Patents to Products:
Product Innovation and Firm Dynamics*

David Argente — Yale
Salomé Baslandze — FRB Atlanta
Douglas Hanley — U. of Pittsburgh
Sara Moreira — Northwestern

December 18, 2025

Abstract

We match patents to products using natural language methods applied to detailed product descriptions and patent texts in the consumer goods sector. While more than half of product innovations originate from non-patenting firms, patent filings are on average followed by subsequent product introductions. Yet this relationship weakens with firm size. Patents held by market leaders also yield revenue premiums beyond what can be explained by their own product introductions and are associated with stronger deterrence of competitors' innovations. To interpret these findings, we develop a simple growth model in which larger firms have stronger incentives to engage in strategic patenting—filing for protection rather than market innovation—which dampens innovation and slows creative destruction.

JEL Classification: O31, O34, O4
Keywords: growth, creative destruction, strategic patents, innovation, products.

*Emails: david.argente@yale.edu; salome.baslandze@atl.frb.org; sara.moreira@kellogg.northwestern.edu. We thank Ufuk Akcigit, Fernando Alvarez, Antonin Bergeaud, Toni Braun, Laurent Frésard, Pete Klenow, Hugo Hopenhayn, Benjamin F. Jones, Claudio Michelacci, Jesse Perla, Juan Rubio-Ramirez, and numerous seminar and conference participants for their helpful comments. Researcher(s)' own analyses calculated (or derived) based in part on data from Nielsen Consumer LLC and marketing databases provided through the NielsenIQ Datasets at the Kilts Center for Marketing Data Center at The University of Chicago Booth School of Business. The conclusions drawn from the NielsenIQ data are those of the researcher(s) and do not reflect the views of NielsenIQ. NielsenIQ is not responsible for, had no role in, and was not involved in analyzing and preparing the results reported herein. The views expressed in this paper do not necessarily represent the views of the Federal Reserve System or the Federal Reserve Bank of Atlanta. First draft: March 2019.

1 Introduction

Product innovation is a key engine of long-run growth (Romer, 1990; Grossman and Helpman, 1991). By expanding varieties and improving quality, new products raise productivity, enhance consumer welfare, and spur creative destruction. Yet because detailed data on new product introductions are scarce, empirical work has relied heavily on patents as the primary innovation metric. However, the extent to which patented ideas are reflected in market innovations remains largely unknown. This measurement gap is critical, since patents also serve a defensive purpose—used strategically to block rivals and extend incumbency[1]—and may not always be associated with innovation brought to market. Understanding the extent to which firms grow through product innovation or strategic use of patents is especially important in light of recent trends: patenting has surged even as productivity growth and business dynamism have slowed (Decker, Haltiwanger, Jarmin and Miranda, 2016; Crouzet and Eberly, 2019; Bloom, Jones, Van Reenen and Webb, 2020; Akcigit and Ates, 2023; Aghion, Bergeaud, Boppart, Klenow and Li, 2023), raising concerns that the patent system may increasingly serve to sustain market power rather than foster innovation.

In this paper, we address this gap by developing the first large-scale dataset that links patents to product innovations in the U.S. consumer goods sector. We combine NielsenIQ scanner data with USPTO patent records and apply natural language processing techniques to match detailed product descriptions with patent texts. We show that while more than half of product innovations in the sector originate from firms that do not patent, patent filings are, on average, associated with product introductions. However, this association between product introduction and patenting weakens with firm size. Furthermore, relative to smaller firms, patents held by large firms generate larger revenue premiums beyond what can be explained by their own product introductions and are associated with stronger deterrence of competitors' future product introductions. To interpret these findings, we develop a stylized Schumpeterian model (Aghion and Howitt, 1992) in which the returns to product innovation decline with firm size while the protective value of patenting increases. Calibrated to our data, the model's counterfactuals imply that a sizable share of large-firm patents are strategic—filed for protective purposes without corresponding market innovations—thereby reducing innovation and reallocation in the consumer goods sector.

We start by constructing a novel dataset that links patents to product introductions at scale. This requires overcoming two main empirical challenges: (i) measuring product

[1]See the classic study by Gilbert and Newbery (1982), along with more recent contributions by Cohen, Nelson and Walsh (2000); Jaffe and Lerner (2004); Bessen and Meurer (2008); Boldrin and Levine (2013). Some of the popular press articles are: "The Experts: Does the Patent System Encourage Innovation?" (WSJ, 2013) and "Save America's Patent System" (NYT, 2022).

innovation in the market and (ii) linking innovations to the patents that underlie them. We address these challenges by combining comprehensive retail scanner data with information on U.S. patents, enriched by modern text-analysis techniques.

The core data source is NielsenIQ scanner data, which provide product information for the consumer goods sector from 2006 to 2015, accounting for approximately 12.5-14% of total goods consumption These data provide detailed product attributes (e.g., formula, style, content) and record sales and prices information. We use this richness to construct measures of product innovation. Our simplest measure is the count of new products (barcodes) introduced by a firm in a given product category-year. To better capture substantive innovations, we also build a quality-adjusted measure that incorporates both the new attributes a product brings to the market and the market response in terms of prices and sales. By drawing on scanner data that cover thousands of consumer goods products (e.g., lamps, batteries, over-the-counter pharmaceuticals, laundry detergents, yogurts) with near-universal coverage, we obtain a comprehensive view of product innovations in the sector, the firms that introduce them, and their market positions.

The second challenge is to match products to firms' patents. We begin at the firm level, linking company names across the NielsenIQ and USPTO data to track overall patenting and product introductions jointly over time. We then develop a more granular match at the patent and product level. To do so, we use detailed product descriptions from NielsenIQ, enriched with text from Wikipedia articles, together with firms' patent texts, and apply natural language processing methods to assign patents to product categories based on text similarity.[2] This approach yields our benchmark patent-to-products dataset at the firm×category×year level, enabling us to observe whether specific firm patents translate into commercial products. We link over 190,000 patent applications (and up to 400,000 when including those filed prior to Nielsen coverage window) to active product categories of NielsenIQ firms. For example, our matching procedure links Procter & Gamble's 2011 patent on water-soluble film, which led to Tide Pods, to laundry detergents, and Beyond Meat's 2014 patent on plant-based meat to subsequent simulated beef product introductions.

We conduct extensive validation and robustness checks to ensure that our matching procedure accurately links patents to products and filters out unrelated patents, such as those outside the consumer goods sector or cost-reducing process patents. The algorithm is robust to perturbations, and we validate its performance through external benchmarks, placebo tests, and checks on filtered patents. Because our data focus exclusively on product

[2] See Manning, Raghavan and Schütze (2008). Similar techniques are used by Younge and Kuhn (2016), Webb (2019), Kelly, Papanikolaou, Seru and Taddy (2021), and Kalyani, Bloom, Carvalho, Hassan, Lerner and Tahoun (2025).

innovations, we ensure that the algorithm explicitly excludes general process and method patents, which are unlikely to be directly associated with product introductions, and we conduct multiple validation exercises using external measures (Bena and Simintzi, 2017; Tham, Baslandze, Sojli and Liu, 2025). This approach guarantees that our analysis consistently centers on product innovations and the patents most plausibly related to them. While cost-reducing process innovations are an important component of firms' broader innovative activities, our study deliberately focuses on product innovation and patenting to maintain a consistent scope.

The resulting dataset tracks product patents and products commercialized for firms in the consumer goods sector. Although our analysis is confined to this sector, patenting intensities and product introduction rates are broadly comparable to those in other manufacturing industries: of 35,000 firms in our data, 15% applied for a patent at least once (9% during the NielsenIQ sample period), a rate in line with manufacturing and at least three times higher than that of the U.S. economy (Graham, Grim, Islam, Marco and Miranda, 2018; Mezzanotti and Simcoe, 2025). Consumer products are also among the most active in patent litigation (PwC, 2018), underscoring the importance of patents as competitive tools. The sector spans a wide range of categories with distinct patenting intensities, enabling us to document heterogeneous patterns across product types.

In the first part of our analysis, we use our patents-to-product dataset to reevaluate the association between patents and product innovation. Patents have long been employed as proxies for innovation because they are observable, frequent, and correlated with firm performance, including stock market value, growth, and productivity (Hall, Thoma and Torrisi, 2007; Balasubramanian and Sivadasan, 2011; Kogan, Papanikolaou, Seru and Stoffman, 2017). Leveraging our novel match between patents and product introductions, we directly test how patenting relates to market innovation. First, we find that firms without patents account for the majority of product innovation in the sector: they introduce more than half of all new products, 65% of quality-adjusted innovations, and account for 58% of sectoral sales growth. Next, exploiting variation within firm $\times$ product category over time, we estimate how new patent filing rates relate to product introduction rates, controlling for category–time and firm–category fixed effects. We find that a 10% increase in the patenting rate is associated with a 0.4% increase in the product introduction rate in the following year, with the strongest effects at one- to two-year lags, and that these effects are persistent. The relationship is stronger for granted and citation-adjusted patents, suggesting that these patent measures better capture meaningful differences in realized innovation. The positive association holds in both food and non-food product categories, with particularly strong effects in health and beauty, non-food grocery, and general merchandise.

Next, leveraging our joint data on product introductions and patents, we examine how the relationship between patents and product introduction varies with firm size. We show that smaller firms within a product category introduce new products at much higher rates, whereas market leaders introduce products at lower rates and file substantially more patents per new product. This higher patent intensity is not explained by a shift toward fewer but more radical innovations, as quality-adjusted product introduction rates also decline with firm size.[3] Regression estimates confirm that the positive association between patenting and subsequent product introductions is strong for small firms but weakens sharply as firms grow: the patents-to-product introduction coefficient is 0.15 for the lowest quintile of firms, but more than six times smaller for those in the top quintile, indicating that leaders' patents are relatively less likely to yield actual product commercialization. The inclusion of a rich set of fixed effects ensures that our estimates are not driven by systematic differences in patentability or innovation across product categories over time, nor by firm-specific, time-invariant predispositions to patent—such as lower costs due to firm size, accumulated experience, or specialization within a category. Extensive robustness checks also demonstrate that the weaker link between patents and product introductions among larger firms is not driven by delayed commercialization, misclassified process-related patents, data coverage issues, international product launches, or size-dependent measurement error in our text-matching algorithm. Additionally, we find that patents by market leaders are, on average, less scientifically novel, receive fewer forward citations (particularly from competitors), and are more likely to be litigated.

The robust pattern of a weaker association between patents and actual innovations among larger firms has two main implications. First, the declining patent-innovation relationship indicates that product patents are less reliable proxies for market innovation among large firms, with important consequences for empirical work that relies on patents to measure innovative activity. Second, the weaker conversion of patents into market innovations suggests that the incentives to patent and to innovate vary systematically with firm size. In particular, large firms may enjoy additional benefits, or face lower marginal costs, of filing patents to deter competition and protect their market position, and patents that are not followed by product commercialization likely serve primarily strategic purposes.

In the second part of our empirical analysis, we turn to the roles of product innovation and patenting in firm growth. Using data on revenues, product introductions, and patent filings, we estimate the revenue premium from patents with and without accompanying product innovations. Patents are associated with higher revenues—a 10% increase in patent

[3] Mezzanotti and Simcoe (2025) provides evidence that large firms systematically use patents more extensively, even after adjusting for product market, and for both the quantity and quality of R&D.

stock raises sales by 1.5% in the following year—but the elasticity of sales with respect to new product introductions is substantially larger, underscoring central role of new product launches in firm growth (Hottman, Redding and Weinstein, 2016; Argente, Lee and Moreira, 2024). For large firms, however, a sizable residual patent revenue premium remains even after conditioning on product launches. We further show that intensified patenting by market leaders is followed by fewer product introductions from competitors, whereas smaller firms' patents have no such effect. Overall, these results suggest that patents contribute to firm growth through both productive and strategic channels: while innovation broadly drives growth, patenting plays a disproportionate role for large incumbents by deterring rivals and sustaining revenues.

To interpret these findings, we conclude the paper with an illustrative growth model that separates firms' incentives to commercialize new products from their incentives to patent. The model rationalizes our empirical finding of a declining association between patents and product innovation as firms grow by showing how patenting and innovation incentives diverge with firm size: as firms expand, the returns to product introduction decline due to rent cannibalization, whereas the benefits of patenting rise because larger incumbents have more to protect. This generates strong incentives for strategic patents— patents filed for protective purposes without corresponding market innovations, which allow incumbents to defend their market position but weaken creative destruction. Calibrated to our empirical moments, the counterfactuals suggest that a sizable share of patents filed by large firms are strategic, reducing creative destruction by about 3% in product categories dominated by large incumbents. Eliminating the option to patent without commercialization increases innovation incentives for large incumbents, who must now innovate to protect their market positions, and raises reallocation, underscoring the cost of strategic patenting for growth.

The model shows that larger firms obtain stronger benefits from patenting. We also consider an extension with size-dependent patenting costs, in which larger firms face lower marginal costs of patenting and enforcement due to scale economies in research and development, legal capacity, and litigation resources. Our findings show that these forces make the incentives for strategic patenting even more prominent. Although deliberately stylized, the model highlights the broader mechanisms through which patents that are not followed by product commercialization can sustain incumbents' market power and weaken both innovation and reallocation in the consumer goods sector.

Related Literature – Our findings contribute to the growing literature on rising market concentration and the role that large firms may play in the productivity slowdown. While patenting has surged, productivity growth and business dynamism have slowed (Decker, Haltiwanger, Jarmin and Miranda, 2016; Gordon, 2016; Akcigit and Ates, 2023). Large firms increasingly invest in intangibles, including intellectual property, yet this shift often coincides with rising concentration rather than faster aggregate innovation (Gutiérrez and Philippon, 2017; Crouzet and Eberly, 2019; De Loecker, Eeckhout and Unger, 2020; Autor, Dorn, Katz, Patterson and Van Reenen, 2020). Recent work highlights a range of strategic behaviors by large incumbents that contribute to weaker competition and slower productivity growth: defensive inventor hiring (Akcigit and Goldschlag, 2023; Fernández-Villaverde, Yu and Zanetti, 2025), political influence (Akcigit, Baslandze and Lotti, 2023; Gutiérrez and Philippon, 2019), acquisitions of potential rivals (Cunningham, Ederer and Ma, 2021), brand reallocation (Pearce and Wu, 2024), and slower knowledge diffusion from leaders to laggards (Akcigit and Ates, 2023).[4] Leveraging data that separately track firm patents and market innovations, we document an additional strategic margin: larger firms increasingly patent without commercializing products, a practice that can limit reallocation and dampen overall innovation. Combined with the rising share of patenting by large firms (Akcigit and Ates, 2021), this pattern also can partly reconcile the coexistence of rising aggregate patenting and relatively modest productivity growth (Bloom et al., 2020).

Our novel data set also speaks to the longstanding question of how well patents measure innovation. A large literature has relied on indirect proxies in the absence of direct measures. Some studies infer innovation from employment or sales growth (Garcia-Macia, Hsieh and Klenow, 2019), while others value innovation directly from patents (e.g., Akcigit and Kerr, 2018) or combined with stock market data (Kogan, Papanikolaou, Seru and Stoffman, 2017). Alternative approaches consider innovations occurring outside the patent system, such as new technical books (Alexopoulos, 2011) or innovations exhibited at World Fairs (Moser, 2012).[5] Our contribution is to link patents to specific product introductions at the firm level—an otherwise unobservable relationship—allowing us to assess how well patent metrics capture true market innovations. We show that the informativeness of patent metrics in the consumer goods sector varies systematically across the firm-size distribution: patents track market innovation much less closely for larger firms.

The rest of the paper is organized as follows. Section 2 describes the data, matching algorithms, and extensive validation checks. Section 3 presents the main empirical results on patents and product innovation. Section 3.2 examines the relationship between patents, innovation, and firm growth. Section 5 introduces a simple model with a quantitative illustration. Section 6 concludes.

[4]See Baslandze (2021) for a review.

[5]See a comprehensive survey of the innovation literature by Bryan and Williams (2021).

2 Data, Matching Algorithms, and Measurement

We face two main challenges in studying the relationship between patents and product innovation. First, while patent data are widely available, large-scale information on new product introductions is rare. Second, linking patents to related products is difficult. This section outlines our empirical strategies to address these challenges.

We build a dataset on product introductions using NielsenIQ scanner data covering consumer goods firms from 2006 to 2015. Products are identified by barcodes, with detailed characteristics that allow us to construct measures of product innovation. Patent data come from the United States Patent and Trademark Office (USPTO). Together, these sources provide sector-wide coverage of both patents and product innovations.

To link patents to products, we begin by using firm names in the patent and product data sets to map firms' patent portfolios to their products. This mapping is too coarse to connect patents with specific products and does not filter out unrelated patents, such as those tied to products outside the consumer goods sector or to general, cost-reducing process improvements. We therefore draw on the rich information contained in patent records and product descriptions from NielsenIQ and Wikipedia, applying natural language processing methods to systematically link sets of patents to related products within firms.

