00)
Service	11.77	12.37	−0.60	11.39	12.42	−1.03
	(3.44)	(3.11)	(0.00)	(4.54)	(4.26)	(0.00)
Resource	9.98	9.86	0.12	9.66	9.64	0.02
	(2.28)	(2.34)	(0.10)	(2.68)	(2.80)	(0.85)
Manufacture	9.73	9.56	0.17	9.63	9.52	0.11
	(2.49)	(2.41)	(0.00)	(3.42)	(3.24)	(0.00)

Notes. The table summarizes cross metrics, reported as counts and fractions by industry group at the CPC Group level. NP and P columns show statistics for non-patenting and patenting firms, respectively, and Diff columns show the difference between them. Means and medians are computed from firm-year observations for the period 1997 to 2023. Panels B31a and B31b report count and fraction statistics, respectively. Standard deviations and interquartile ranges are reported below means and medians, respectively. P-values, reported below differences, are based on 1,000 bootstrap iterations under the null that values for both firm types are drawn from the same distribution. In each iteration, we sample with replacement from the pooled data, compute the difference in means and medians, and report the p-value as $(c+1)/(n+1)$, where c is the count of iterations where the absolute resampled difference exceeds or equals the observed value, and n is the total number of valid bootstrap samples.

Table B32: Cross Metrics by Industry: CPC Patent, in Counts and Fractions

(a) Count Statistics for Non-Patenting (NP) and Patenting (P) Firms

Industry Group	Mean Count			Median Count		
	NP	P	Diff	NP	P	Diff
Finance	49.21	50.56	−1.35	47.36	48.54	−1.19
	(15.68)	(19.53)	(0.06)	(19.97)	(25.66)	(0.19)
Service	45.03	39.27	5.76	43.49	36.97	6.53
	(15.86)	(14.61)	(0.00)	(22.00)	(19.45)	(0.00)
Resource	46.42	44.32	2.10	46.98	44.92	2.07
	(11.23)	(12.86)	(0.00)	(14.19)	(17.59)	(0.00)
Manufacture	37.48	32.49	4.99	37.10	30.40	6.69
	(12.90)	(12.38)	(0.00)	(19.04)	(17.46)	(0.00)

(b) Fraction Statistics for Non-Patenting (NP) and Patenting (P) Firms

Industry Group	Mean Fraction			Median Fraction		
	NP	P	Diff	NP	P	Diff
Finance	11.48	11.94	−0.46	10.94	11.54	−0.59
	(3.58)	(4.50)	(0.00)	(4.37)	(5.69)	(0.00)
Service	10.49	9.20	1.29	10.14	8.76	1.39
	(3.54)	(3.28)	(0.00)	(4.78)	(4.30)	(0.00)
Resource	10.86	10.35	0.51	10.97	10.52	0.45
	(2.43)	(2.83)	(0.00)	(2.99)	(3.76)	(0.00)
Manufacture	8.77	7.58	1.19	8.70	7.08	1.61
	(2.88)	(2.77)	(0.00)	(4.32)	(3.87)	(0.00)

Notes. The table summarizes cross metrics, reported as counts and fractions by industry group at the CPC Patent level. NP and P columns show statistics for non-patenting and patenting firms, respectively, and Diff columns show the difference between them. Means and medians are computed from firm-year observations for the period 1997 to 2023. Panels B32a and B32b report count and fraction statistics, respectively. Standard deviations and interquartile ranges are reported below means and medians, respectively. P-values, reported below differences, are based on 1,000 bootstrap iterations under the null that values for both firm types are drawn from the same distribution. In each iteration, we sample with replacement from the pooled data, compute the difference in means and medians, and report the p-value as $(c+1)/(n+1)$, where c is the count of iterations where the absolute resampled difference exceeds or equals the observed value, and n is the total number of valid bootstrap samples.

Table B33: Cross Metrics by Industry and Size: CPC Section, in Counts

Industry Group	Size Class	Mean Rate			Median Rate		
		NP	P	Diff	NP	P	Diff
Finance	Large	9.22	10.91	−1.69	0.00	0.00	0.00
		(22.37)	(22.51)	(0.30)	(0.00)	(0.00)	(1.00)
	Mid	10.39	21.64	−11.25	0.00	0.00	0.00
		(23.84)	(31.19)	(0.00)	(0.00)	(59.50)	(1.00)
	Small	9.09	18.21	−9.12	0.00	0.00	0.00
		(22.64)	(26.55)	(0.00)	(0.00)	(44.96)	(1.00)
	Private	10.28	21.47	−11.19	0.00	0.00	0.00
		(23.54)	(27.88)	(0.00)	(0.00)	(48.05)	(1.00)
Service	Large	42.22	48.46	−6.24	52.00	58.50	−6.50
		(30.17)	(24.86)	(0.00)	(64.50)	(24.27)	(0.00)
	Mid	41.10	48.58	−7.48	50.60	59.00	−8.40
		(30.41)	(25.30)	(0.00)	(65.00)	(24.46)	(0.00)
	Small	41.91	48.97	−7.05	52.00	58.50	−6.50
		(30.33)	(23.89)	(0.00)	(65.00)	(24.67)	(0.00)
	Private	41.29	48.41	−7.11	51.50	58.50	−7.00
		(29.09)	(25.02)	(0.00)	(64.00)	(24.94)	(0.00)
Resource	Large	50.06	46.99	3.07	49.00	47.33	1.67
		(18.17)	(17.72)	(0.03)	(19.50)	(13.80)	(0.17)
	Mid	46.04	48.02	−1.98	45.25	47.33	−2.08
		(19.63)	(17.75)	(0.20)	(19.00)	(17.50)	(0.15)
	Small	45.81	43.47	2.33	45.33	46.00	−0.67
		(19.91)	(18.32)	(0.03)	(19.50)	(18.25)	(0.69)
	Private	46.56	49.42	−2.85	47.33	51.00	−3.67
		(18.11)	(17.87)	(0.01)	(16.40)	(21.00)	(0.00)
Manufacture	Large	50.10	52.60	−2.50	46.71	53.50	−6.79
		(21.22)	(17.81)	(0.00)	(27.67)	(23.67)	(0.00)
	Mid	51.89	52.64	−0.75	49.33	54.12	−4.79
		(18.61)	(16.85)	(0.05)	(24.62)	(22.75)	(0.00)
	Small	51.32	50.30	1.02	50.50	52.00	−1.50
		(19.24)	(16.99)	(0.00)	(25.71)	(25.00)	(0.00)
	Private	49.92	49.93	−0.01	49.60	49.50	0.10
		(17.20)	(15.39)	(0.97)	(20.62)	(21.06)	(0.86)

Notes. The table summarizes cross metrics, reported as counts by industry group and size class at the CPC Section level. NP and P columns show statistics for non-patenting and patenting firms, respectively, and Diff columns show the difference between them. Means and medians are computed from firm-year observations for the period 1997 to 2023. Standard deviations and interquartile ranges are reported below means and medians, respectively. P-values, reported below differences, are based on 1,000 bootstrap iterations under the null that both groups are drawn from the same distribution. In each iteration, we sample with replacement from the pooled data, compute the difference in means and medians, and report the p-value as the proportion of iterations where the absolute resampled difference exceeds or equals the observed value.

Table B34: Cross Metrics by Industry and Size: CPC Class, in Counts

Industry Group	Size Class	Mean Rate			Median Rate		
		NP	P	Diff	NP	P	Diff
Finance	Large	57.06	58.76	−1.70	68.45	65.67	2.78
		(36.13)	(36.24)	(0.50)	(51.00)	(49.50)	(0.40)
	Mid	56.00	62.03	−6.03	65.33	73.30	−7.97
		(36.05)	(36.21)	(0.21)	(54.28)	(50.15)	(0.14)
	Small	64.87	64.91	−0.05	72.50	69.40	3.10
		(31.82)	(35.36)	(0.99)	(34.00)	(34.40)	(0.44)
	Private	57.58	47.87	9.70	66.77	47.63	19.13
		(33.85)	(32.58)	(0.03)	(41.40)	(34.42)	(0.00)
Service	Large	61.99	66.97	−4.99	62.18	69.18	−7.00
		(19.64)	(18.59)	(0.00)	(26.67)	(23.67)	(0.00)
	Mid	61.61	68.89	−7.28	62.33	71.45	−9.12
		(21.21)	(18.80)	(0.00)	(29.17)	(19.00)	(0.00)
	Small	62.78	66.65	−3.87	64.62	68.50	−3.88
		(20.30)	(17.88)	(0.00)	(27.64)	(22.44)	(0.00)
	Private	58.13	65.49	−7.36	59.29	67.11	−7.83
		(21.40)	(17.31)	(0.00)	(27.87)	(21.25)	(0.00)
Resource	Large	48.15	46.29	1.86	46.37	45.77	0.59
		(12.06)	(11.89)	(0.05)	(17.30)	(10.15)	(0.43)
	Mid	46.99	47.09	−0.10	45.68	45.23	0.45
		(13.12)	(11.04)	(0.93)	(13.26)	(11.25)	(0.47)
	Small	46.79	46.11	0.68	44.56	43.90	0.66
		(13.80)	(11.06)	(0.35)	(19.53)	(14.15)	(0.28)
	Private	46.89	51.24	−4.35	44.62	49.15	−4.53
		(12.74)	(12.00)	(0.00)	(14.98)	(15.47)	(0.00)
Manufacture	Large	50.87	54.05	−3.18	47.26	52.51	−5.25
		(16.08)	(14.65)	(0.00)	(25.08)	(22.52)	(0.00)
	Mid	51.16	53.74	−2.58	47.35	52.69	−5.34
		(14.87)	(14.85)	(0.00)	(24.01)	(22.35)	(0.00)
	Small	51.30	51.66	−0.37	48.81	50.52	−1.71
		(15.51)	(14.33)	(0.03)	(25.02)	(21.53)	(0.00)
	Private	50.03	52.02	−1.99	47.28	50.12	−2.85
		(14.16)	(13.26)	(0.00)	(19.39)	(20.20)	(0.00)

Notes. The table summarizes cross metrics, reported as counts by industry group and size class at the CPC Class level. NP and P columns show statistics for non-patenting and patenting firms, respectively, and Diff columns show the difference between them. Means and medians are computed from firm-year observations for the period 1997 to 2023. Standard deviations and interquartile ranges are reported below means and medians, respectively. P-values, reported below differences, are based on 1,000 bootstrap iterations under the null that both groups are drawn from the same distribution. In each iteration, we sample with replacement from the pooled data, compute the difference in means and medians, and report the p-value as the proportion of iterations where the absolute resampled difference exceeds or equals the observed value.