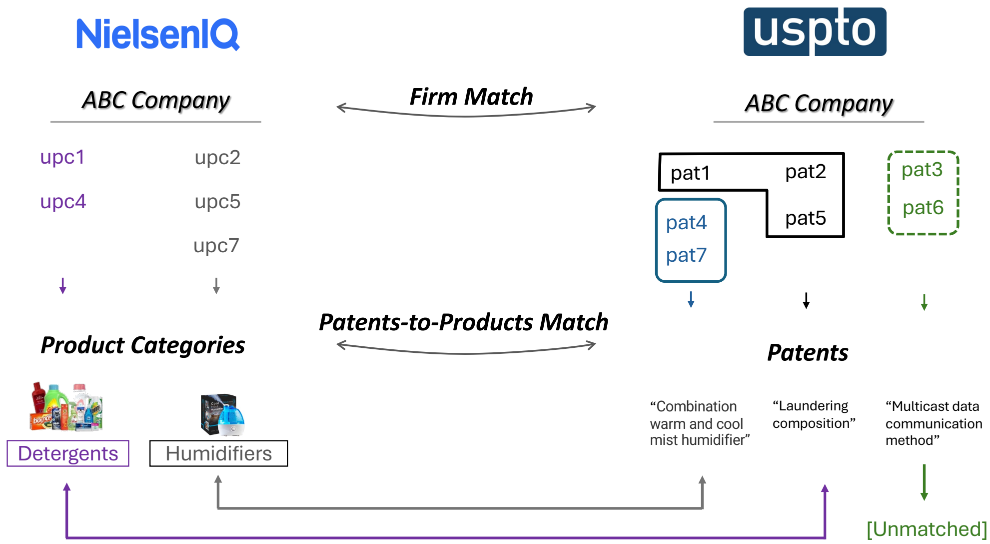
Because a patent may generate multiple products (or none) and a product may draw on multiple patents (or none), one-to-one matching is neither feasible nor desirable. Instead, we group similar products into categories and assign patents based on textual similarity between patent texts and product descriptions within those categories. This yields our benchmark patents-to-products dataset. Figure 1 illustrates the construction, with algorithmic details provided below.

To our knowledge, this algorithm produces a unique dataset. de Rassenfosse (2018) collect data on virtual patent markings for 100 firms, but cannot measure product introduction, sales, or prices, nor do they offer sector-wide coverage. Some private firms also link patents to products for clients (e.g., FairTech, IPStrategy, Powering Ideas, IntellectPeritus), but their datasets are confidential and limited to client portfolios.

2.1 Data

Product Data – Our primary source of product information is the scanner data set from NielsenIQ Retail Measurement Services (RMS), provided by the Kilts Data Center at the University of Chicago Booth School of Business. This data set is collected from point-of-sale systems in grocery, drug, and general-merchandise stores. The data set consists of more than one million distinct products identified by Universal Product Codes (UPCs), which

Figure 1: Product and Patent Data Sets



Notes: This diagram exemplifies the construction of the data sets linking products and patents. In this example, under Firm-level Match, all patents assigned to ABC Company are matched to all products sold by ABC Company. With Patent-to-Products Match, *pat*4 and *pat*7 match to *upc*1 and *upc*4 products that comprise a product category "Detergents"; *pat*1, *pat*2, and *pat*5 match to *upc*2, *upc*5, and *upc*7 that comprise a product category "Humidifiers"; patents *pat*3 and *pat*6 are unmatched.

are scanned at the point of sale. Each UPC consists of 12 numerical digits that are uniquely assigned to each product, and we use these to identify products. UPCs carry information about the brand and a rich set of product attributes like size, packaging, and flavor.

The data focus on the consumer packaged goods (CPG) sector, which accounts for 14% of total goods consumption in the U.S.[6] This sector includes both food and non-food categories, such as health and beauty aids, over-the-counter pharmaceuticals, household supplies, and general merchandise—including cookware, small electronics, gardening products, and other durable consumer goods. Our data cover the years from 2006-2015 and combine all sales, quantities, and prices at the national and annual levels. We use the panel structure of each UPC to measure its entry year. This product data set covers about 40% of the CPG sector sales, and nearly the universe of firms and new products in the sector. Appendix A.1 provides additional details about the coverage and representativeness of NielsenIQ RMS to measure product innovation in the consumer goods sector.

Patent Data – Our main source of data for patent analysis is the USPTO data on the universe of published patent applications, granted or not. We use the original bulk data files provided by USPTO's Bulk Data Storage System for our analysis. Our sample

[6] Consumer Brands Association (2024) attributes 10% of the U.S. economy to the CPG industry in 2022.

initially contains information on more than 7 million patent applications filed by more than 500 thousand patent assignees in the years 1975-2017. For each patent, we use information about the patent application year, patent technology classifications, forward patent citations received, the number of claims on a patent, patent status (granted, pending, or abandoned in order to capture additional quality heterogeneity), and the type of patent (utility or design). We also extensively use text from patent applications, as we describe in more detail below. Appendix A.2 gives more detail about our sample and the variables we use.

2.2 Matching Algorithms

2.2.1 Matching Firms

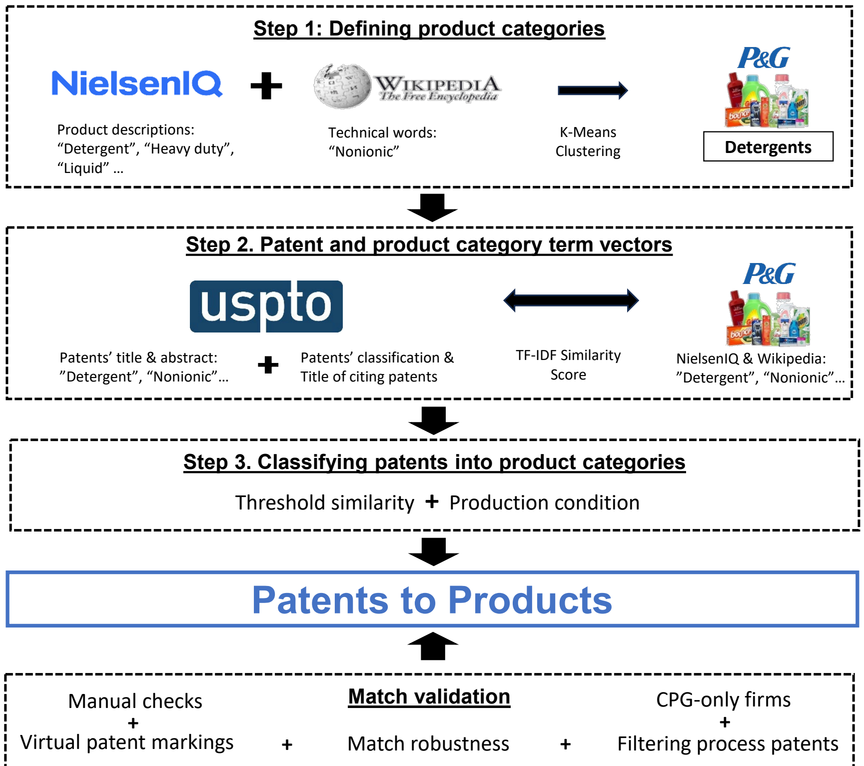
We match patents to products at the firm level using firm names from the patent and product data sets. For products, we obtain firm names from GS1 US, the official source of UPCs, which connects barcodes to the selling firm. For patents, we start with assignee names, typically the original patent holders, which may not necessarily reflect the current owners due to sales or reorganizations. We therefore combine USPTO re-assignment data with Thomson Reuters Mergers & Acquisition data to identify the most current patent holders. This step assumes that when firms merge or one acquires another, the surviving firm inherits all prior patents.

Firm names often differ across data sets due to formatting, abbreviations, or misspellings, complicating the matching. To address this, we developed a name-cleaning algorithm to standardize firm names, building on methods from the NBER Patent Data Project (Hall, Jaffe and Trajtenberg, 2001) and Akcigit, Celik and Greenwood (2016). Details are provided in Appendix A.3.

2.2.2 Patents-to-Products Match

To link patents to products, we use detailed product and patent descriptions and apply modern text analysis tools that are increasingly being used in economic research (Younge and Kuhn, 2016; Kelly et al., 2021; Webb, 2019; Kalyani et al., 2025). We proceed in three steps. In the first step, we group NielsenIQ products into product categories—a set of similar products, to which a patent can be linked. In the second step, we create a vector of terms describing product categories and patents. In the third step, similarity scores between each patent and every product category vector are computed to classify each patent into a product category and filter out patents unrelated to CPG products. Extensive validation and robustness checks of the procedure are then explored.

Figure 2: Patents-to-Products Match

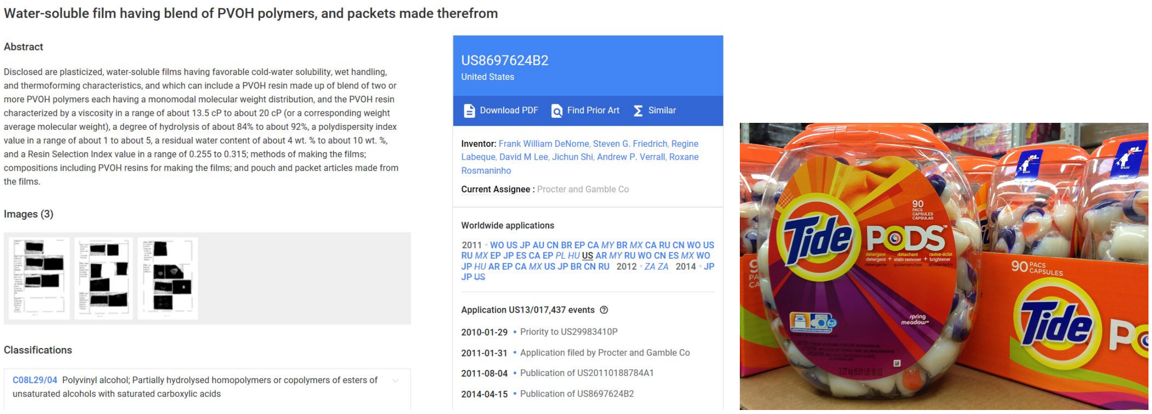


Notes: This diagram illustrates the steps undertaken for the patent-to-products match underlying the construction of the firm-product category-level data over time.

In what follows, we concisely describe each step with more technical details delegated to Appendix A.4. Figure 2 provides a schematic summary of the matching steps and validation exercises to further aid the reader. An example of a patent and its corresponding product introduction is shown in Figure 3.

Step 1. Defining Product Categories in the Product Data– Our first goal is to carve out well-defined sets of products in the NielsenIQ data—which we call product categories—that collect distinct and sufficiently large sets of similar products that would meaningfully relate to a distinct set of patents. NielsenIQ provides a classification, but its levels are either too detailed or too broad for our purposes. In the original data, each product belongs to one of 1,070 detailed product modules. These modules aggregate products with similar technological features, but very similar products may still fall into different modules. For example, "Detergents-packaged" and "Detergents-heavy duty-liquid" are distinct modules, though too similar to assign patents to one but not the other. By contrast, Nielsen also

Figure 3: Patent and Product: Procter & Gamble



(a) Patent application in 2011

(b) The first Tide Pods in 2012

Notes: Additional examples are presented in Appendix A.6.

provides a more aggregated scheme with 114 product groups, which are often too broad. For instance, "Bathroom scale," "Blender appliance," "Breadmaker appliance," "Vacuum and carpet cleaner appliance," and "Coffee and tea maker appliance" are all combined in the group "Housewares, appliances," despite being quite distinct. Hence, we aim to create an intermediate categorization—more aggregated than modules but less than groups—to associate patents with a well-defined set of products.

To this end, we aggregate product modules using a clustering procedure. We first expand the short module descriptions in NielsenIQ data with hand-collected Wikipedia articles. For each of the 1,070 modules, we manually assign one or two closely matching Wikipedia articles.[7] For example, "Detergent-packaged" and "Detergent-heavy duty-liquid" are both assigned to the "Laundry detergent" page, while "Cleaners-humidifiers/vaporizers" module is linked to the "Humidifier" and "Dehumidifier" entries. The use of Wikipedia to encode textual knowledge is common in the machine learning literature (e.g., BERT, fastText). Its advantage for our purposes is that it provides technical product descriptions and comprehensive texts for our analysis.

For each module, we then construct a representative document from the module title, the Wikipedia title, and the full article text.[8] We parse the text into single words and two-word phrases, apply lemmatization, and weight terms with the standard term-frequency-inverse-document-frequency (tf-idf) method (Aizawa, 2003), which downweights common words less

[7]The Wikipedia assignments were independently done by three research assistants, then cross-checked and finalized by the authors. The mapping is available upon request.

[8]Module and Wikipedia titles, as well as the first 10% of the article, are weighted 10 times more since they contain the most informative content.

useful for classification. We then l^2-normalize the resulting term-frequency vectors.

With representative vectors in place, we cluster modules using k-means (Lloyd, 1982). This partitions the vector space into clusters of similar products by minimizing within-cluster variance. As a baseline, we aggregate the 1,070 modules into 400 clusters, which we call product categories. This strikes a balance between merging very similar products and maintaining distinctiveness across categories. For instance, "Detergent-packaged," "Detergent-heavy duty-liquid," and "Detergent-light duty" form the "Laundry detergent" category, while "Cleaners-humidifiers/vaporizers" and "Humidifier and Vaporizer appliance" form the "Humidifiers" category.

Appendices A.4.1 and A.4.2 describe our NLP methods in detail, including clustering, quality checks, and robustness exercises. Appendix B.1 shows the sensitivity of our results to using NielsenIQ groups instead of our product categories, bypassing clustering altogether.

Step 2. Constructing term vectors for patents and product categories and corresponding similarity scores – Our goal is to assign patents to product categories using text analysis. To do so, we collect representative documents for patents and product categories and convert them into term-frequency vectors.

For patents, we use texts from the title, abstract, international patent classification description, and titles of cited patents. These fields are most informative about content and possible applications. We concatenate them into one document for each patent, then apply the same parsing, lemmatizing, tf-idf weighting, and normalization as before.

For product categories, we use the titles of all modules comprising the category and their corresponding Wikipedia articles to construct weighted, normalized term-frequency vectors. To improve matches to consumer products and filter out high-tech patents owned by NielsenIQ firms but outside our coverage, we add pseudo-product categories. These help prevent misclassification of unrelated patents.[9] After reviewing patents held by firms in our sample, we selected 19 diverse pseudo-categories (e.g., "computers," "software," "touch-screen") and created their respective term vectors from their Wikipedia descriptions.

We then compute similarity scores between the vectors of each patent and product category. For a patent p with vector f_p and a product category j with vector f_j, similarity is defined as the cosine similarity $s_{jp} = f_j \times f_p$. This measure lies in $[0, 1]$, with zero indicating no word overlap and one indicating identical texts. Appendix A.4.3 provides further details.

[9]For example, Samsung Electronics produces phones and TVs not covered by NielsenIQ; we want to exclude patents associated with these products. See Appendix A.4.2 for more details.

Step 3. Defining patent-product category match and non-matches – Next, we use the similarity scores between products and patents, and assign each patent to a product category or designate it as a non-match. Given that some patents may relate to general production processes rather than specific products or pertain to products outside the consumer goods sector NielsenIQ covers, we provide the option to classify a patent as a "non-match," meaning it is not assigned to any product category.

For each patent, we narrow the set of potential product categories to those with similarity scores above the threshold similarity or those ranked in the five highest similarity scores. We tested different threshold levels, and in our baseline algorithm, we set a threshold similarity of 0.025, which we use to filter out the patents unrelated to the products we consider. We further use NielsenIQ data to exclude categories where firms never sell any products (i.e., production condition). If the remaining set of potential categories is not empty, we assign the patent to the category with the highest similarity score. Otherwise, the patent is designated as a non-match.[10]

Our methodology assumes one product category match for each patent. However, some patents may be more general in nature, relating to multiple categories. Our baseline algorithm abstracts from this possibility. Nonetheless, our procedure for defining product categories is designed to ensure that they encompass a broad range of technically similar products, making it plausible for a patent to relate to only this range of products.[11] In Appendix A.4.4, we present more details of this procedure.

2.2.3 Match Statistics and Validation

Table 1 provides statistics of the data used in our analysis. We have annual data for all 34,665 firms that sold at least one product in our consumer goods sector data (CPG firms). The raw USPTO patent data cover information from 1975 to 2017, but because our product data only cover years from 2006 to 2015, our analysis can only consider annual variation from 2006 to 2015. In this period, the USPTO data include about 3.4 million patent applications in total and about 500 thousand patent applications filed by CPG firms. The firm $\times$ category data set resulting from the patents-to-products match includes 40% of those patent applications. The remaining 60% of patents, while filed by CPG firms, have been filtered out as they are not associated with products covered by our data, highlighting the importance of allowing for non-matches.