Table B35: Cross Metrics by Industry and Size: CPC Subclass, in Counts

Industry Group	Size Class	Mean Rate			Median Rate		
		NP	P	Diff	NP	P	Diff
Finance	Large	71.12	70.71	0.41	69.00	69.20	−0.20
		(30.22)	(30.23)	(0.86)	(29.68)	(30.23)	(0.91)
	Mid	68.67	82.72	−14.05	66.44	81.12	−14.69
		(29.24)	(27.81)	(0.00)	(30.58)	(22.36)	(0.00)
	Small	66.67	77.31	−10.65	66.20	72.33	−6.12
		(23.77)	(29.70)	(0.00)	(27.67)	(33.62)	(0.05)
	Private	67.16	62.71	4.44	64.75	59.92	4.83
		(28.02)	(32.06)	(0.21)	(28.60)	(32.74)	(0.13)
Service	Large	58.73	61.45	−2.72	57.34	61.82	−4.48
		(17.19)	(16.39)	(0.00)	(22.52)	(19.97)	(0.00)
	Mid	59.12	62.60	−3.48	58.43	63.93	−5.51
		(18.45)	(16.06)	(0.00)	(24.28)	(20.14)	(0.00)
	Small	60.11	60.70	−0.58	60.11	61.43	−1.32
		(17.66)	(15.15)	(0.11)	(23.89)	(18.88)	(0.00)
	Private	56.32	60.18	−3.85	55.20	59.57	−4.37
		(18.62)	(15.11)	(0.00)	(22.41)	(19.42)	(0.00)
Resource	Large	48.66	47.49	1.17	46.72	46.06	0.66
		(12.42)	(11.60)	(0.25)	(18.11)	(10.14)	(0.41)
	Mid	48.19	49.27	−1.08	46.70	47.44	−0.74
		(13.27)	(12.78)	(0.28)	(14.08)	(15.31)	(0.47)
	Small	48.28	45.46	2.82	45.98	44.26	1.72
		(14.48)	(12.24)	(0.00)	(22.73)	(13.45)	(0.03)
	Private	47.95	51.21	−3.27	45.32	49.43	−4.11
		(13.07)	(12.19)	(0.00)	(14.40)	(16.80)	(0.00)
Manufacture	Large	51.18	51.92	−0.73	49.34	51.39	−2.05
		(15.14)	(13.35)	(0.11)	(24.49)	(20.12)	(0.00)
	Mid	51.24	51.71	−0.47	49.58	51.32	−1.74
		(14.41)	(13.58)	(0.12)	(23.20)	(19.60)	(0.00)
	Small	50.22	49.31	0.91	49.66	49.33	0.33
		(15.03)	(13.74)	(0.00)	(23.56)	(18.31)	(0.11)
	Private	49.11	51.49	−2.38	47.40	50.80	−3.40
		(13.49)	(13.18)	(0.00)	(17.48)	(18.86)	(0.00)

Notes. The table summarizes cross metrics, reported as counts by industry group and size class at the CPC Subclass level. NP and P columns show statistics for non-patenting and patenting firms, respectively, and Diff columns show the difference between them. Means and medians are computed from firm-year observations for the period 1997 to 2023. Standard deviations and interquartile ranges are reported below means and medians, respectively. P-values, reported below differences, are based on 1,000 bootstrap iterations under the null that both groups are drawn from the same distribution. In each iteration, we sample with replacement from the pooled data, compute the difference in means and medians, and report the p-value as the proportion of iterations where the absolute resampled difference exceeds or equals the observed value.

Table B36: Cross Metrics by Industry and Size: CPC Group, in Counts

Industry Group	Size Class	Mean Rate			Median Rate		
		NP	P	Diff	NP	P	Diff
Finance	Large	60.29	59.39	0.90	57.46	58.30	−0.84
		(18.73)	(18.92)	(0.49)	(21.99)	(21.75)	(0.56)
	Mid	57.20	66.18	−8.99	53.88	64.60	−10.72
		(17.70)	(19.54)	(0.00)	(20.00)	(19.91)	(0.00)
	Small	53.87	63.61	−9.74	51.43	59.80	−8.37
		(14.81)	(24.01)	(0.00)	(16.77)	(27.82)	(0.00)
	Private	54.33	53.15	1.18	51.60	47.63	3.97
		(16.31)	(17.24)	(0.58)	(17.61)	(20.09)	(0.05)
Service	Large	51.01	53.63	−2.62	49.85	53.65	−3.80
		(14.45)	(14.07)	(0.00)	(19.45)	(16.91)	(0.00)
	Mid	51.41	54.58	−3.17	50.23	55.15	−4.93
		(16.25)	(13.99)	(0.00)	(21.14)	(18.51)	(0.00)
	Small	52.08	52.05	0.04	51.03	52.07	−1.04
		(15.48)	(13.61)	(0.92)	(20.99)	(17.87)	(0.01)
	Private	48.88	51.82	−2.94	47.46	51.45	−3.99
		(15.00)	(13.25)	(0.00)	(18.92)	(17.72)	(0.00)
Resource	Large	43.52	43.32	0.20	43.03	42.23	0.80
		(10.15)	(12.52)	(0.81)	(12.35)	(10.09)	(0.31)
	Mid	43.08	42.97	0.11	42.69	41.84	0.84
		(10.08)	(10.42)	(0.88)	(11.96)	(14.69)	(0.39)
	Small	42.41	39.45	2.96	40.90	38.65	2.25
		(11.59)	(10.60)	(0.00)	(16.49)	(13.46)	(0.00)
	Private	42.89	45.07	−2.18	41.06	44.24	−3.19
		(10.59)	(10.37)	(0.00)	(12.09)	(12.62)	(0.00)
Manufacture	Large	43.54	43.00	0.54	42.59	42.74	−0.16
		(12.14)	(11.03)	(0.17)	(17.26)	(15.48)	(0.74)
	Mid	43.05	42.44	0.60	42.67	42.20	0.46
		(11.69)	(11.26)	(0.02)	(17.20)	(15.52)	(0.13)
	Small	41.55	39.90	1.65	41.61	39.93	1.68
		(12.39)	(11.35)	(0.00)	(17.69)	(14.76)	(0.00)
	Private	41.61	42.05	−0.44	40.63	41.88	−1.26
		(10.83)	(10.94)	(0.04)	(14.76)	(15.58)	(0.00)

Notes. The table summarizes cross metrics, reported as counts by industry group and size class at the CPC Group level. NP and P columns show statistics for non-patenting and patenting firms, respectively, and Diff columns show the difference between them. Means and medians are computed from firm-year observations for the period 1997 to 2023. Standard deviations and interquartile ranges are reported below means and medians, respectively. P-values, reported below differences, are based on 1,000 bootstrap iterations under the null that both groups are drawn from the same distribution. In each iteration, we sample with replacement from the pooled data, compute the difference in means and medians, and report the p-value as the proportion of iterations where the absolute resampled difference exceeds or equals the observed value.

Table B37: Cross Metrics by Industry and Size: CPC Patent, in Counts

Industry Group	Size Class	Mean Rate			Median Rate		
		NP	P	Diff	NP	P	Diff
Finance	Large	52.35	51.35	1.01	50.82	50.35	0.47
		(15.89)	(17.65)	(0.38)	(20.81)	(24.66)	(0.77)
	Mid	51.13	51.75	−0.62	49.13	48.37	0.76
		(16.23)	(21.53)	(0.79)	(21.66)	(27.29)	(0.81)
	Small	49.40	49.23	0.18	47.38	49.16	−1.78
		(15.35)	(23.31)	(0.91)	(19.83)	(32.88)	(0.38)
	Private	47.91	48.44	−0.53	46.37	44.48	1.89
		(15.75)	(18.98)	(0.77)	(19.20)	(25.54)	(0.40)
Service	Large	46.42	39.83	6.59	44.61	37.33	7.28
		(15.96)	(15.14)	(0.00)	(23.38)	(20.12)	(0.00)
	Mid	46.63	39.38	7.25	45.34	37.17	8.17
		(16.11)	(15.05)	(0.00)	(22.44)	(19.23)	(0.00)
	Small	46.00	38.52	7.48	44.95	36.34	8.60
		(15.93)	(14.05)	(0.00)	(22.32)	(18.71)	(0.00)
	Private	43.48	40.45	3.03	41.66	38.07	3.59
		(15.57)	(14.87)	(0.00)	(20.97)	(21.05)	(0.00)
Resource	Large	46.93	46.62	0.31	48.08	45.38	2.69
		(9.68)	(10.72)	(0.68)	(11.62)	(13.04)	(0.00)
	Mid	47.51	46.01	1.50	48.36	45.51	2.85
		(10.44)	(12.44)	(0.06)	(12.59)	(16.15)	(0.00)
	Small	46.62	41.47	5.14	46.84	41.42	5.42
		(11.39)	(13.75)	(0.00)	(14.97)	(21.88)	(0.00)
	Private	46.13	45.29	0.83	46.68	46.11	0.57
		(11.38)	(12.69)	(0.22)	(14.41)	(15.27)	(0.48)
Manufacture	Large	42.23	36.01	6.22	43.22	35.38	7.84
		(12.84)	(13.08)	(0.00)	(16.29)	(18.85)	(0.00)
	Mid	41.16	33.88	7.28	41.66	32.50	9.17
		(12.48)	(12.84)	(0.00)	(16.78)	(18.80)	(0.00)
	Small	36.59	30.56	6.03	36.04	28.36	7.68
		(13.18)	(11.68)	(0.00)	(20.10)	(15.78)	(0.00)
	Private	37.24	34.45	2.79	36.53	32.21	4.32
		(12.46)	(12.33)	(0.00)	(17.99)	(17.16)	(0.00)

Notes. The table summarizes cross metrics, reported as counts by industry group and size class at the CPC Patent level. NP and P columns show statistics for non-patenting and patenting firms, respectively, and Diff columns show the difference between them. Means and medians are computed from firm-year observations for the period 1997 to 2023. Standard deviations and interquartile ranges are reported below means and medians, respectively. P-values, reported below differences, are based on 1,000 bootstrap iterations under the null that both groups are drawn from the same distribution. In each iteration, we sample with replacement from the pooled data, compute the difference in means and medians, and report the p-value as the proportion of iterations where the absolute resampled difference exceeds or equals the observed value.