We perform an extensive set of validation exercises to evaluate the robustness and quality

[10] Appendix B.1 shows the sensitivity of our main findings to higher similarity thresholds.

[11] In this sense, the methodology delivers a many-to-many patent-to-product match, where each patent can be matched to multiple products of the firm.

Table 1: Match Statistics

	Period	
	1975-2017	2006-2015
Number of patent applications		
All assignees in USPTO	7,304,072	3,386,208
CPG firms: firm-level match	1,046,030	505,544
CPG firms: patents-to-products match	399,684	190,575
Number of firms		
All CPG firms		34,665
CPG with at least one patent applied in 1975-2017		5,209
CPG with a patent applied in 2006-2015		3,266

Notes: Match statistics for the baseline firm-level and firm × category level data sets.

of our patents-to-products match. We use four main types of validation exercises: manual checks, external validations using online-collected data on patent markings, analysis of the robustness of the algorithm-implied similarity scores and placebo tests, and validation of non-matches with respect to non-CPG products and process patents.

i. Manual checks – We manually verified many of the patent-to-products matches to evaluate whether our judgment on the best product category for a patent aligned with the algorithm's chosen category. Table A1 in the Appendix lists 100 patent applications from top-selling firms within the largest product categories, as per NielsenIQ's data. It is evident that the patent titles reflect the assigned product categories well. We also drew a random sample of 150 matched patents and examined their content. We found that about 80% of the cases presented a clear and convincing match, while the remaining cases displayed weaker or more indirect associations in our assessment.[12] In summary, although some noise is inevitable, our visual assessments reassure us of the match's adequate quality.

ii. External validation: Virtual patent markings – Next, we employ virtual patent markings to validate our matches. Some firms use virtual patent markings to notify the public that their product is patented by publishing the products and the corresponding patents protecting them online. Our website searches revealed that very few firms in our dataset utilized virtual patent markings. Even when they did, only a small subset of products and patents were included in the markings. Nevertheless, these data present a unique opportunity for externally validating our matching algorithm for a subsample of patents.

[12]The evaluation was carried out by a research assistant and the authors based on our best judgment. Given the complexity of many patents, some errors are unavoidable. The corresponding table is available upon request.

For Procter & Gamble (P&G) and Kimberly Clark (KC), we collected 400 virtual patent markings from the company websites and mapped the products listed on the websites to our product categorization. We then validate this markings-based patents-to-product category correspondence against our match. For the sample of matched patents, we see that for 79% of patents, the product category from virtual markings is the most or the second-most preferred product category based on our similarity scores. Appendix A.5.2 provides more details about the analysis.

iii. Robustness of the match and placebo tests – We evaluate the robustness of the product category choice by our matching algorithm to potential small perturbations in the algorithm. For the algorithm to be robust against small changes, we should observe that the highest-ranked product categories have substantially higher similarity scores with the patents than lower-rank product categories do. Appendix A.5.3 shows this is the case. Next, we verify that we are indeed carving out well-defined neighborhoods in the technological space by matching patents into distinct categories. For that, we compare the actual distribution of similarity scores between patents classified in the same product category versus a placebo group of patents drawn at random. Appendix A.5.4 shows that the distribution of similarity scores between pairs of patents within product categories is indeed very different and first-order stochastically dominates that of the placebo group.

iv. Validating non-matches – In the last step of the matching algorithm, multiple criteria were used to allow for the possibility that some patents filed by CPG firms were not directly associated with any of the consumer-good categories and, hence, would need to be filtered out. Indeed, as mentioned above, 60% of patents did not match using our preferred algorithm. A valid "non-match" can arise for two main reasons. First, a patent may relate to goods that the firm produces outside the CPG sector; second, a patent may be about a general process or method that does not strongly affect the introduction of products. We validate our non-matches against these two possibilities.

Non-CPG products–
Since our data do not cover products outside the CPG sector, we should expect our patents-to-products match to produce more non-matches for firms that more heavily produce non-CPG products. For example, we should expect patents of Samsung Electronics, which only has a subset of products present in our data (e.g., phone accessories and smaller consumer electronics), to have more non-matches than the patents owned by Procter and Gamble, whose product lines are entirely covered by Nielsen. Indeed, 87% of P&G patents and only 35% of Samsung patents match to our products. To systematically explore this pattern, in

the spirit of Hoberg and Phillips (2016), we use information from publicly traded companies' 10K reports to manually identify public firms whose output is mainly in the consumer-goods sector.[13] By classifying firms into CPG-only firms and those that are not CPG-only (selling many products outside NielsenIQ coverage), we find that 92% of patents of CPG-only firms match based on patents-to-products match. As expected, this share is lower for non-CPG-only firms (36%)—Figure 4. This provides additional support that the algorithm filters out irrelevant patents not associated with our product data.

Process patents–

Our match of patents to products should also filter out the general process and method patents that have a low association with product introductions. For example, some process patents may be aimed at reducing production costs and be less associated with product creation. Although distinguishing such general process patents from product-related patents is challenging, we use existing external procedures to proxy for process patents and compare them with the algorithm's non-matches. Firstly, we argue that the design patents are clearly product-related and should have the highest match rate in our patents-to-products match. Next, for utility patents, following Bena and Simintzi (2017), we use patent claims text to create proxies for process-related and product-related patents. We classify patents whose claims start with "methods" and "processes" as process patents, while the rest are product patents.[14] As shown in Figure 4, design patents–those classified as product patents with the highest certainty, have the highest match rate of 56%. This is followed by product patents at 41% and process patents at 33%–supporting the idea that process-related patents have the lowest association with product introduction.

Still, one may wonder why a sizable share of process-related patents match our products. This is not surprising considering that many patents with "method" and "process" claims pertain to product introductions rather than just improving general efficiency and cost reduction.[15] For example, Frito-Lay's "Method for reducing the oil content of potato chips" (ID 11777839) may lead to higher-quality, low-fat chips, and P&G's "Method for whitening teeth" (ID 13150392) may lead to the introduction of new teeth-whitening products.

Overall, these exercises offer reassurance that our algorithm successfully filters out patents unrelated to the products in our data.[16] This ensures that our analysis consistently

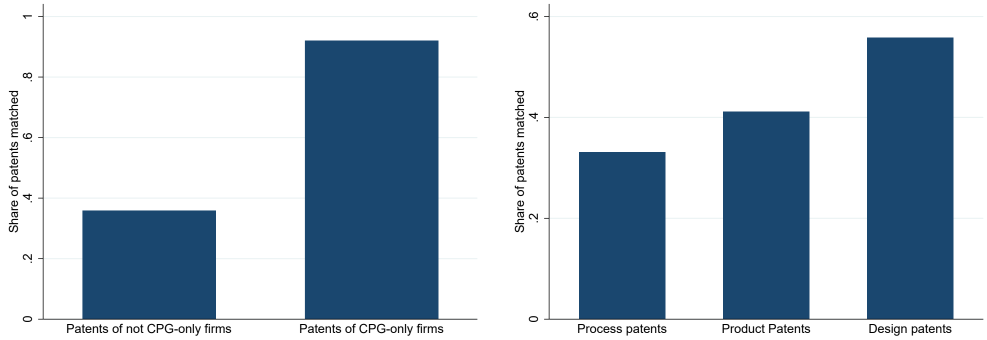
[13]We matched 270 publicly traded companies over our sample period and classified 23% of them as CPG-only firms. See Appendix A.5.5 for more details.

[14]Additional details about the exercise are presented in Appendix A.5.5.

[15]Tham, Baslandze, Sojli and Liu (2025) show that process improvements incentivize subsequent product innovation.

[16]Nevertheless, we also later show the robustness of our analysis, completely excluding these process-related patents with qualitatively similar results.

Figure 4: Match Validation. CPG-only Firms and Product-Related Patents



Notes: Panel (a) shows the share of patents that match with product categories in which firms ever sell a product. The left figure compares patents of the CPG-only and non-CPG-only firms, while the right figure compares process, product-related, and design patents. CPG-only firms and non-CPG-only firms refer to the sample of firms defined in Appendix A.1. Process and product-related patents are defined in Appendix A.2.

centers on product innovations and the patents most plausibly related to them, maintaining a consistent scope.

2.3 Measures of Product Introduction and Patenting

2.3.1 Product introduction

Our measures of product introduction are based on the number of products that firms introduce to the market and the quality improvements in those products. We use the product data described above to identify the entry dates of products in the market and their respective characteristics and performance. Our first innovation measure at the firm×category level is the number of **new products** of firm i (in product category j) in year t, as in Broda and Weinstein (2010) and Argente, Lee and Moreira (2018):

$$\mathrm{N}_{ijt} \equiv \sum_{u=1}^{T_{ijt}} \mathbb{1}[u \text{ is entrant}],$$

where product u is sold by firm i in product category j, T_{ijt} is the number of products that firm i sells in j as of period t, and $\mathbb{1}[u \text{ is entrant}]$ is an indicator that takes the value of one if u is a new barcode in t. This measure is simple and parsimonious but does not distinguish major product innovations from innovations that make relatively minor changes to a product's characteristics. In contrast to the previous literature, we construct the second set of measures of **quality-adjusted new products** that deals with this potential drawback by explicitly accounting for differences in characteristics across new products:

$$\mathrm{qN}_{ijt} \equiv \sum_{u=1}^{T_{ijt}} \mathrm{q}_u \mathbb{1}[u \text{ is entrant}],$$

where $\mathrm{q}_u \in [0, 1]$ is a measure of quality that we describe below. Together these two metrics allow us to account for differences in both the quantity and quality of product innovation across firms and over time.

Our baseline measure of product quality aims at capturing differences in novelty and economic impact across new products. We build on Argente and Yeh (2022) and use detailed information on product attributes that is available from the product data. Products can then be compared on the basis of characteristics associated with their attributes $\{v_{u,1}, ..., v_{u,A}\}$.[17] We test if each new product has characteristics distinct from those of all existing products available in the market, and we compute the quality of a new product as a weighted sum of its novel characteristics across all product attributes:

$$q_u \equiv \sum_{a=1}^{A} \omega_a \mathbb{1}[v_{ua} \text{ is new}].$$

where ω_a are weights that reflect the economic value associated with a particular attribute. We develop a novel approach to estimate weights that capture the importance of each attribute by using "shadow prices" from hedonic pricing regressions (Bresnahan and Gordon, 1996). The underlying assumptions are that the degree of novelty of a product should be reflected in its price, and that the price, in turn, captures the value of the product's embodied characteristics as determined by their respective shadow prices. A new product has a high novelty score if it has many new characteristics and/or if its characteristics are associated with high implicit prices. We provide details on the properties of this procedure in Appendix A.7, along with some evidence that the novelty score is strongly associated with the performance of the firm and its products.[18]

We use three alternative measures of new product quality to evaluate the robustness of our empirical results. First, we use a simpler version of the quality measure that weighs each attribute equally (quality *q1*). This measure only captures variation in the share of new product characteristics contained in a product. Second, we use a weighted quality

[17]For example, "children" and "regular" are two mutually exclusive characteristics associated with the attribute "formula" for "pain remedies-headache" products. Naturally, the number and type of attributes varies across product categories. For example, the product category "pain remedies-headache" includes 10 attributes: brand, flavor, container, style (i.e. children, regular), form, generic, formula (i.e. regular, extra strength, rapid release), type (i.e. aspirin), consumer (i.e. trauma, migraine), and size. On average, we observe that the different product categories include between 5 to 12 attributes. Appendix A.7 gives details.

[18]We show that our measure is correlated with the growth rate of the firm, the share of sales generated by new products, and the average duration of new products in the market even after conditioning on the number of products being introduced by the firm (Table A2 in the Appendix).

measure using weights that reflect "shadow sales" (quality *q2*). This measure assigns lower quality to new products that are associated with high shadow prices but do not reach many customers. Finally, we use a measure of residual demand taken from Hottman, Redding and Weinstein (2016) and Argente, Lee and Moreira (2024) (quality *q3*). This measure does not use information about the degree of novelty of a product and instead captures the appeal of new products relative to other products sold in the market, under some functional-form assumptions. Overall, our baseline measure and these alternative metrics allow us to consider many critical dimensions of the quality of new products and to assess the robustness of our results.

2.3.2 Patent Measures

Using an approach similar to how we measured product introduction, we compute measures that allow us to account for differences in the quantity and quality of patent applications across firms and over time. Our baseline measure is the number of **patent applications** (P_{ijt}). Using our patent-product category match, we are also able to measure the number of patent applications filed by firm i in product category j in year t as follows:

$$\text{P}_{ijt} \equiv \sum_{p=1}^{P_{it}} \mathbb{1}[p \text{ is match to } j].$$

Throughout the paper, we use information about whether a patent was granted and information about patent citation counts to compute our measures of patent quality. Patent applications that become **granted patents** (gP_{ijt}) are perceived as high-quality patents because the patent office deemed them novel enough to not be rejected. We compute the number of patent applications that are granted as:

$$\text{gP}_{ijt} \equiv \sum_{p=1}^{P_{it}} \mathbb{1}[p \text{ is granted}] \times \mathbb{1}[p \text{ is match to } j].$$

We also define **patent citations** (cP_{ijt}) as the total number of patents weighted by forward citations received in the first five years since the application was filed:[19]

$$\text{cP}_{ijt} \equiv \sum_{p=1}^{P_{it}} c_p \times \mathbb{1}[p \text{ is match to } j].$$

Measures based on forward citations have traditionally been used to assess the economic and technological significance of a patent (for earlier contributions, see Pakes (1986), Schanker-

[19] A 5-year citations measure attempts to reduce the truncation issue inherent to citations—the fact that patents filed more recently have had less time to accumulate citations (Hall et al., 2001).

man and Pakes (1986), Trajtenberg (1990)).

2.4 Summary Statistics

Table 2 provides summary statistics on the product- and patent-related variables for firms in our sample, grouped by their patenting activity. We split firms into three groups: (i) firms that never filed a patent application, (ii) firms whose last application was filed before 2006 (the start of the NielsenIQ data), and (iii) firms that filed a patent application between 2006 and 2015.

The share of patenting firms and the rate of product introductions in the consumer goods sector are comparable to those in other manufacturing sectors. More than 5,000 firms (15%) applied for at least one patent, and over 3,000 firms (9.5%) filed during 2006-2015. For comparison, Graham, Grim, Islam, Marco and Miranda (2018) link Census data to the USPTO and find that 6.3% of manufacturing firms had at least one granted patent between 2000 and 2011—a comparable statistic in our sample is 7.6%.[20] Table 2 shows that the average product introduction rate is 19%. While no equivalent comprehensive data exist for other sectors, Goolsbee and Klenow (2018) use Adobe Analytics data on online transactions and report product introduction rates comparable to those in other non-durable consumer manufacturing sectors.[21]

Firms with patent applications between 2006 and 2015 file more than six patents per year, on average. Because many patents receive no citations, especially in the first five years, the average number of citation-weighted applications (cP_{ijt}) is close to the raw number (P_{ijt}). These firms may hold some design patents, but the majority are utility patents. Patenting firms are on average larger: they sell more products, operate in more categories, and have higher sales.

Substantial innovation is associated with firms that never patented. Table 3 shows that 54% of new products were introduced by never-patenting firms. Accounting for novelty, about 65% of quality-adjusted product introduction comes from these firms.[22] On average, patenting firms introduce more incremental products, while non-patenting firms contribute a larger share of novel innovations. Since these statistics rely on firm-level matches, they implicitly attribute all new products of a patenting firm to its patents. Highly diversified

[20]These patenting rates are at least three times higher relative to the entire US economy (Graham, Grim, Islam, Marco and Miranda, 2018; Mezzanotti and Simcoe, 2025).

[21]Goolsbee and Klenow (2018) show that some durable goods (e.g., furniture), not covered in our dataset, have higher entry rates than non-durables (e.g., food).

[22]This result holds across different quality adjustments. For example, never-patenting firms account for 65% of $q1N$ and 77% of $q2N$. For $q3$, no good counterpart to $q3N$ can be constructed, but Table 2 shows that $q3$ is not necessarily higher for patenting firms.