Table B38: Cross Metrics by Industry and Size: CPC Section, in Fractions

Industry Group	Size Class	Mean Rate			Median Rate		
		NP	P	Diff	NP	P	Diff
Finance	Large	2.20	2.70	−0.50	0.00	0.00	0.00
		(5.31)	(5.56)	(0.16)	(0.00)	(0.00)	(1.00)
	Mid	2.51	5.07	−2.56	0.00	0.00	0.00
		(5.75)	(7.31)	(0.00)	(0.00)	(14.31)	(1.00)
	Small	2.20	4.32	−2.11	0.00	0.00	0.00
		(5.47)	(6.31)	(0.00)	(0.00)	(10.48)	(1.00)
	Private	2.43	5.03	−2.60	0.00	0.00	0.00
		(5.54)	(6.49)	(0.00)	(0.00)	(11.45)	(1.00)
Service	Large	9.88	11.49	−1.61	12.40	13.99	−1.58
		(7.00)	(5.86)	(0.00)	(15.16)	(5.96)	(0.00)
	Mid	9.59	11.49	−1.90	11.98	14.16	−2.17
		(7.03)	(5.96)	(0.00)	(15.25)	(5.97)	(0.00)
	Small	9.76	11.48	−1.72	12.33	13.56	−1.23
		(7.00)	(5.58)	(0.00)	(15.25)	(5.94)	(0.00)
	Private	9.65	11.28	−1.62	12.13	13.56	−1.43
		(6.75)	(5.80)	(0.00)	(15.05)	(5.66)	(0.00)
Resource	Large	11.64	10.98	0.67	11.81	11.01	0.80
		(4.05)	(4.07)	(0.03)	(4.42)	(2.90)	(0.03)
	Mid	10.73	11.18	−0.45	10.85	11.12	−0.27
		(4.44)	(4.04)	(0.17)	(4.29)	(4.02)	(0.22)
	Small	10.69	10.25	0.44	10.84	10.87	−0.04
		(4.46)	(4.20)	(0.08)	(4.29)	(3.84)	(0.51)
	Private	10.87	11.45	−0.58	11.14	12.03	−0.89
		(4.11)	(4.05)	(0.02)	(3.86)	(4.71)	(0.00)
Manufacture	Large	11.64	12.24	−0.60	11.15	12.53	−1.38
		(4.73)	(3.99)	(0.00)	(5.80)	(5.42)	(0.00)
	Mid	12.07	12.23	−0.16	11.67	12.71	−1.04
		(4.11)	(3.76)	(0.08)	(5.14)	(4.96)	(0.00)
	Small	11.93	11.73	0.20	11.80	12.11	−0.32
		(4.25)	(3.80)	(0.00)	(5.63)	(5.49)	(0.00)
	Private	11.64	11.57	0.07	11.69	11.58	0.11
		(3.86)	(3.41)	(0.29)	(4.79)	(4.77)	(0.24)

Notes. The table summarizes cross metrics, reported as fractions by industry group and size class at the CPC Section level. NP and P columns show statistics for non-patenting and patenting firms, respectively, and Diff columns show the difference between them. Means and medians are computed from firm-year observations for the period 1997 to 2023. Standard deviations and interquartile ranges are reported below means and medians, respectively. P-values, reported below differences, are based on 1,000 bootstrap iterations under the null that both groups are drawn from the same distribution. In each iteration, we sample with replacement from the pooled data, compute the difference in means and medians, and report the p-value as the proportion of iterations where the absolute resampled difference exceeds or equals the observed value.