Table 2: Summary Statistics by Firm's Patenting Status

	No Patents	Patents before 2006	Patents 2006-2015
Product data			
Number of products	15.49	31.08	78.35
Number of new products (N)	2.58	5.26	13.45
Average quality of new products (q)	0.27	0.20	0.20
Quality-adjusted number of new products (qN)	0.46	0.62	1.48
Product introduction rate (n)	0.19	0.17	0.22
Quality-adjusted product introduction rate (qn)	0.07	0.04	0.06
Sales from all products	2371.59	9392.09	37094.71
Sales from new products	454.74	1811.01	8130.00
Number of product categories	2.36	3.07	5.46
Average quality of new products ($q1$)	0.13	0.10	0.10
Average quality of new products ($q2$)	0.18	0.11	0.12
Average quality of new products ($q3$)	0.06	0.32	0.10
Patent data			
Number of patent applications (P)	0.00	0.00	6.34
Number of granted patent applications (gP)	0.00	0.00	4.57
Number of citations-weighted patent applications (cP)	0.00	0.00	5.88
Stock of patent applications	0.00	11.33	125.36
Stock of granted patent applications	0.00	11.02	107.63
Stock of citations-weighted patent applications	0.00	17.97	215.24
Number of firms	29215	1943	3266
Observations	186934	15803	29052

Notes: The table shows averages of product-based and patent-based variables of the firm-level dataset. The first column groups firms with no patents; the second column includes firms that patented before appearing in NielsenIQ RMS (before 2006); the third column covers firms with patents in 20062015. Product introduction statistics can be computed only for 20072015 because entries for products introduced in 2006 cannot be determined (left-censoring). Sales are reported in thousands of dollars, deflated by the CPI for all urban consumers. Patent statistics are highly skewed; averages are reported after winsorizing patent variables at the top 0.1%.

firms may patent in one category while introducing unrelated products in others, overstating the role of patents in product introduction. This highlights the importance of establishing a closer patent-product link. We therefore replicate the exercise at the firm × category level. Table 3 shows that firms that never patented in a category account for the majority of new products in that category.

We also decompose sectoral growth from 2006 to 2015 by firms' patenting status:

$$\underbrace{\text{Growth}_{\text{`}06-\text{`}15}}_{7\%} = \underbrace{\text{Growth}^{\text{Patent}}_{\text{`}06-\text{`}15}}_{4\%} \times \underbrace{s^{\text{Patent}}_{2006}}_{0.72} + \underbrace{\text{Growth}^{\text{No Patent}}_{\text{`}06-\text{`}15}}_{14.4\%} \times \underbrace{s^{\text{No Patent}}_{2006}}_{0.28} \tag{1}$$

where *Growth* refers to sales growth, and s^{Patent}_{2006} and $s^{\text{No Patent}}_{2006}$ are sales shares of firm×categories with or without patents.[23] These decompositions show that although non-patenting firms

[23]We first write $Rev^{CPG}_t = \sum_j \sum_{i\in\Omega^j_{\text{Patent}}} Rev_{ijt} + \sum_h \sum_{i\in\Omega^j_{\text{No Patent}}} Rev_{ijt}$, where the second sum is across categories and Ω denotes the set of firms with and without patents in category j. Taking percentage

Table 3: Share of New Products Accounted for by Patenting Firms

	New Products, N	Quality-adjusted New Products, qN
Match 1 (firm-level)		
Firms with patents in 2006-2015	0.38	0.28
Firms with patents before 2006	0.08	0.07
Firms with no patents	0.54	0.65
Match 2 (patents-to-products)		
Firm × category with patents in 2006-2015	0.23	0.16
Firm × category with patents before 2006	0.07	0.05
Firm × category with no patents	0.71	0.79

Notes: The table shows the share of product innovation measured by our two benchmarksproduct introduction (column 1) and quality-adjusted product introduction (column 2)accounted for by firms and firm×categories with or without patents.

are smaller and account for less sales, they contribute more to aggregate growth–58% of sectoral growth.

Our data cover categories with substantial heterogeneity in entry rates and patenting intensity. We classify dry grocery, frozen foods, dairy, deli, packaged meat, fresh produce, and alcoholic beverages as food categories. Non-food categories include health and beauty (including over-the-counter drugs), non-food grocery, and general merchandise (cookware, electronics, household supplies). Appendix Figure B1 shows that food and non-food categories have similar entry rates but distinct patent intensities. Patenting firms and patents per product are more prevalent in non-food categories. It is therefore not surprising that many product introductions, particularly in food, are not directly linked to patents. Some products represent only minor upgrades and may not meet the patentability requirement of "novelty and non-obviousness," so raw patent counts naturally miss such incremental innovations.

3 Patents and Product Innovation on the Market

Product innovation—the introduction of new and improved products to the market—is a key contributor to economic growth and a central aspect of endogenous growth models (Romer, 1990; Grossman and Helpman, 1991). Absent the direct innovation measures, the researchers have relied on indirect inference (using employment and sales) or patent metrics to proxy for firm and aggregate innovation in the market. The use of patents to proxy for firm innovation has a long tradition and is supported by studies showing a positive relationship between patents and a firm's stock market value, growth, and productivity (Hall, Thoma

changes in sales yields (1).

and Torrisi, 2007; Balasubramanian and Sivadasan, 2011; Kogan, Papanikolaou, Seru and Stoffman, 2017). However, skeptics of patent metrics argue that patents are a tool for legal protection against competitors, and the prevalence of strategic patents, especially in recent decades, makes it hard to infer how well these measures reflect true innovations in the market (Cohen, Nelson and Walsh, 2000; Jaffe and Lerner, 2004; Boldrin and Levine, 2013).

Leveraging our novel match between patents and product introductions, we directly evaluate how patenting relates to actual innovation in the market. Since product innovation is a key driver of firm growth in the consumer goods sector (Hottman, Redding and Weinstein, 2016; Argente, Lee and Moreira, 2024), and our data provide detailed product-level information, we focus primarily on the link between patents and product introductions, while also discussing process innovation in Sections 2.2.3 and 3.2.1. The richness of our data further allows us to examine heterogeneity in this relationship across product types and firm characteristics. We find that patent filings are, on average, positively associated with product introductions, with the strongest relationship observed at a one-year lag following the patent application. The association persists over several years and is notably stronger in non-food product categories. These results imply that, on average, patent-based measures capture meaningful variation in realized innovation.

However, the relationship between patents and product innovation weakens with firm size. As firms grow, they file more patents relative to the number of products they introduce, resulting in a lower observed conversion rate of patents into market innovations. Extensive robustness checks demonstrate that this weaker patent-product introduction link among larger firms is not driven by delayed commercialization, misclassified process-related patents, data coverage issues, international product launches, or size-dependent measurement error in our text-matching algorithm. Additional evidence from patent characteristics further supports this pattern: large-firm patents tend to be less novel, receive fewer forward citations (especially from other firms), and are more likely to be litigated. Taken together, these results indicate that patent-based measures are weaker proxies for realized market innovation for larger firms—patterns consistent with more prevalent strategic patenting, in which firms file for protection without corresponding market innovation.

3.1 Baseline Estimates

To estimate the relationship between patents and subsequent product introduction by firms, we use the following baseline specification at the firm × category level over time:

$$Y_{ijt} = \beta\, P_{ijt-1} + \alpha_{ij} + \gamma_{jt} + u_{ijt} \tag{2}$$

where Y_{ijt} is the product introduction rates for firm i in category j in year t, and P_{ijt-1} is the patenting rates by the firm i in category j in year $t-1$. Product introduction rate is measured as the number of new products or quality-adjusted new products over the total number of the firm's products; patenting rates are measured as the number of patent filings (all, granted, or citation-adjusted) over the cumulative stock of patents, net of depreciation.[24] The key coefficient of interest is β that captures how product introduction activities relate to patenting activities.

By using firm × category × year-level data, we control for a wide range of potential confounding factors. The inclusion of product category-year fixed effects, γ_{jt}, absorbs category-specific shocks such as changes in market-wide demand or cyclical fluctuations, including those affecting particular categories. Firm-category fixed effects, α_{ij}, capture persistent differences across firm-category pairs, including variation in market power, intellectual property strategies, or innovation capabilities. This rich set of fixed effects ensures that our estimates are not confounded by systematic differences in patentability or data coverage across product categories, nor by firm-specific, time-invariant predispositions to patent—such as lower patenting/enforcement costs due to firm size, accumulated experience, or specialization within a category. Therefore, the coefficient β can be interpreted as capturing withinfirm-category variation in product introductions that is associated with recent changes in patenting activity, net of both market-wide time-varying factors and firm-specific time-invariant influences.

Table 4 shows the estimates of equation (2) for both product introduction and quality-adjusted product introduction rates. The rows present results based on patent applications, granted patents, and citation-adjusted patents. The estimates reveal a significant positive relationship between product introduction—both in quantity and in quality-adjusted terms—and prior patenting activity. On average, a ten–percentage-point increase in a firm's patenting rate within a given product category is associated with roughly a 0.4–percentage-point increase in its subsequent product introduction rate (or 0.2 percentage points for the quality-adjusted measure). This relationship is somewhat stronger when using granted and

[24]Using rates for product innovation and patenting removes scale effects, but in Appendix Table B1, we also show similar results in (log) levels.

Table 4: Product Introduction on Patenting Rates

	Product Introduction			Product Introduction Quality-Adjusted		
	(1)	(2)	(3)	(4)	(5)	(6)
Patents(t-1)	0.0445***			0.0174***		
	(0.008)			(0.003)		
Patents granted(t-1)		0.0469***			0.0200***	
		(0.009)			(0.003)	
Patents citations adj.(t-1)			0.0562***			0.0264***
			(0.013)			(0.005)
Observations	409,641	409,641	408,493	409,641	409,641	408,493
R-squared	0.357	0.357	0.358	0.302	0.302	0.303
Time-Category	Y	Y	Y	Y	Y	Y
Firm-Category	Y	Y	Y	Y	Y	Y

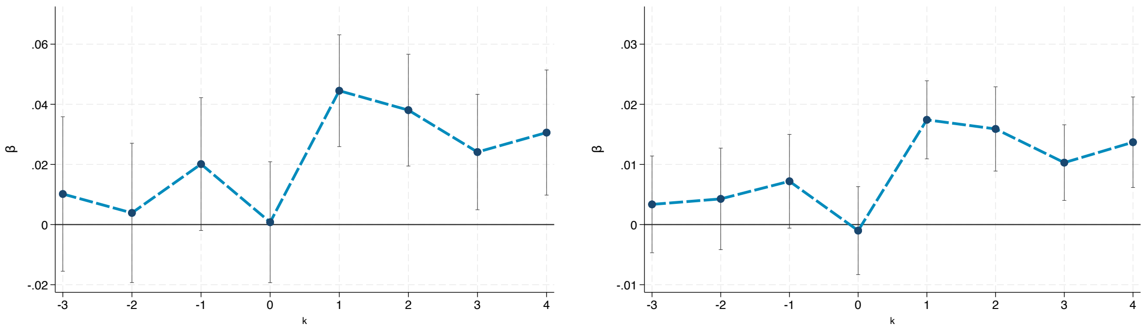
Notes: The table shows regressions of the product introduction rates and quality-adjusted product introduction rates as a function of the patenting rates, using firm × category × year data. Product and quality-adjusted product introduction rates are defined as the number of new products or quality-adjusted new products over the total number of products in the firm × category × year. Product quality measures are defined in Section 2.3.1. *Patents* is the ratio of the firm's number of patent applications in a particular category-year over the total number of cumulative patents in that category-year, net of depreciation; *Patents granted* is the ratio of the firm's number of granted patent applications in a particular category-year over the total number of cumulative patents in that category-year; *Patents citation-adjusted* is the ratio of the firm's number of citations-weighted granted patents in a particular category-year over the total number of citation-weighted granted patents in that category-year. A standard 15% annual depreciation rate is used (Hall et al., 2005). Observations at the firm × category × year level with zero patents are included in the regression. Standard errors are clustered at the firm × category level.

citation-adjusted patents. Hence, our results support the view that granted and citation-adjusted patent measures better capture meaningful variation in realized new product innovations in the market.

We also explore the association between product introduction and patents for various product categories in Appendix Table B2. Our analysis indicates that the association between product introduction and patents holds across both food-related and non-food categories—including health and beauty care, non-food grocery, and general merchandise—and is comparatively stronger for the latter.

Exploring different lag structures in our baseline regression specification can also help us gain insights into the dynamic relationship between patent filings and product introduction. Our baseline specification uses a one-year lag between patent filing and product introduction to account for the fact that it may take longer for firms to develop and commercialize a new product after they apply for patents. Figure 5 plots the estimated coefficients for different lags, where k refers to the lag (in years) between patenting and product introduction. We find the strongest positive association between patent filings and product introduction with one and two-year lags, with lower but persistent effects in subsequent years.

Figure 5: Product Introduction and Patenting Rates: Dynamics



(a) Product-to-patent introduction, β_k (b) Q-adj. product-to-patent introduction, β_k

Note: The figure plots the estimated coefficients after estimating equation $Y_{ijt+k} = \beta_k P_{ijt} + \alpha_{ij} + \gamma_{jt} + u_{ijt+k}$, $k = -3, ..., 0, ..., 4$ for product introduction rates n in (a) and quality-adjusted product introduction rates qn in (b). Product and quality-adjusted product introduction rates are defined as the number of new products or quality-adjusted new products over the total number of products in the firm × category × year. Product quality measures are defined in Section 2.3.1. The variable P is the ratio of the firm's number of patent applications in a particular category-year over the total number of cumulative patents in that category-year, net of depreciation. The regression includes firm × category and category × year fixed effects. Standard errors are clustered at firm × category. The vertical bands represent +/- 2.45 × st. error of each point estimate.

3.2 Firm Size Heterogeneity

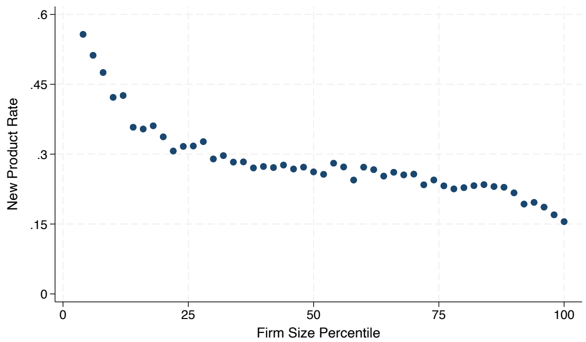
Next, we analyze how patents-to-product introduction patterns vary by firms' market positions. We begin by examining product introduction and patenting in relation to firm size within a market. Panels (a) and (b) of Figure 6 plot the average product introduction rates—the ratio of new products and quality-adjusted new products over a firm's stock of existing products—for firms in different firm size percentiles. Firm size percentiles for each category are defined based on the distribution of the average firm-category sales. Market leaders—larger firms within product categories, have lower product innovation rates. On average, firms in the top sales quintile have annual introduction rates of about 20%, while firms in the bottom quintile have rates more than twice as large. Larger firms do not compensate for this decline in the rate of new product introduction with innovations of higher quality. On average, firms in the top sales quintile have quality-adjusted product introduction rates of 9.7%, while firms in the bottom sales quintile have rates of approximately 16%.[25]

Panels (c) and (d) show that larger firms, on average, file more patents relative to new products they introduce. In particular, both the ratio of new patents filed over quality-adjusted new products introduced (Panel (c)) and the ratio of patenting rate over quality-adjusted product introduction rate (Panel (d)) both increase with firm size in the market.[26]

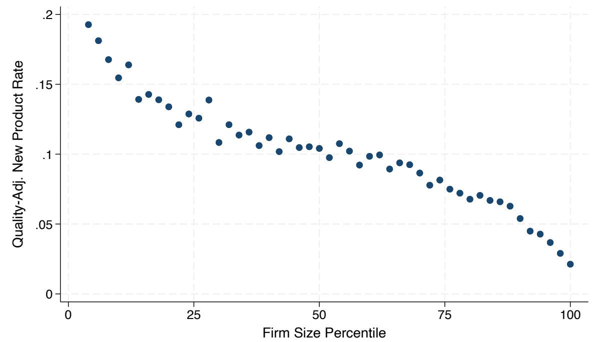
[25]Figure B2 in the Appendix confirms similar patterns using alternative product quality metrics.

[26]If we do not scale our measures of patenting, results are even starker: the unconditional probability of patenting and the total number of patents filed by large firms are much higher than they are for small firms

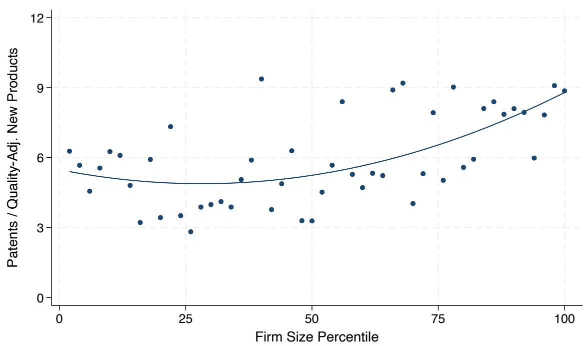
Figure 6: Product Introduction and Patents-to-Products by Firm Size



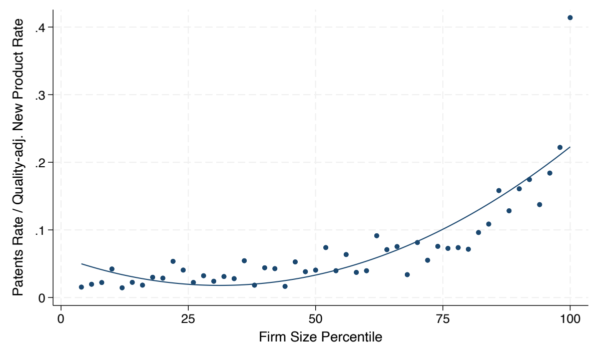
(a) Product Introduction Rate



(b) Quality-Adj. Product Introduction Rate



(c) Patents-to-Quality-Adj. New Products



(d) Patent Rate-to-Quality-Adj. New Product Rate

Note: These figures plot the production innovation rates and patents per innovation measures over firm size percentiles. For each firm × product category, we compute average sales and define firm size percentiles based on the average sales distribution in that product category. Each panel plots the average value of the respective variables in each percentile. Panel (a) shows the average product introduction rate (new products divided by the total number of products sold); Panel (b) shows the quality-adjusted product introduction rate (quality-adjusted new products divided by the total number of products sold); Panel (c) shows the ratio of patent applications (×1000) per quality-adjusted new products; and Panel (d) shows the patenting rate over quality-adjusted product introduction, where the patenting rate is calculated including observations at the firm × category × year level with zero patents.

Note that this higher intensity of patenting activity relative to product introduction is not explained by the possibility that larger firms introduce fewer but more novel products as the presented product introduction numbers are adjusted for their quality.

The patterns above suggest that the relationship between product introduction and patenting changes with firm size. We now, more formally, explore how the relationship varies with firm size by estimating equation (2), interacting patenting with firm size (firm sales in product category). As before, we control for time × product category and firm × product category fixed effects to ensure that potential confounders, such as differences in patentability across firms and product categories, do not drive our results. Table 5

(see Figure B3 in the Appendix).

Table 5: **Product Introduction on Patenting: by Size**

	(1)	(2)	(3)	(4)
	Product Introduction		Product Introduction Quality-Adjusted	
Patents(t-1)	0.0445***	0.0860***	0.0174***	0.0459***
	(0.008)	(0.023)	(0.003)	(0.008)
Size(t)		0.0102***		0.0012***
		(0.000)		(0.000)
Patents(t-1) x Size(t)		-0.0037**		-0.0025***
		(0.002)		(0.001)
Observations	409,641	409,641	409,641	409,641
R-squared	0.357	0.362	0.302	0.303
Time-Category	Y	Y	Y	Y
Firm-Category	Y	Y	Y	Y

Notes: The table shows regressions of the product introduction rates and quality-adjusted product introduction rates as a function of the patenting rates, using firm × category × year data, similar to Table 4 but introducing size (firm sales in category-year) and size interaction with patenting. Product and quality-adjusted product introduction rates are defined as the number of new products or quality-adjusted new products over the total number of products in the firm × category × year. Product quality measures are defined in Section 2.3.1. *Patents* is the ratio of the firm's number of patent applications in a particular category-year over the total number of cumulative patents in that category-year, net of depreciation. Observations at the firm × category × year level with zero patents are included in the regression. Standard errors are clustered at the firm × category.

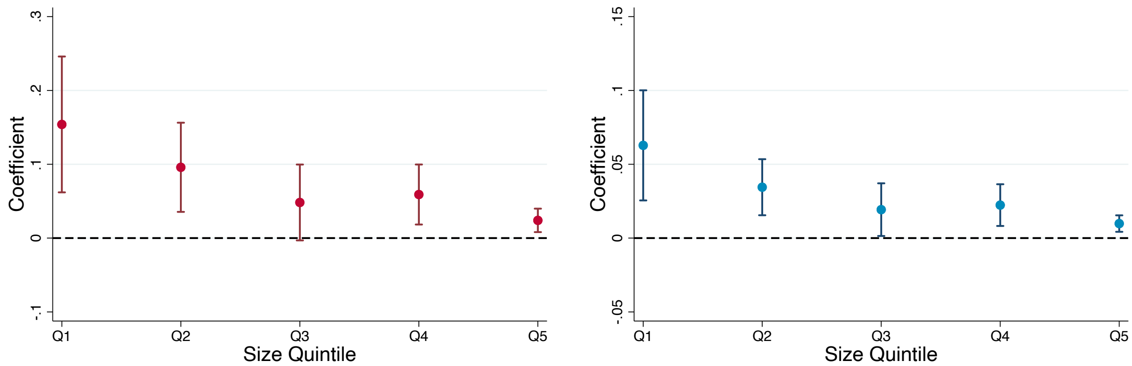
shows the negative interaction term with size for both product introduction and quality-adjusted product introduction rate.[27] Figure 7, in addition, visualizes the differences in patents-to-product introduction relationship across different-size firms. Panel (a) plots the coefficients from estimating equation (2), interacting patenting rate with firm size quintile dummies. As before, firms are ranked within product categories based on their average sales within each category. While the patents-to-product introduction coefficient is 0.15 for the lowest quintile of firms, it is 6.4 times smaller for the firms in the top quintile. A similar comparison holds for quality-adjusted product introduction regressions from Panel (b). Finally, we also demonstrate that this relationship remains robust across different types of products. Appendix Figure B4 confirms that the link between patent filing and product introduction is weaker for larger firms in both food and non-food categories.

3.2.1 Declining Patents-to-Innovation Relation with Size: Robustness

In this section, we show that the observed decline in the patents-to-innovation relationship with size is not driven by our matching algorithm, data coverage, or econometric specifications, indicating that this is a robust pattern in the data: patents held by larger firms are less likely to translate into product innovations in the market. We first examine several

[27]Appendix Table B3 shows similar results with different quality adjustments.

Figure 7: Product Introduction on Patenting: by Size



(a) Product Introduction Rate (b) Quality-Adj. Product Introduction Rate

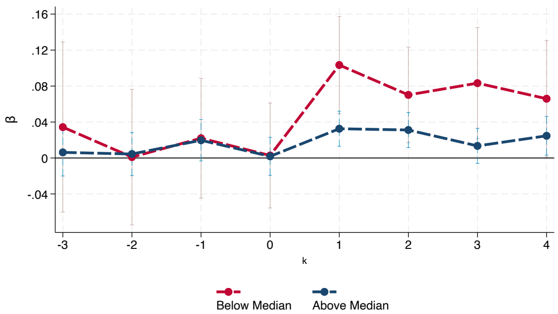
Notes: The figure plots the β_k coefficients from estimating the regressions $Y_{ijt} = \sum_{k=1}^{5} \beta_k \, P_{ijt-1} \times Q_{ijk} + \alpha_{ij} + \gamma_{jt} + u_{ijt}$, where Y_{ijt} is product introduction rate (Panel (a)) and quality-adjusted product introduction rate (Panel (b)) of firm i in product category j in year t; P_{ijt-1} is the patenting rate of firm i in product category j in year $t-1$; Q_{ijk} are dummies equal to one if the firm i's average sales in product category j are in the k^{th} quintile of firm sales distribution in j. The regression includes firm $\times$ category and category $\times$ year fixed effects. Standard errors are clustered at the firm $\times$ category level.

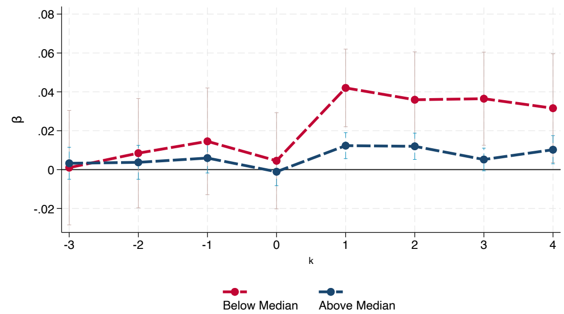
potential confounding explanations for the weakening of the patents-to-product relationship with firm size: differences in the timing of product introductions by firm size, mismeasurement of process patents, differential data coverage, and size-dependent measurement error. We then conclude by presenting additional evidence from patent characteristics that further reinforces the empirical pattern that patents play a diminishing role in supporting product innovation at larger firms.

Timing of product introductions by size. Our baseline results explored the patents-to-product innovation relation with a one-year lag; however, if larger firms take longer to commercialize their inventions—for example, due to conducting more experimental research—this could weaken the contemporaneous association between patents and product introductions for them. To assess this possibility, we use the dynamic specifications of equation (2) for large and small firms. Figure 8 shows that the patents-to-products relation for smaller firms–those with average revenue below the median revenue in their respective category—is stronger across all time lags. This suggests that, at least for the sample frame in our data, we do not find evidence that patents held by larger firms are associated with product innovation with a longer delay.

Process innovations by size. One possible concern with interpreting the decline in regression coefficients with firm size as evidence of a weaker relationship between patents and innovation among larger firms is the potential inclusion of process patents. Specifically,

Figure 8: Product Introduction and Patenting Rates Dynamics: By Size





(a) Product-to-patent introduction, β_k (b) Q-adj. product-to-patent introduction, β_k

Note: The figure plots the estimated coefficients after estimating equation $Y_{ijt+k} = \beta_k P_{ijt} + \alpha_{ij} + \gamma_{jt} + u_{ijt+k}$, $k = -3, ..., 0, ..., 4$ for product introduction rates n in (a) and quality-adjusted product introduction rates qn in (b). Product and quality-adjusted product introduction rates are defined as the number of new products or quality-adjusted new products over the total number of products in the firm × category × year. Product quality measures are defined in Section 2.3.1. The variable P is the ratio of the firm's number of patent applications in a particular category-year over the total number of cumulative patents in that category-year, net of depreciation. The regression includes firm × category and category × year fixed effects. Standard errors are clustered at the firm × category level. The vertical bands represent +/- 2.45 × st. error of each point estimate.

some patents may focus on cost reduction or improving production efficiency rather than directly promoting product innovation. If these process patents, which are unrelated to product innovation, are disproportionately utilized by larger firms (Cohen and Klepper, 1996) and inadvertently included in our analysis, this could explain the observed decrease in regression coefficients for larger firms.[28]

We address this concern in several ways. First, we note that our algorithm explicitly filters out general process and method patents that are not product-specific, increasing the likelihood that matched patents correspond to products in our data. Recall that in validating our matching algorithm (Section 2.2.3), we use independent proxies for product- and process-related patents drawn from claims texts, following Bena and Simintzi (2017),[29] and show that the algorithm successfully filters out the unrelated process patents while retaining those plausibly tied to product introductions (e.g., "method for whitening teeth").

Second, because the margin of error in filtering out unrelated process patents may vary with firm size, we examine the share of process-matched patents—according to our algorithm—that are also classified as process-related by Bena and Simintzi (2017), and assess how they evolve with firm size. We find no systematic relationship between the share

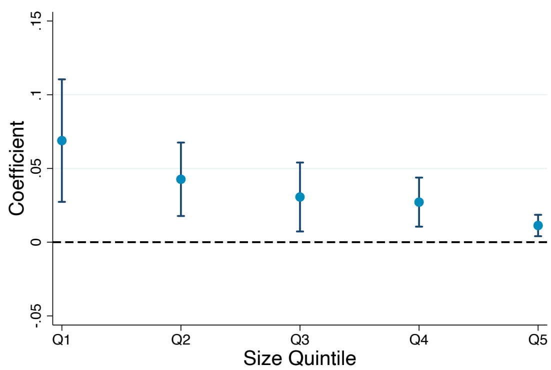
[28] A related concern is that larger firms may file patents primarily for licensing rather than for their own product commercialization. While our data attribute patents to the original holders and lack comprehensive information on temporary licensing agreements, prior research (Fosfuri, 2004; Gambardella, Giuri and Luzzi, 2006) finds that larger firms are less likely to license out their patents–if anything, reinforcing the observed decline in the patents-to-product relationship with firm size.

[29] About 25% of the matched patents are classified as process patents using this classification method.

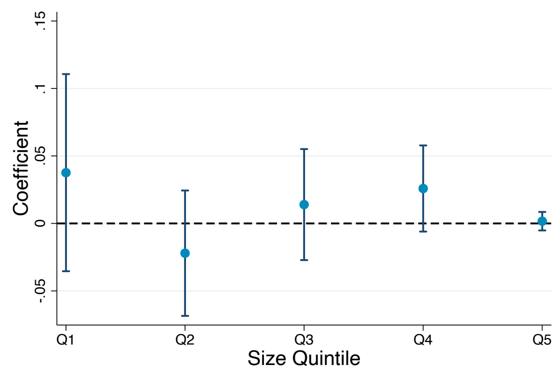
of process-related patents in a firm's portfolio and firm size (Appendix Figure B5).[30]

Third, we entirely exclude patents classified as process-related using the above proxy and repeat our analysis. Panel (a) of Figure 9 demonstrates that the positive relationship between patents and products, along with its diminishing gradient as firm size increases, remains robust when these process-related patents are omitted. Panel (b), which includes only process-related patents, shows that these patents do not have a statistically significant association with product introduction.

Figure 9: Product Introduction and Patenting by Size: Process and Product Patents (Quality-Adjusted)



(a) Product Patents



(b) Process Patents

Notes: The figures use a specification equivalent to Figure 7, for quality-adjusted product introduction rates, splitting matched patents into process- and product-related patents based on classification in Bena and Simintzi (2017). The left panel plots regression coefficients including only product-related patents. The right panel plots regression coefficients including only (matched) process-related patents. Standard errors are clustered at the firm $\times$ category level. Appendix Figure B6 reports corresponding results using unadjusted for quality product introduction rates.

Finally, if the weaker association between process patents and product introductions among large firms reflects a greater focus on cost-reducing innovations, we would expect this to manifest in price dynamics. To test this, we examine whether process patenting by larger firms is associated with lower price growth. Appendix Figure B7 plots coefficients from regressions of firm-product category-year average prices on the patent introduction rate, interacted with firm size quintiles. The specification includes firm-category and category-year fixed effects, mirroring the approach in Figure 7. We find no evidence that patenting leads to subsequent price declines within firms—nor is such an effect present specifically among larger firms.

Data coverage by size. Another potential concern is that the relatively weaker as-

[30] We obtain similar results using the process patent classification from Tham, Baslandze, Sojli and Liu (2025). These additional results are available upon request.

sociation between patents and product introduction of larger firms that we estimate could be explained by differences in data coverage across firms of different sizes. The patent data set encompasses the entire portfolio of USPTO patents held by firms, whereas our product data set does not cover products outside the consumer goods sector or products sold internationally. This could, in turn, result in lower rates of patent attribution to new products in our data for firms that also operate in other sectors or internationally.

To address the concern about products outside the CPG sector, we first note that, as discussed earlier, our empirical findings are based on the patents-to-products match, which filters out patents unrelated to the consumer goods sector, reducing this concern. However, we also reestimate our results on a smaller sample of firms that sell exclusively CPG products. Appendix Figure B8 shows broadly similar patterns of the weakening relationship between patents and product innovation for larger firms, albeit the estimates are more noisy because of a smaller sample size.

Internationalization may also contribute to differential data coverage by firm size. Patent protection is always local and only pertains to the jurisdiction in which it is filed. Larger firms, in particular, may introduce products abroad that are not captured in our product data, even if they file patents in the U.S. To investigate this, we incorporate two new datasets. First, we merge our data with the EPO's PATSTAT data to capture international patent filings.[31] About 60% of patents in our sample are filed in countries beyond the U.S., with the average patent filed in 3.2 countries. This suggests that NielsenIQ firms actively protect intellectual property internationally and likely sell abroad. However, controlling for the share of international patents does not affect our baseline results, and this variable is not statistically significant in explaining U.S. product innovation rates (Appendix Table B4). This finding aligns with our analysis of the Mintel Global New Products Database (MNPD), which tracks product launches across major retail channels worldwide.[32] We find that most products launched abroad by top global firms (as identified by NielsenIQ) are also introduced in the U.S., typically within 1.2 years of their initial international launch. This suggests that our product data provides a reasonably comprehensive view of global product introductions. Thus, the weakening patents-to-product introduction relationship for larger firms is unlikely to be driven by international launches that are missing from our data.

Measurement error by size. Lastly, we also assess the possibility that our textual analysis of patents inadvertently weakens the relationship between patents and products of large firms. A potential concern is that the text of patents filed by firms of different sizes may be systematically different, and that our matching algorithm may be less effective

[31] We thank Francesca Lotti for sharing the data with us.
[32] See Argente, Oberfield and Van Patten (2025).

in ascribing patents filed by larger firms to specific product categories. To better gauge this concern, we study the textual characteristics of patents, including patent document length, number of unique words, textual diversity, and the relative entropy of patents' word distributions. We evaluate whether these characteristics vary systematically across firms of different sizes within the same product categories. We do not find systematic differences in the textual characteristics of patents filed by large and small firms in these metrics. Furthermore, we also do not observe significant differences in the share of matched patents across firm size (see Figure B9 in the Appendix for details). Overall, our exercises suggest that differences in data coverage and the properties of the matching algorithm are unlikely to explain the weaker association between patents and product innovation for large firms.