Table B39: Cross Metrics by Industry and Size: CPC Class, in Fractions

Industry Group	Size Class	Mean Rate			Median Rate		
		NP	P	Diff	NP	P	Diff
Finance	Large	13.23	13.92	−0.68	16.04	15.53	0.51
		(8.32)	(8.51)	(0.22)	(11.48)	(11.78)	(0.43)
	Mid	13.09	14.58	−1.49	15.57	17.58	−2.01
		(8.35)	(8.51)	(0.21)	(12.06)	(11.58)	(0.09)
	Small	15.12	15.34	−0.21	16.91	16.00	0.91
		(7.28)	(8.38)	(0.80)	(7.37)	(7.55)	(0.20)
	Private	13.36	11.24	2.12	15.55	11.27	4.28
		(7.77)	(7.63)	(0.03)	(9.17)	(7.88)	(0.00)
Service	Large	14.44	15.83	−1.38	14.55	16.54	−1.99
		(4.40)	(4.25)	(0.00)	(5.84)	(4.94)	(0.00)
	Mid	14.34	16.26	−1.92	14.50	16.78	−2.28
		(4.79)	(4.36)	(0.00)	(6.42)	(4.56)	(0.00)
	Small	14.58	15.59	−1.01	14.89	16.25	−1.36
		(4.57)	(4.10)	(0.00)	(6.15)	(5.09)	(0.00)
	Private	13.55	15.21	−1.66	13.81	15.51	−1.70
		(4.87)	(3.93)	(0.00)	(6.16)	(4.86)	(0.00)
Resource	Large	11.19	10.80	0.39	10.81	10.69	0.12
		(2.56)	(2.60)	(0.05)	(3.47)	(2.08)	(0.46)
	Mid	10.94	10.95	−0.00	10.66	10.50	0.16
		(2.86)	(2.34)	(0.99)	(2.76)	(2.27)	(0.24)
	Small	10.91	10.85	0.06	10.37	10.35	0.03
		(2.94)	(2.35)	(0.73)	(3.85)	(3.02)	(0.82)
	Private	10.94	11.85	−0.91	10.42	11.49	−1.07
		(2.77)	(2.57)	(0.00)	(3.14)	(3.16)	(0.00)
Manufacture	Large	11.83	12.58	−0.74	11.06	12.29	−1.23
		(3.44)	(3.17)	(0.00)	(5.22)	(4.81)	(0.00)
	Mid	11.90	12.48	−0.58	11.05	12.22	−1.16
		(3.13)	(3.20)	(0.00)	(4.89)	(4.78)	(0.00)
	Small	11.93	12.04	−0.11	11.36	11.74	−0.39
		(3.29)	(3.07)	(0.00)	(5.19)	(4.59)	(0.00)
	Private	11.67	12.04	−0.38	11.04	11.56	−0.52
		(3.07)	(2.83)	(0.00)	(4.22)	(4.31)	(0.00)

Notes. The table summarizes cross metrics, reported as fractions by industry group and size class at the CPC Class level. NP and P columns show statistics for non-patenting and patenting firms, respectively, and Diff columns show the difference between them. Means and medians are computed from firm-year observations for the period 1997 to 2023. Standard deviations and interquartile ranges are reported below means and medians, respectively. P-values, reported below differences, are based on 1,000 bootstrap iterations under the null that both groups are drawn from the same distribution. In each iteration, we sample with replacement from the pooled data, compute the difference in means and medians, and report the p-value as the proportion of iterations where the absolute resampled difference exceeds or equals the observed value.

Table B40: Cross Metrics by Industry and Size: CPC Subclass, in Fractions

Industry Group	Size Class	Mean Rate			Median Rate		
		NP	P	Diff	NP	P	Diff
Finance	Large	16.50	16.67	−0.17	15.84	16.09	−0.25
		(6.97)	(6.96)	(0.73)	(6.53)	(6.69)	(0.49)
	Mid	16.04	19.50	−3.46	15.42	18.97	−3.54
		(6.73)	(6.48)	(0.00)	(6.68)	(6.06)	(0.00)
	Small	15.54	18.28	−2.74	15.27	17.20	−1.93
		(5.41)	(6.99)	(0.00)	(5.91)	(7.92)	(0.00)
	Private	15.59	14.74	0.85	14.98	13.80	1.18
		(6.46)	(7.51)	(0.30)	(6.20)	(7.28)	(0.10)
Service	Large	13.66	14.51	−0.84	13.36	14.79	−1.43
		(3.75)	(3.66)	(0.00)	(4.91)	(4.52)	(0.00)
	Mid	13.74	14.76	−1.02	13.51	15.12	−1.61
		(4.08)	(3.64)	(0.00)	(5.50)	(4.72)	(0.00)
	Small	13.95	14.19	−0.24	13.87	14.39	−0.53
		(3.89)	(3.40)	(0.00)	(5.32)	(4.40)	(0.00)
	Private	13.12	13.97	−0.85	12.82	13.89	−1.08
		(4.17)	(3.38)	(0.00)	(4.90)	(4.37)	(0.00)
Resource	Large	11.29	11.06	0.24	10.84	10.75	0.10
		(2.60)	(2.45)	(0.27)	(3.57)	(1.98)	(0.61)
	Mid	11.21	11.44	−0.23	10.83	10.98	−0.15
		(2.82)	(2.70)	(0.30)	(2.88)	(3.04)	(0.46)
	Small	11.24	10.68	0.56	10.68	10.42	0.26
		(3.04)	(2.56)	(0.00)	(4.54)	(2.69)	(0.14)
	Private	11.17	11.82	−0.65	10.57	11.35	−0.78
		(2.80)	(2.53)	(0.00)	(2.95)	(3.48)	(0.00)
Manufacture	Large	11.91	12.07	−0.16	11.49	11.90	−0.41
		(3.19)	(2.81)	(0.09)	(4.97)	(4.20)	(0.00)
	Mid	11.91	12.00	−0.09	11.51	11.87	−0.36
		(2.99)	(2.86)	(0.14)	(4.68)	(4.05)	(0.00)
	Small	11.67	11.48	0.19	11.52	11.46	0.06
		(3.15)	(2.89)	(0.00)	(4.81)	(3.87)	(0.20)
	Private	11.45	11.91	−0.46	11.08	11.73	−0.65
		(2.88)	(2.77)	(0.00)	(3.70)	(4.00)	(0.00)

Notes. The table summarizes cross metrics, reported as fractions by industry group and size class at the CPC Subclass level. NP and P columns show statistics for non-patenting and patenting firms, respectively, and Diff columns show the difference between them. Means and medians are computed from firm-year observations for the period 1997 to 2023. Standard deviations and interquartile ranges are reported below means and medians, respectively. P-values, reported below differences, are based on 1,000 bootstrap iterations under the null that both groups are drawn from the same distribution. In each iteration, we sample with replacement from the pooled data, compute the difference in means and medians, and report the p-value as the proportion of iterations where the absolute resampled difference exceeds or equals the observed value.