Additional evidence from patent characteristics. We provide additional evidence that larger firms' patents tend to be less novel and scientifically valuable, but more likely to be associated with strategic motives. We compare patent characteristics of market leaders, the largest firms in each product category, to other firms in the same category. Table 6 presents cross-sectional comparisons of average firm-level patent characteristics. Odd columns show raw comparisons; even columns add controls for the number of firms in the category and firm fixed effects for multi-category firms.

Columns 1-2 focus on text-based novelty (similarity to prior filings), showing that leaders' patents are significantly less novel.[33] Columns 3-4 show leaders' patents receive fewer forward citations, suggesting weaker follow-on innovation. Columns 5-6 show a smaller share of those citations come from competitors, implying limited spillovers. Columns 7-8 show leaders' patents are somewhat more likely to involve litigation.[34] Together, these differences in patent characteristics reinforce the empirical pattern that patents play a diminishing role in supporting product innovation at larger firms.

3.2.2 Discussion

There are several potential reasons why the relationship between patents and innovation weakens with firm size. Larger firms may have stronger benefits from using patents to defend their market position and deter competition, without necessarily commercializing the patented ideas. They may also face lower marginal costs of patenting and enforcement due to scale economies in R&D, in-house legal teams, and litigation resources.[35] Overall,

[33]This measure applies only to firms with multiple patents.

[34]Litigation is rare–0.04 share of patents–but economically meaningful. The consumer products sector led patent litigation from 1998-2017, surpassing biotech, pharma, and electronics (PwC, 2018).

[35]As discussed above, our specifications include a rich set of fixed effects that absorb time-invariant predispositions to patent—such as lower baseline patenting or enforcement costs stemming from firm size, accumulated experience, or category-specific specialization. These controls imply that such scale economies

Table 6: Patent Characteristics: Leaders vs Followers

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Text Novelty		Citations		Share Cited Others		Share Litigated	
Leader	-0.042***	-0.033***	-0.538**	-0.675***	-0.062**	-0.076**	0.004	0.008**
	(0.008)	(0.008)	(0.247)	(0.257)	(0.027)	(0.030)	(0.005)	(0.003)
Observations	2,633	1,830	2,179	1,526	1,569	1,095	2,179	1,526
R-squared	0.015	0.405	0.002	0.247	0.003	0.335	0.000	0.094
Controls	N	Y	N	Y	N	Y	N	Y

Notes: The table compares the average patent characteristics of leaders and the other firms. "Leader" is a dummy equal to one for the firm with the highest sales in a given category. *Text Novelty* is a patent-level metric between zero and one. The text-based patent novelty measure is equal to one minus the text similarity between a given patent and its most similar predecessor within a firm (patent text similarity is computed using the same methods outlined in Section A.4.3); *Citations* is the mean number of citations received by patents in the first five years after the application; *Share Cited Others* is the share of forward citations accounted for by citations from other firms different from the patent owner; *Share Litigated* is the share of patents involved in litigation. Data on litigations come from the USPTO Patent Litigation Dataset. Due to truncation concerns, we provide statistics for the patents filed in 2005. Controls include the total number of firms in a category and fixed effects at the firm level.

greater benefits and/or lower costs of patenting increase the incentives of larger firms to file patents without product commercialization—a pattern we refer to as *strategic patenting*.

In practice, strategic patenting can take multiple forms that have been discussed in the literature. These include "sleeping patents"—ideas that are patented but not commercialized or licensed;[36] "patent thickets"—dense clusters of overlapping patents; "patent evergreening"—the practice of filing patents on incremental or secondary features to extend exclusivity beyond the original 20-year term;[37] and "defensive patents"—portfolios accumulated primarily as insurance against litigation, particularly relevant for large firms that are more attractive targets for lawsuits.

are unlikely to be the main drivers of our estimated patterns. Nonetheless, if some cost advantages evolve over time and are not fully captured by the fixed effects, this could contribute to the estimated patents-to-products association.

[36] Driscoll's, which controls a third of the U.S. berry market, invests heavily in patented berry varieties it often does not commercialize. It has one of the highest patent-to-product ratios in our data and has recently pursued lawsuits to defend its portfolio (Rosenbaum, 2017).

[37] An example is the P&G patent for Swiffer Wet Jet mops. Rather than patenting the mop's features, P&G patented the specific function of the disposable cloths. The original and over 80 follow-up patents have made market entry difficult for competitors. During our sample period, generic alternatives were virtually absent, and P&G held about 95% of the sweeper mop market, well above the 40-50% average share of leaders in other categories. For recent work on the topic, see Torrisi, Gambardella, Giuri, Harhoff, Hoisl and Mariani (2016); Shapiro (2000); Hall, Graevenitz and Helmers (2021); Righi and Simcoe (2020); Mezzanotti and Simcoe (2025).

4 Patents, Innovation, and Firm Growth

In Schumpeterian growth models, firms can expand through two channels. First, they can innovate by introducing new products to the market. Second, they can protect their existing positions by reducing competitive pressure, for example, through strategic patenting. These two strategies differ significantly in their social value: while innovation contributes directly to aggregate productivity growth, the second strategy often results in reallocation without a comparable productivity gain. With our data, we can shed light on the relative importance of these strategies for firm growth. We observe firm revenues, actual product introductions, and patent filings, which may either reflect genuine innovation or serve primarily as a defensive mechanism to safeguard market position.

We estimate the relationship between patent filings and firm revenue growth, both unconditionally and conditional on product innovation. In doing so, we also contribute to the interpretation of private patent value estimates derived from stock market data, as in Hall, Thoma and Torrisi (2007) and Kogan et al. (2017). These existing estimates conflate two components of patent value: the productive component—where patents signal true technological advances that translate into new products and revenue gains—and the strategic component —where patents serve primarily to deter competitors and facilitate reallocation without innovation. By relating patent filings to revenue growth, with and without product introductions, we disentangle the productive and strategic roles of patents in firm growth.

To estimate the revenue premium associated with firm patent portfolio and product introduction, we estimate the following baseline relationship:

$$\log \text{Sales}_{ijt} = \psi \log \text{Cum Patents}_{ijt-1} + \rho \log N_{ijt} + \log \text{Sales}_{ijt-1} + \alpha_{ij} + \gamma_{jt} + \varepsilon_{ijt}, \qquad (3)$$

where the dependent variable is the logarithm of firm i sales in product category j at time t, $\log \text{Cum Patents}_{ijt-1}$ is the cumulative number of firm's patent applications in product category j in year $t-1$ net of depreciation, and N_{ijt} is the log number of product introduction using, as before, both product counts and quality-adjusted product counts. All specifications control for lagged firm sales, category-year, and firm-category fixed effects, and a hyperbolic sine transformation is used for logs to account for zeros.

Table 7 presents results from alternative specifications around the baseline model. Column (1) excludes controls for product introductions and shows that an increase in patent application stock is associated with higher revenue in the following year. Specifically, a 10% increase in the stock of patents raises sales by 1.5%. In contrast, the elasticity of sales with respect to new product introductions is much larger, estimated at 0.517, which aligns with prior research highlighting that sustained product launches are a key driver of firm growth

Table 7: Sales on Patents, and Product Introduction

	(1)	(2)	(3)	(4)	(5)	(6)
		Log Sales (t)			Log Sales (t)	
Log Cum Patents(t-1)	0.151***		0.118***	0.139***	-0.101*	-0.028
	(0.031)		(0.029)	(0.030)	(0.058)	(0.062)
Log New Products		0.517***	0.517***		2.026***	
		(0.006)	(0.006)		(0.023)	
Log Q-new Products				0.760***		3.453***
				(0.013)		(0.054)
Log Cum Patents(t-1) × Size(t-1)					0.018***	0.014***
					(0.004)	(0.004)
Log New Products × Size(t-1)					-0.130***	
					(0.002)	
Log Q-new Products × Size(t-1)						-0.219***
						(0.004)
Observations	408,161	408,161	408,161	408,161	408,161	408,161
R-squared	0.900	0.905	0.905	0.902	0.910	0.905
Time-Category	Y	Y	Y	Y	Y	Y
Firm-Category	Y	Y	Y	Y	Y	Y
Log Sales(t-1)	Y	Y	Y	Y	Y	Y

Notes: The table shows regressions of log sales on patent stock in the previous period and the product introduction, conditional on firm sales in the previous period, using firm × category × year data. *Log Cum Patents(t-1)* is the log number of patent applications by time t-1–net of depreciation, and the *Log New Products (Q-new Products)* is the number of new products (quality-adjusted new products) introduced at time t. To account for zeros, logs are calculated using the hyperbolic sine transformation. Baseline regression specifications in (3) and (4) are based on equation (3). The last two columns interact patent stock and product introduction variables with firm log size (sales) in the previous period. All regressions additionally control for log sales (t-1), time-category and firm-category fixed effects. Standard errors are clustered at the firm × category level.

in this sector (Argente, Lee and Moreira, 2024).

Columns (3) and (4) present our baseline specification, including both patent filings and product introductions, with and without quality adjustments. As expected, the coefficient on patents declines relative to column (1), reflecting that realized product introductions account for part of the revenue premium associated with patenting. However, the remaining premium remains sizable. The residual revenue premium not explained by market innovations is consistent with the view that patent portfolios serve a protective role, limiting competition and reallocating market share to the patenting firm. It may also reflect consumers' perceptions of patented products as being of higher quality.[38]

The last two columns of Table 7 examine firm heterogeneity in the relationship between sales, product introduction, and patenting. Columns (5) and (6) extend the baseline specification by interacting patents and product introductions with lagged firm revenue (in logs). The results reveal striking size-dependent patterns in the sources of growth. Among smaller

[38]Table B5 in the Appendix replicates these specifications for prices and quantities, showing that the incremental revenue from patenting arises from both higher quantities sold and higher prices.

firms, product introductions drive revenue growth, while for larger firms, patents yield a significant revenue premium even after accounting for new products. For example, the elasticity of sales to patents (product introductions) is 0.03 (1.04) for firms in the bottom quintile of the size distribution, compared to 0.12 (0.43) for firms in the top quintile. These results suggest that patents contribute to revenue growth through channels beyond product innovation, particularly for larger firms. This is consistent with the view that patents help larger firms deter competition and reduce creative destruction, enabling them to expand market share without continual product innovation.[39]

Lastly, our baseline specification uses the cumulative count of patents to reflect the strength of a firm's patent portfolio, recognizing that multiple interrelated patents may jointly contribute to defending the firm's market position. Appendix Table B6 replicates the analysis using patent flows—i.e., annual applications—instead of stocks. While the results remain qualitatively similar, the estimated elasticity of revenue with respect to patent flows is smaller.[40] This suggests that it is the accumulation of patent stocks, particularly for larger firms, that plays a key role in sustaining revenue growth.

4.1 Patents and Creative Destruction

The previous results showed that firm patenting is associated with revenue growth beyond its effect through product introduction—especially for larger firms within product categories. This aligns with longstanding views that patents help firms deter competitors and limit creative destruction (e.g., Jaffe and Lerner, 2004; Cockburn and J. MacGarvie, 2011; Williams, 2013; Lampe and Moser, 2015; Galasso and Schankerman, 2015). Below, we provide additional evidence supporting this mechanism and explore why the patent-related revenue premium is larger for bigger firms: by reducing the threat of creative destruction, patents may help incumbents sustain market share without continual innovation.

We examine whether firm patenting is associated with reduced product entry by rivals. Using our data, we test whether patents filed by market leaders are followed by declines in product introductions by their competitors—whom we refer to, for simplicity, as market followers. We start by identifying the market leader in each category as the firm with the highest sales in that category and the followers as the remaining firms operating in that market.[41] Then, for each year t and product category j, we compute the number of new

[39]Table B7 shows that the patent revenue premium for larger firms increases after 2010, supporting the view that strategic patent use has been growing over time (Akcigit and Ates, 2023).

[40]The estimated patent revenue premium in these specifications is in the ballpark of the private patent value estimates reported in Kogan, Papanikolaou, Seru and Stoffman (2017).

[41]To have a static firm-level measure, we define leaders as of 2006, which is the first year of our data. However, the results are not sensitive to a different choice, like using average sales over all years. Moreover,

products introduced by the leader N^{L}_{jt} and the average product introduction by its followers N^{F}_{jt} in t, and we compute the number of patent applications introduced by the leader P^{L}_{jt} and by its followers P^{F}_{jt} until t. We evaluate how product introduction by followers responds to patenting (and product introduction) of the leaders using the following specification:

$$\log N^{\mathrm{F}}_{jt} = \eta^{\mathrm{F}} \log P^{\mathrm{L}}_{jt-1} + \alpha^{\mathrm{F}} \log N^{\mathrm{L}}_{jt-1} + \theta^{\mathrm{F}}_{j} + \gamma^{\mathrm{F}}_{t} + \varepsilon^{\mathrm{F}}_{jt}, \tag{4}$$

where η^{F} is our coefficient of interest, measuring the association of patents of leaders with the product introduction by followers. We control for $\ln N^{\mathrm{L}}_{jt-1}$ to ensure that the relationship between leaders' patents and followers' product introduction is not driven by possible direct interactions between the leader's and followers' product offerings (such as learning from new products on the market).[42] We also include time- and category-fixed effects to control for time trends and differences in the intensities of patenting and product innovation across product categories. Likewise, we estimate a symmetric regression that estimates the relationship between leaders' innovation and the followers' patenting:

$$\log N^{\mathrm{L}}_{jt} = \eta^{\mathrm{L}} \log P^{\mathrm{F}}_{jt-1} + \alpha^{\mathrm{L}} \log N^{\mathrm{F}}_{jt-1} + \theta^{\mathrm{L}}_{j} + \gamma^{\mathrm{L}}_{t} + \varepsilon^{\mathrm{L}}_{jt} \tag{5}$$

These regressions help us test whether the relation between competitors' patents and product introduction is affected by whether we focus on leaders or followers.

Table 8 presents the estimated coefficients. Column (1) shows that product introduction by followers is negatively correlated with the size of the leader's patent portfolio. This result suggests that followers reduce the introduction of new products in categories where the leader intensifies its patenting efforts. By contrast, we do not find evidence that leaders' product introductions crowd out those of followers, conditional on their patenting activity. In column (2), we control for total market sales to account for potential shifts over time in the importance of different product types. Columns (3) and (4) show that product innovation by leaders is not significantly related to the followers' patenting activity. Hence, while patents can be thought of as a protective tool used to hinder product-market competition, our results indicate that this mechanism is most relevant for large market leaders. These findings suggest that leaders are more likely to accumulate patents in ways that limit competition, helping to explain the sizable revenue premium from patents among larger firms.

we consider alternative definitions of market leaders (e.g., top quintile), and the results are robust.

[42]We also use quality-adjusted new products in all of these regressions, and the results are similar.

Table 8: **Patenting of Market Leaders and Followers**

	Followers Log N^{F} (1)	(2)		Leaders Log N^{L} (3)	(4)
Leaders			**Followers**		
Log P^{L} (t-1)	-0.071***	-0.059***	Log P^{F} (t-1)	-0.015	-0.012
	(0.018)	(0.018)		(0.061)	(0.061)
Log N^{L} (t-1)	0.010*	0.005	Log N^{F} (t-1)	0.215***	0.186**
	(0.006)	(0.006)		(0.079)	(0.079)
Observations	3,192	3,192	Observations	3,188	3,188
Category	Y	Y	Category	Y	Y
Time	Y	Y	Time	Y	Y
Controls	N	Y	Controls	N	Y

Notes: The table shows the relationship between the patents of leaders (followers) and the product introduction of followers (leaders). The leader is defined as the firm with the highest sales in a given category in 2006; the followers are defined as the rest of the firms in the category. In columns (1) and (2), the dependent variable is the log average number of products introduced by followers at time t, and the independent variables are the log number of patent applications by leaders until time $t-1$ and the log number of new products introduced by the leader at time $t-1$. In columns (3) and (4), the dependent variable is the log number of products introduced by leaders at time t, and the independent variables are the log average number of patent applications filed by followers until time $t-1$ and the log average number of new products introduced by the followers at time $t-1$. Columns (2) and (4) also control for total sales in the category-time. The inverse hyperbolic sine transformation is used for logarithms.

5 Conceptual Framework

We develop an illustrative Schumpeterian model that underscores the importance of measuring both product innovation and patenting activity. The model rationalizes our findings that patents held by larger firms are less likely to result in product innovations and that the revenue premium from patents used to deter creative destruction is greater for larger firms. This framework allows us to assess the implications of patents filed for protective purposes without corresponding market innovations, which, again, we refer to as *strategic patents*.

The model distinguishes between a firm's decision to innovate and its decision to patent, thereby highlighting how innovation and protection incentives diverge with firm size. As firms expand, the attractiveness of introducing new products declines due to rent cannibalization, whereas the incentive to patent increases as firms seek to defend their larger market positions. Consequently, larger firms are more likely to engage in strategic patenting—filing patents without subsequent product commercialization. Although such behavior may help incumbents preserve market dominance, it ultimately weakens creative destruction and slows the pace of innovation.

Product Introduction and Patenting – We consider a partial equilibrium framework of innovation in a single product line. The product line is held by an incumbent who produces a product of quality q and earns profit $\Pi = \pi q^{\gamma}$ with $0 < \gamma < 1$. Hence, an incumbent with a higher-quality product is larger and earns higher profits, but faces diminishing returns to quality.[43] Incumbents can improve their products and file patents. We model a one-time decision of product introduction and patenting for an incumbent with quality q who exogenously obtains an idea of size λ.[44]

Product introduction is uncertain. The firm chooses the probability of product introduction z_m by incurring a commercialization cost $\frac{c_m z_m^2}{2}$. If successful, it brings a higher-quality product $q + \lambda$ to market and earns higher profits. At the same time, the firm chooses the probability of patenting z_p, incurring cost $\frac{c_p z_p^2}{2}$, which covers, for example, research, filing, legal, and enforcement expenses. A patent provides protection against replacement by competitors. Even without commercialization, a patent filing makes the idea public, so the highest available quality in the economy becomes $q + \lambda$ (Hegde, Herkenhoff and Zhu, 2022). Thus, the highest quality available in the economy may differ from what is commercialized to consumers. In this sense, firms' product and patent activities are distinct: product introduction need not imply patenting, and patenting need not imply product introduction, an important departure from standard models of innovation and growth.

Creative Destruction – Incumbents can be displaced by entrants through creative destruction. For that, entrants must introduce a higher-quality product and overcome any patent barriers. Entrants arrive at an exogenous rate p and draw an innovation step λ^e from a uniform distribution on $(0, 1)$ relative to the highest available quality. Building on "the shoulders of giants," entrants learn both from existing products and patents. Thus, the quality frontier is $q + \lambda$ if the incumbent introduces and/or patents, and q otherwise.

Normally, an entrant wins the market by offering higher quality than the incumbent. This may not hold if the incumbent holds patent protection. Patents protect the incumbent's quality by creating a "wall" of height ε $(0 < \varepsilon < 1)$ that entrants must clear to enter, similar to Abrams, Akcigit and Grennan (2013). The parameter ε reflects how different an entrant's innovation must be, which can depend on IP strength and patent scope. Thus, the probability of creative destruction is p without patents and $p\,(1 - \varepsilon)$ with patents (see Appendix C.2). Unlike standard models, not all quality improvements reach the market: entrants may be blocked by incumbents' patents, highlighting the separation between product and patent spaces.

[43] See Appendix C.1 for microfoundations and discussion of decreasing returns to quality. An assumption of $\gamma < 1$ is also supported by our calibration.

[44] For simplicity, the idea is either used or disappears. A dynamic version would yield similar qualitative predictions but require tracking the firm's evolution in product and patent spaces.

Value Function – Consider an incumbent with initial product quality q.[45] Let V^{11} denote its (gross) value when it both introduces a new product and patents, V^{10} when it only introduces, V^{01} when it only patents, and V^{00} when it does neither. Then we obtain

$$\begin{aligned} V^{11}(q) &= \frac{\pi(q+\lambda)^\gamma}{r+p(1-\varepsilon)}, \qquad V^{10}(q) = \frac{\pi(q+\lambda)^\gamma}{r+p}, \\ V^{01}(q) &= \frac{\pi q^\gamma}{r+p(1-\varepsilon)}, \qquad V^{00}(q) = \frac{\pi q^\gamma}{r+p}, \end{aligned} \tag{6}$$

where r is the interest rate. The value of the incumbent firm with existing quality q is then an expectation over these values, net of product introduction and patenting costs:

$$\begin{aligned} \mathbb{V}(q) = \max_{z_m, z_p} \Big\{ & z_m z_p V^{11}(q) + z_m(1-z_p)V^{10}(q) + (1-z_m)z_p V^{01}(q) \\ & + (1-z_m)(1-z_p)V^{00}(q) - \frac{c_m z_m^2}{2} - \frac{c_p z_p^2}{2} \Big\}. \end{aligned} \tag{7}$$

Strategic Patenting – A central feature of this economy is that incumbents can pursue *strategic patenting*—patenting without product introduction—represented by option $V^{01}(q)$. We show that this option is increasingly valuable for larger firms. To evaluate how the incentives for patenting and product introduction vary with firm size, consider the revenue premium from strategic patenting: $(V^{01} - V^{00})$:

$$V^{01} - V^{00} = \frac{\pi \varepsilon p q^\gamma}{(r+p(1-\varepsilon))(r+p)}.$$

The revenue premium rises with firm size q: larger incumbents have more value to protect and thus reap greater returns. By contrast, incentives for product introduction fall with size ($\frac{\partial(V^{10}-V^{00})}{\partial q} < 0$), reflecting the *Arrow-replacement effect*: incremental returns from new products diminish with firm size. These opposing forces explain why larger firms gain more from strategic patenting. Incentives for strategic patenting also increase with stronger patent protection (ε) and higher profit flows (π).

Although all patents slow creative destruction and prolong incumbents' market presence, only non-strategic ones foster aggregate innovation. Strategic patents instead block reallocation without delivering offsetting gains, potentially hindering growth. These dynamics underscore the need to assess their broader impact in richer models. The next section provides a back-of-the-envelope calculation to illustrate these trade-offs within our framework.

[45] We analyze a single incumbent in a product category facing potential entry. While the model can be generalized to multiple firms with different market shares, we focus on a single incumbent and study comparative statics of product introduction and patenting with respect to firm size (sales), captured in the model by q.

Table 9: **Implications of Strategic Patenting. Model Counterfactuals**

	Benchmark	No Strategic & benchm. innov.	No Strategic
Innovation (z_m)			
Median incumbent	0.2102	0.2102	0.2103
Large incumbent	0.1550	0.1550	0.2794
Creative destruction (τ)			
Median incumbent	0.0948	0.0949	0.0949
Large incumbent	0.0915	0.0944	0.0940
Patenting (z_p)			
Median incumbent	0.0092	0.0024	0.0024
Large incumbent	0.1614	0.0252	0.0477

Notes: The first column shows optimal z_m, z_p, and τ from the model. The second column considers the economy with z_m as in the benchmark, but where no strategic patents are allowed (no V^{01} option). The third economy again shuts down the strategic patents option, but calculates optimal innovation and patent choices.

5.1 Implications of Strategic Patenting: Quantitative Illustration

We solve the model and present counterfactuals to illustrate how the option to patent without product introduction—strategic patenting—can dampen creative destruction and innovation. Given the stylized nature of the framework, we interpret the results as highlighting broad qualitative mechanisms rather than providing precise quantitative predictions. The numerical values should be viewed as rough, back-of-the-envelope calculations for the consumer goods sector we study.

We calibrate the model to match firm growth by patenting status, innovation rates, patents per product, and sales growth across firm-size percentiles. Appendix C.3 details the procedure and shows that, despite its parsimony, the model reproduces key patterns in the data quite well. We then conduct two counterfactuals. First, we provide a back-of-the-envelope estimate of strategic patents; second, we analyze a policy regime without strategic patents, where protection is granted only if a patent is tied to a market innovation.

The first counterfactual isolates the "excess" patents filed by incumbents for strategic rather than innovative purposes. For that, we compare the benchmark economy to a counterfactual where product introduction rates are *held at benchmark values*, but the strategic patenting channel is shut down: firms can patent only when introducing a new product improvement (shutting down V^{01} option). Column (1) of Table 9 reports innovation, patenting, and implied creative destruction in the benchmark economy for two product lines: one held by a large incumbent (p90-p95 size percentile in the data) and another by a typical incumbent (median size). Column (2) shows the counterfactual. For the same innovation rate, large incumbents patent 84% less when strategic patents are disallowed–

highlighting prevalence of strategic patents with no product innovation among large firms. This excessive patenting reduces creative destruction in the benchmark by 3%. These effects are negligible for markets with smaller incumbents, suggesting that large firms rely most heavily on strategic patenting.

Second, we analyze a policy regime without strategic patents where, unlike the previous exercise, firms optimally choose both patenting and innovation rates. Comparing Column (3) for this counterfactual with the benchmark shows that, again, incumbents patent less, which raises equilibrium creative destruction in markets with large incumbents. A key new margin is that large incumbents innovate substantially more. In the benchmark economy, incumbents' product introduction is lower because they rely on strategic patenting to protect market shares, reducing incentives to innovate. In the no-strategic patents economy, maintaining market position requires genuine innovation rather than defensive patenting.[46]

Extension: Heterogeneity in Costs – The patenting cost parameter c_p is assumed to be independent of firm size. One might argue that larger firms face lower effective costs due to their "deeper pockets," legal teams, and resources. However, this would only strengthen their incentives for strategic patenting: lower costs would lead to more patents without product introduction, further reducing creative destruction. Counterfactual experiments with size-dependent costs confirm that large firms' share of strategic patents remains similar to the baseline (see Appendix C.4). While technological differences make it easier for large firms to patent without innovating, their fundamental strategic motive, using patents to defend their large market shares rather than to introduce new products, remains unchanged.

Overall, our exercises highlight that large incumbents have high incentives (benefits and/or costs of patents) to file strategic patents that reduce creative destruction and do not meaningfully contribute to the advancement of market innovations, and that a patent system in which protection can be granted without firms commercializing their inventions stifles not only competitors innovation but also dampens incumbents incentives to innovate.

6 Conclusion

This paper develops the first large-scale patent-to-product match, linking patents to products through textual analysis of USPTO documents and NielsenIQ- and Wikipedia-based product descriptions. This new dataset allows us to move beyond traditional statistics

[46] So, patent protection incentivizes innovation, consistent with the innovation-enhancing role of patents Bryan and Williams (2021)

and uncover several insights. We show that most new products are introduced by firms that never patent. Among patenting firms, filings are positively associated with subsequent product introductions, but the link between patents and innovation weakens with firm size. Larger firms file more patents relative to the products they introduce, resulting in a lower conversion rate of patents into market innovations. At the same time, patenting yields sizable revenue premiums not explained by product innovation, consistent with the view that patent portfolios serve a protective role by limiting competition and reallocating market share. In line with this, we find that followers scale back product introductions in categories where leaders intensify their patenting.

We rationalize these findings with a simple innovation model showing how larger firms have stronger incentives to engage in strategic patenting—filing for protection rather than market innovation. Counterfactual experiments confirm that a large amount of patents by large firms are strategic, and that their prevalence dampens innovation by leaders and slows creative destruction in the consumer goods sector.

Our results suggest that reforms tying patent protection more closely to genuine market innovation could substantially enhance the overall innovation process. While distinguishing strategic from *bona fide* patents is inherently difficult, our findings provide guidance. One option is a working-prototype requirement, forcing applicants—especially large firms—to undertake part of the development process before securing patent protection—a principle with historical precedent in the 19th-century USPTO model requirement (Khan, 2005). A complementary approach is to enhance examination-stage scrutiny—examining text-based novelty with particular attention to applicants' prior portfolios and overarching claim breadth—and to direct these efforts toward large incumbents, where, as we show, the connection between patents and products is weakest and strategic motives most pronounced.[47] We believe that extending our framework into richer general equilibrium quantitative models of strategic patenting—capable of evaluating such policies and assessing their aggregate implications—is an important avenue for future research.

[47] An analogy is trademark law: under the Lanham Act (15 U.S.C. 1051 et seq.), applicants must demonstrate a *bona fide* intent to use the mark in commerce.

References

Abrams, David S, Ufuk Akcigit, and Jillian Grennan, "Patent Value and Citations: Creative Destruction or Strategic Disruption?," Working Paper 19647, National Bureau of Economic Research November 2013.

Aghion, Philippe and Peter Howitt, "A Model of Growth through Creative Destruction," *Econometrica*, 1992, *60* (2), 323–51.

__ , Antonin Bergeaud, Timo Boppart, Peter J. Klenow, and Huiyu Li, "A Theory of Falling Growth and Rising Rents," *The Review of Economic Studies*, 2023, *90*, 2675–2702.

Aizawa, Akiko, "An information theoretic perspective of TF-IDF measures," *Information Processing and Management*, 01 2003, *39*, 45–65.

Akcigit, Ufuk and Nathan Goldschlag, "Where Have All the Creative Talents Gone? Employment Dynamics of US Inventors," Working Paper 31085, National Bureau of Economic Research March 2023.

__ and Sina T Ates, "Ten facts on declining business dynamism and lessons from endogenous growth theory," *American Economic Journal: Macroeconomics*, 2021, *13* (1), 257–98.

__ and __ , "What Happened to U.S. Business Dynamism?," *Journal of Political Economy*, 2023.

__ and William R. Kerr, "Growth through Heterogeneous Innovations," *Journal of Political Economy*, 2018, *126* (4), 1374 – 1443.

__ , Murat Celik, and Jeremy Greenwood, "Buy, Keep, or Sell: Economic Growth and the Market for Ideas," *Econometrica*, 2016, *84*, 943–984.

__ , Salome Baslandze, and Francesca Lotti, "Connecting to Power: Political Connections, Innovation, and Firm Dynamics," *Econometrica*, 2023.

Alexopoulos, Michelle, "Read All about It!! What Happens Following a Technology Shock?," *American Economic Review*, June 2011, *101* (4), 1144–79.

Argente, David and Chen Yeh, "Product Life Cycle, Learning, and Nominal Shocks," *The Review of Economic Studies*, 2022, *89* (6), 2992–3054.

__ , Ezra Oberfield, and Diana Van Patten, "On the Dynamics of International Product Innovation," 2025.

—, Munseob Lee, and Sara Moreira, "Innovation and Product Reallocation in the Great Recession," *Journal of Monetary Economics*, 2018, *93*, 1–20.

—, —, and —, "The life cycle of products: Evidence and implications," *Journal of Political Economy*, 2024, *132* (2), 337–390.

Autor, David, David Dorn, Lawrence F Katz, Christina Patterson, and John Van Reenen, "The Fall of the Labor Share and the Rise of Superstar Firms," *The Quarterly Journal of Economics*, 2020, *135* (2), 645–709.

Balasubramanian, Natarajan and Jagadeesh Sivadasan, "What Happens When Firms Patent? New Evidence from U.S. Economic Census Data," *The Review of Economics and Statistics*, 2011, *93* (1), 126–146.

Baslandze, Salome, "Barriers to Creative Destruction: Large Firms and Nonproductive Strategies," FRB Atlanta Working Paper 2021-23, Federal Reserve Bank of Atlanta September 2021.

Bena, Jan and Elena Simintzi, "Globalization of work and innovation: Evidence from doing business in china," Technical Report, Discussion paper 2017.

Bessen, James and Michael J. Meurer, *Patent Failure: How Judges, Bureaucrats, and Lawyers Put Innovators at Risk*, Princeton, NJ: Princeton University Press, 2008.

Bloom, Nicholas, Charles I Jones, John Van Reenen, and Michael Webb, "Are Ideas Getting Harder to Find?," *American Economic Review*, 2020, *110* (4), 1104–44.

Boldrin, Michele and David K. Levine, "The Case Against Patents," *Journal of Economic Perspectives*, February 2013, *27* (1), 3–22.

Bresnahan, Timothy F and Robert J Gordon, "Introduction to" The Economics of New Goods"," in "The Economics of New Goods," University of Chicago Press, 1996, pp. 1–26.

Broda, Christian and David E Weinstein, "Product Creation and Destruction: Evidence and Price Implications," *American Economic Review*, 2010, *100* (3), 691–723.

Bryan, Kevin A and Heidi L Williams, "Innovation: Market Failures and Public Policies," in "Handbook of Industrial Organization," Vol. 5, Elsevier, 2021, pp. 281–388.

Cavenaile, Laurent and Pau Roldan-Blanco, "Advertising, Innovation, and Economic Growth," *American Economic Journal: Macroeconomics*, July 2021, *13* (3), 251–303.

Cockburn, Iain and Megan J. MacGarvie, "Entry and Patenting in the Software Industry," *Management Science*, 05 2011, *57*, 915–933.

Cohen, Wesley M. and Steven Klepper, "Firm Size and the Nature of Innovation Within Industries: The Case of Process and Product R&D," *The Review of Economics and Statistics*, 1996, *78* (2), 232–243.

Cohen, Wesley M, Richard R Nelson, and John P Walsh, "Protecting their Intellectual Assets: Appropriability Conditions and Why US Manufacturing Firms Patent (or not)," NBER Working Paper 2000.

Consumer Brands Association, "CPG Industry Drives U.S. Economy with $2.5 Trillion Contribution and 22.3 Million Jobs," https://consumerbrandsassociation.org/press-releases/cpg-industry-drives-u-s-economy-with-2-5-trillion-contribution-and-22-3-million-jobs/ October 2024. Accessed August 2025.

Crouzet, Nicolas and Janice C Eberly, "Understanding Weak Capital Investment: the Role of Market Concentration and Intangibles," NBER Working Paper 25869 May 2019.