Table B41: Cross Metrics by Industry and Size: CPC Group, in Fractions

Industry Group	Size Class	Mean Rate			Median Rate		
		NP	P	Diff	NP	P	Diff
Finance	Large	13.99	14.02	−0.02	13.28	13.37	−0.09
		(4.30)	(4.32)	(0.93)	(5.03)	(4.87)	(0.78)
	Mid	13.37	15.62	−2.25	12.54	15.25	−2.70
		(4.03)	(4.60)	(0.00)	(4.50)	(4.70)	(0.00)
	Small	12.57	15.04	−2.47	11.95	13.81	−1.87
		(3.33)	(5.65)	(0.00)	(3.55)	(6.36)	(0.00)
	Private	12.62	12.53	0.09	11.96	11.02	0.95
		(3.72)	(4.06)	(0.85)	(3.86)	(4.80)	(0.03)
Service	Large	11.88	12.67	−0.79	11.61	12.76	−1.15
		(3.17)	(3.14)	(0.00)	(4.35)	(3.94)	(0.00)
	Mid	11.96	12.88	−0.92	11.58	13.05	−1.47
		(3.62)	(3.18)	(0.00)	(4.83)	(4.33)	(0.00)
	Small	12.09	12.17	−0.08	11.78	12.21	−0.42
		(3.44)	(3.08)	(0.28)	(4.76)	(4.31)	(0.00)
	Private	11.40	12.03	−0.64	11.01	11.93	−0.91
		(3.37)	(2.97)	(0.00)	(4.21)	(4.08)	(0.00)
Resource	Large	10.11	10.08	0.03	10.01	9.79	0.22
		(2.14)	(2.73)	(0.88)	(2.51)	(2.16)	(0.17)
	Mid	10.04	9.98	0.06	9.95	9.67	0.28
		(2.16)	(2.20)	(0.75)	(2.46)	(3.12)	(0.13)
	Small	9.89	9.27	0.62	9.55	9.13	0.42
		(2.42)	(2.22)	(0.00)	(3.27)	(2.83)	(0.01)
	Private	10.00	10.41	−0.41	9.62	10.25	−0.63
		(2.26)	(2.15)	(0.01)	(2.54)	(2.71)	(0.00)
Manufacture	Large	10.14	10.00	0.14	9.97	9.92	0.05
		(2.60)	(2.34)	(0.08)	(3.53)	(3.25)	(0.63)
	Mid	10.01	9.85	0.16	9.91	9.77	0.15
		(2.43)	(2.40)	(0.00)	(3.49)	(3.25)	(0.03)
	Small	9.66	9.29	0.37	9.68	9.27	0.40
		(2.62)	(2.42)	(0.00)	(3.64)	(3.14)	(0.00)
	Private	9.71	9.73	−0.02	9.49	9.67	−0.18
		(2.32)	(2.32)	(0.68)	(3.15)	(3.31)	(0.00)

Notes. The table summarizes cross metrics, reported as fractions by industry group and size class at the CPC Group level. NP and P columns show statistics for non-patenting and patenting firms, respectively, and Diff columns show the difference between them. Means and medians are computed from firm-year observations for the period 1997 to 2023. Standard deviations and interquartile ranges are reported below means and medians, respectively. P-values, reported below differences, are based on 1,000 bootstrap iterations under the null that both groups are drawn from the same distribution. In each iteration, we sample with replacement from the pooled data, compute the difference in means and medians, and report the p-value as the proportion of iterations where the absolute resampled difference exceeds or equals the observed value.

Table B42: Cross Metrics by Industry and Size: CPC Patent, in Fractions

Industry Group	Size Class	Mean Rate			Median Rate		
		NP	P	Diff	NP	P	Diff
Finance	Large	12.16	12.12	0.04	11.74	11.87	−0.14
		(3.58)	(4.00)	(0.87)	(4.49)	(5.23)	(0.68)
	Mid	11.97	12.24	−0.27	11.39	11.82	−0.44
		(3.68)	(5.03)	(0.57)	(4.62)	(6.14)	(0.45)
	Small	11.55	11.65	−0.10	10.97	11.89	−0.92
		(3.50)	(5.44)	(0.78)	(4.28)	(7.08)	(0.03)
	Private	11.14	11.42	−0.28	10.67	10.21	0.45
		(3.62)	(4.41)	(0.53)	(4.31)	(5.80)	(0.35)
Service	Large	10.82	9.40	1.42	10.36	8.95	1.41
		(3.54)	(3.36)	(0.00)	(4.96)	(4.46)	(0.00)
	Mid	10.87	9.28	1.58	10.53	8.78	1.74
		(3.57)	(3.38)	(0.00)	(4.79)	(4.29)	(0.00)
	Small	10.69	9.00	1.69	10.45	8.55	1.91
		(3.54)	(3.16)	(0.00)	(4.79)	(4.21)	(0.00)
	Private	10.14	9.38	0.76	9.74	8.85	0.89
		(3.51)	(3.35)	(0.00)	(4.68)	(4.72)	(0.00)
Resource	Large	10.95	10.88	0.06	11.13	10.55	0.58
		(2.03)	(2.34)	(0.69)	(2.33)	(2.87)	(0.00)
	Mid	11.11	10.73	0.38	11.24	10.65	0.59
		(2.25)	(2.78)	(0.03)	(2.63)	(3.23)	(0.00)
	Small	10.92	9.77	1.14	10.93	9.81	1.12
		(2.41)	(3.05)	(0.00)	(3.10)	(4.93)	(0.00)
	Private	10.78	10.47	0.31	10.89	10.68	0.21
		(2.50)	(2.76)	(0.05)	(3.11)	(3.43)	(0.23)
Manufacture	Large	9.91	8.41	1.50	10.10	8.24	1.86
		(2.90)	(2.94)	(0.00)	(3.68)	(4.25)	(0.00)
	Mid	9.64	7.90	1.74	9.77	7.54	2.23
		(2.80)	(2.89)	(0.00)	(3.71)	(4.25)	(0.00)
	Small	8.55	7.14	1.41	8.42	6.62	1.80
		(2.92)	(2.61)	(0.00)	(4.53)	(3.40)	(0.00)
	Private	8.71	7.98	0.74	8.56	7.49	1.08
		(2.80)	(2.72)	(0.00)	(4.10)	(3.81)	(0.00)

Notes. The table summarizes cross metrics, reported as fractions by industry group and size class at the CPC Patent level. NP and P columns show statistics for non-patenting and patenting firms, respectively, and Diff columns show the difference between them. Means and medians are computed from firm-year observations for the period 1997 to 2023. Standard deviations and interquartile ranges are reported below means and medians, respectively. P-values, reported below differences, are based on 1,000 bootstrap iterations under the null that both groups are drawn from the same distribution. In each iteration, we sample with replacement from the pooled data, compute the difference in means and medians, and report the p-value as the proportion of iterations where the absolute resampled difference exceeds or equals the observed value.

B.5 Technology Momentum

Table B43: Technology Momentum Monthly Alpha by Factor Model

		Non-Patenting Firms					Patenting Firms				
	Decile	One	Three	Four	Five	Six	One	Three	Four	Five	Six
Equal-Weighted	High	0.90	0.84	0.89	0.83	0.87	0.71	0.68	0.74	0.81	0.86
		(3.10)	(3.55)	(3.75)	(3.38)	(3.55)	(1.71)	(2.07)	(2.24)	(2.56)	(2.71)
	Low	−1.34	−1.39	−1.24	−1.27	−1.17	−0.37	−0.40	−0.28	−0.22	−0.14
		(−4.38)	(−5.07)	(−4.66)	(−4.55)	(−4.32)	(−1.24)	(−1.46)	(−1.04)	(−0.76)	(−0.50)
	High–Low	2.08	2.07	1.97	1.94	1.88	0.93	0.93	0.87	0.87	0.84
		(5.01)	(4.98)	(4.74)	(4.57)	(4.43)	(1.75)	(1.82)	(1.69)	(1.69)	(1.63)
Value-Weighted	High	0.57	0.56	0.51	0.49	0.46	0.22	0.25	0.24	0.15	0.16
		(2.46)	(2.41)	(2.22)	(2.05)	(1.93)	(0.83)	(0.94)	(0.92)	(0.58)	(0.61)
	Low	−0.45	−0.46	−0.38	−0.43	−0.37	0.18	0.20	0.27	0.24	0.29
		(−1.87)	(−1.89)	(−1.58)	(−1.71)	(−1.51)	(0.72)	(0.78)	(1.06)	(0.93)	(1.13)
	High–Low	0.86	0.86	0.74	0.76	0.68	−0.11	−0.10	−0.18	−0.25	−0.29
		(2.19)	(2.17)	(1.88)	(1.86)	(1.68)	(−0.26)	(−0.24)	(−0.42)	(−0.57)	(−0.66)

Notes. The table shows monthly alpha in percentage points from factor models estimated using monthly returns from 1998 to 2023 for equal-weighted and value-weighted technology momentum portfolios, for patent-owning and non-owning firms. The five models we estimate are: 1) a market model (MKT), 2) a three-factor market, size, and value model (MKT, SMB, HML), 3) a four-factor market, size, value, and momentum model (MKT, SMB, HML, MOM), 4) a five-factor market, size, value, profitability, and investment model (MKT, SMB, HML, RMW, CMA), and 5) a six-factor market, size, value, profitability, investment, and momentum model (MKT, SMB, HML, RMW, CMA, MOM). Each sub-table shows factor loadings for high-decile and low-decile portfolios, as well as for a high-minus-low portfolio. T-statistics are reported in parentheses under each alpha estimate.