Cunningham, Colleen, Florian Ederer, and Song Ma, "Killer Acquisitions," *Journal of Political Economy*, 2021, *129* (3), 000–000.

de Rassenfosse, Gaétan, "Notice Failure Revisited: Evidence on the Use of Virtual Patent Marking," NBER Working Paper 2018.

Decker, Ryan A., John Haltiwanger, Ron S. Jarmin, and Javier Miranda, "Declining Business Dynamism: What We Know and the Way Forward," *American Economic Review*, May 2016, *106* (5), 203–07.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2018.

Fernández-Villaverde, Jess, Yang Yu, and Francesco Zanetti, "Defensive Hiring and Creative Destruction," Working Paper 33588, National Bureau of Economic Research March 2025.

Fosfuri, Andrea, "The Licensing Dilemma: Understanding the Determinants of the Rate of Licensing," *Strategic Management Journal*, 03 2004, *27*.

Galasso, Alberto and Mark Schankerman, " Patents and Cumulative Innovation: Causal Evidence from the Courts," *The Quarterly Journal of Economics*, 2 2015, *130* (1), 317–369.

Gambardella, Alfonso, Paola Giuri, and Alessandra Luzzi, "The Market for Patents in Europe," LEM Papers Series 2006/04, Laboratory of Economics and Management (LEM), Sant'Anna School of Advanced Studies, Pisa, Italy February 2006.

Garcia-Macia, Daniel, Chang-Tai Hsieh, and Peter J. Klenow, "How Destructive is Innovation?," *Econometrica*, 2019, *87* (5), 1507–1541.

Gilbert, Richard J. and David M. G. Newbery, "Preemptive Patenting and the Persistence of Monopoly," *The American Economic Review*, 1982, *72* (3), 514–526.

Goolsbee, Austan D. and Peter J. Klenow, "Internet Rising, Prices Falling: Measuring Inflation in a World of E-Commerce," *AEA Papers and Proceedings*, May 2018, *108*, 488–92.

Gordon, Robert J., "The Rise and Fall of American Growth: The U.S. Standard of Living since the Civil War," *Journal of Regional Science*, 2016, *56* (5), 921–924.

Graham, Stuart JH, Cheryl Grim, Tariqul Islam, Alan C Marco, and Javier Miranda, "Business dynamics of innovating firms: Linking US patents with administrative data on workers and firms," *Journal of Economics & Management Strategy*, 2018, *27* (3), 372–402.

_ , Robert P Merges, Pam Samuelson, and Ted Sichelman, "High Technology Entrepreneurs and the Patent System: Results of the 2008 Berkeley Patent Survey," *Berkeley Technology Law Journal*, 2009, *24* (4), 1255.

Grossman, Gene M. and Elhanan Helpman, "Quality Ladders in the Theory of Growth," *The Review of Economic Studies*, 1991, *58* (1), 43–61.

Gutiérrez, Germán and Thomas Philippon, "Declining Competition and Investment in the U.S.," NBER Working Paper 23583 July 2017.

_ and _ , "The Failure of Free Entry," NBER Working Paper 26001 2019.

Hall, Bronwyn, Adam Jaffe, and Manuel Trajtenberg, "The NBER Patent Citations Data File: Lessons, Insights and Methodological Tools," NBER Working Paper 8498 2001.

_ , _ , and _ , "Market Value and Patent Citations," *RAND Journal of Economics*, 2005, *36* (1), 16–38.

_ , Grid Thoma, and Salvatore Torrisi, "The Market Value of Patents and R&D: Evidence from European Firms," *Academy of Management Proceedings 2007*, 10 2007, *2007*.

Hall, Bronwyn H, Georg von Graevenitz, and Christian Helmers, "Technology Entry in the Presence of Patent Thickets," *Oxford Economic Papers*, 2021, *73* (2), 903–926.

Hegde, Deepak, Kyle Herkenhoff, and Chenqi Zhu, "Patent Publication and Innovation," NBER Working Paper 29770 2022.

Hoberg, Gerard and Gordon Phillips, "Text-based Network Industries and Endogenous Product Differentiation," *Journal of Political Economy*, 2016, *124* (5), 1423–1465.

Hottman, Colin J, Stephen J Redding, and David E Weinstein, "Quantifying the Sources of Firm Heterogeneity," *The Quarterly Journal of Economics*, 2016, *131* (3), 1291–1364.

Jaffe, Adam B. and Josh Lerner, "Innovation and Its Discontents: How Our Broken Patent System is Endangering Innovation and Progress, and What to Do About It.," *Princeton University Press, Princeton, NJ*, 2004.

Joulin, Armand, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov, "Bag of Tricks for Efficient Text Classification," in "Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Vol 2" April 2017, pp. 427–431.

Kalyani, Aakash, Nicholas Bloom, Marcela Carvalho, Tarek Hassan, Josh Lerner, and Ahmed Tahoun, "The Diffusion of New Technologies*," *The Quarterly Journal of Economics*, 01 2025, p. qjaf002.

Kelly, Bryan, Dimitris Papanikolaou, Amit Seru, and Matt Taddy, "Measuring Technological Innovation over the Long Run," *American Economic Review: Insights*, September 2021, *3* (3), 303–20.

Khan, B. Zorina, *The Democratization of Invention: Patents and Copyrights in American Economic Development, 1790–1920*, Cambridge, UK: Cambridge University Press, 2005.

Kogan, Leonid, Dimitris Papanikolaou, Amit Seru, and Noah Stoffman, "Technological Innovation, Resource Allocation, and Growth*," *The Quarterly Journal of Economics*, 03 2017, *132* (2), 665–712.

Lampe, Ryan and Petra Moser, "Patent Pools, Competition, and Innovation–Evidence from 20 US Industries under the New Deal," *The Journal of Law, Economics, and Organization*, 08 2015, *32* (1), 1–36.

Lerner, Josh, "Patenting in the Shadow of Competitors," *The Journal of Law Economics*, 1995, *38* (2), 463–495.

Lloyd, Stuart, “Least Squares Quantization in PCM,” *IEEE transactions on information theory*, 1982, *28* (2), 129–137.

Loecker, Jan De, Jan Eeckhout, and Gabriel Unger, “The Rise of Market Power and the Macroeconomic Implications,” *The Quarterly Journal of Economics*, 2020, *135* (2), 561–644.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze, *Introduction to Information Retrieval*, New York, NY, USA: Cambridge University Press, 2008.

Mezzanotti, Filippo and Timothy Simcoe, “Innovation, Patenting and Appropriability: Survey Evidence from a Nationally Representative Sample of US Firms,” 2025.

Moser, Petra, “Innovation without Patents: Evidence from World’s Fairs,” *Journal of Law and Economics*, 2012, *55* (1), 43 – 74.

Pakes, Ariel, “Patents as Options: Some Estimates of the Value of Holding European Patent Stocks,” *Econometrica*, 1986, *54* (4), 755–784.

Pearce, Jeremy and Liangjie Wu, “Brand Reallocation and Market Concentration,” Staff Report 1116, Federal Reserve Bank of New York August 2024.

PwC, “2018 Patent Litigation Study,” https://www.pwc.com/us/en/services/forensics/library/2018-patent-litigation-study.html May 2018. Accessed August 2025.

Righi, Cesare and Timothy Simcoe, “Patenting Inventions or Inventing Patents? Strategic Use of Continuations at the USPTO,” NBER Working Paper 2020.

Romer, Paul M, “Endogenous Technological Change,” *Journal of Political Economy*, October 1990, *98* (5), 71–102.

Rosenbaum, Matthew, “How Driscoll’s Reinvented the Strawberry,” *The New Yorker*, August 2017.

Schankerman, Mark and Ariel Pakes, “Estimates of the Value of Patent Rights in European Countries During the Post-1950 Period,” *The Economic Journal*, 1986, *96* (384), 1052–1076.

Shapiro, Carl, “Navigating the Patent Thicket: Cross Licenses, Patent Pools, and Standard Setting,” *Innovation Policy and the Economy*, 2000, *1*, 119–150.

Tham, W, S Baslandze, E Sojli, and L Liu, "Foundational Processes and Growth," Discussion Paper 19858, CEPR Press, Paris & London 2025.

Torrisi, Salvatore, Alfonso Gambardella, Paola Giuri, Dietmar Harhoff, Karin Hoisl, and Myriam Mariani, "Used, Blocking and Sleeping patents: Empirical Evidence from a Large-scale Inventor Survey," *Research policy*, 2016, *45* (7), 1374–1385.

Trajtenberg, Manuel, "A Penny for Your Quotes: Patent Citations and the Value of Innovations," *The RAND Journal of Economics*, 1990, *21* (1), 172–187.

Webb, Michael, "The Impact of Artificial Intelligence on the Labor Market," *Available at SSRN 3482150*, 2019.

Williams, Heidi L., "Intellectual Property Rights and Innovation: Evidence from the Human Genome," *Journal of Political Economy*, 2013, *121* (1), 1–27.

Younge, Kenneth A and Jeffrey M Kuhn, "Patent-to-patent Similarity: a Vector Space Model," *Available at SSRN 2709238*, 2016.

Patents to Products:
Product Innovation and Firm Dynamics

Appendix

For Online Publication

A Additional Data Information

A.1 Product Data

Coverage.— The main advantage of the RMS data set is its size and coverage. Overall, the RMS data consists of more than 100 billion unique sales observations at the week × store × UPC level. The data set comprises around 12 billion transactions per year which are worth $220 billion dollars on average. Over our sample period, 2006-2015, the total sales across all retail establishments are worth approximately $2 trillion and represent 53% of all sales in grocery stores, 55% in drug stores, 32% in mass merchandisers, 2% in convenience stores, and 1% in liquor stores. A key distinctive feature of this database is that the collection points include more than 40,000 distinct stores from around 90 retail chains, across 371 metropolitan statistical areas (MSAs) and 2,500 counties. We keep a balanced set of stores throughout the entire period under the analysis.

Because of its size, the data provides good coverage of the universe of products in the consumer goods sector. Our assessment is based on three considerations. First, comparisons with other scanner data sets reveals that NielsenIQ RMS covers more product introductions and provides more accurate information on product entry time. Argente et al. (2018) compares NielsenIQ RMS with other scanner data sets collected at the store level and shows that NielsenIQ RMS covers a much wider range of products and stores. In comparison to scanner data sets collected at the household level, NielsenIQ RMS also has a wider range of products because it reflects the universe of all transactions for the categories it covers, as opposed to the purchases made by a sample of households. For example, NielsenIQ Homescan covers less than 60% of the products the NielsenIQ RMS covers in a given year.

Second, while the data only covers sales in traditional retail channels and not e-commerce, we do not expect this to substantially affect the total level of innovations in the sector. Between 2000 and 2014, the fraction of all retail sales accounted for by e-commerce went from 0.9 to 6.4 percent, according to figures from the US Census Bureau (Hortaçsu and Syverson, 2015). Thus, during our sample period, online commerce is still a small part of retail activity and will affect innovation numbers by firms that only sell online.

Finally, the data covers sales in food and non-food categories (health and beauty aids, non-food grocery, and general merchandise). However, because the data set has higher coverage of grocery stores, food categories have relatively higher coverage than some general merchandise categories (see, for example, Jaravel (2019) for a thorough comparison of

NielsenIQ RMS and Homescan with the Consumer Expenditure Survey). We assess the impact of this differential coverage of product categories on our measures of product innovation by comparing product introduction rates in our data with those in NielsenIQ Homescan and other sources (e.g. Goolsbee and Klenow, 2018). We do not find a significant association between sales coverage and the differences in product introductions between data sets across various product categories. Nevertheless, throughout the paper we evaluate the robustness of the results when we keep only products that have high coverage.

NielsenIQ Product Classification System.— The data is organized into detailed product modules that are aggregated into product groups. The product groups are then grouped into ten major departments. These departments are: Health and Beauty Aids, General Merchandise, Dry Grocery (e.g., baby food, canned vegetables), Frozen Foods, Dairy, Deli, Packaged Meat, Fresh Produce, Non-Food Grocery, and Alcohol. For example, a 31-ounce bag of Tide Pods has UPC 037000930389, is produced by Procter & Gamble, and belongs to the product module "Detergent-Packaged" in product group "Detergent," which belongs to the "Non-Food Grocery" department. The product group "Detergent" includes several product modules, including automatic dishwasher compounds, detergents heavy duty liquid, detergents light duty, detergents packaged, dishwasher rinsing aids, and packaged soap.

Over time, NielsenIQ expanded coverage of certain product modules (for instance, some in-store food goods), but we keep a consistent set of product modules that are available throughout the period. This leaves us with 10 departments, 114 product groups and 1,070 modules.

Defining a Product.— Defining products by their UPCs has some important advantages. First, UPCs are by design unique to every product: changes in any attribute of a good (e.g. forms, sizes, package, formula) result in a new UPC.This offers a unique opportunity for economists to identify products at the finest level of disaggregation.

Second, UPCs are so widespread that our data is likely to cover all products sold in the consumer goods sector. Producers have a strong incentive to purchase UPCs for all products that have more than a trivial amount of sales because the codes are inexpensive, and they allow sellers to access stores with scanners and internet sales.

For each product in a year, we define its sales as the total sales across all stores and weeks in the year. Likewise, quantity is defined as total quantities sold across all stores and weeks in the year. Price is defined by the ratio of revenue to quantity, which is equivalent to the quantity-weighted average price.[48] To minimize concerns about potential measurement error caused by Nielsen's treatment of private-label products to protect the identity of the retailers, we exclude all private-label goods from the data.

Assigning Products to Firms.— NielsenIQ RMS data does not include information on manufacturing firms. However, products can be linked with firms using information obtained from the GS1 US Data Hub. In order to issue a UPC, firms must first obtain a

[48]We use the weight and the volume of the product to compute unit values.

GS1 company prefix. The prefix is a five- to ten-digit number that identifies firms in their products' UPCs. Argente et al. (2018) provide more details on how to use a subset of the product UPCs to link producers with products.

The GS1 data include the name and address of the firm associated with each prefix, which allows us to append a firm name and location to the UPCs included in the Nielsen-RMS data. A "firm" in the database is defined based on the entity that purchased the barcodes from GS1, which is typically the manufacturer, such as Procter & Gamble.

Constructing a sample of CPG-only firms.— Any firm that produces at least one product in the NielsenIQ RMS data is included in our analysis. We refer to these as CPG firms. However, some of these CPG firms also produce products outside the CPG sector (e.g. Toshiba, Samsung, Whirlpool), while others produce mostly products included in the NielsenIQ RMS data (e.g. Procter & Gamble, Kimberly Clark, Kraft). A part of our analysis is focused on identifying a sample of firms that are solely in the CPG sector. Inspired by Hoberg and Phillips (2016), we use the firm's 10-K reports, which are available from Compustat. The 10-K is a comprehensive summary of a firm's performance that must be submitted annually to the Securities and Exchange Commission, in addition to the annual report. It includes an overview of the firm's main operations, including its products and services. We manually classify each business line reported on the 10-K's into CPG/non-CPG comparing its description with the description of NielsenIQ modules, and classify each publicly traded CPG firm into CPG-only if the majority of the firm's sales results from CPG business lines. We matched 270 publicly traded companies over our sample period; we classify 23% of them as CPG-only firms.

As a robustness check, we also use the National Establishment Time Series (NETS) data, provided by Walls & Associates, which comprise annual observations on specific lines of business at unique locations over the period 1990–2014. These data allow us to track sales, employment, and industry classifications of establishments. After matching Nielsen firm names to those in NETS, we use the industry information of each establishment to classify firms as primarily operating in the CPG industry or as operating in both CPG and other sectors. The advantage of these data is that they cover nonpublicly traded firms.

A.2 Patent Data

Data Details.— Unlike other standard patent data sources such as NBER patent data (Hall et al., 2001) and the data from the Harvard Dataverse Network (Lai et al., 2014), we make use of all patents published in the USPTO, including non-granted patent applications. Using all patent applications, as opposed to just granted applications, offers us two advantages. First, since patents are usually granted with a lag of roughly two years, the more recent years of the sample suffer from severe truncation. Looking at all patent applications alleviates this problem. Second, we can then differentiate between patents that are granted, pending, or abandoned. We use this as one of the patent quality measures, as discussed below. Adding non-granted patent information increases the number of patents in our sample by 1.7 million.

Assigning Patents to Firms.— We begin by selecting all patents that have a valid assignee name.[49] We assign patents to their most recent assignee(s). For this assignment, we use the *current* assignee variable from the USPTO (as of 2017, our patent data vintage). The current assignee variable is missing for some of the patents included in our sample. In such a case, we start with the name of the original assignee and leverage the USPTO reassignment data to track any change of patent ownership due to a patent sale or firm reorganizations. To further track patent ownership through corporate reorganizations, we rely on Thomson Reuters Mergers & Acquisition data. Our underlying assumption is that patent ownership is transferred to the acquiring firm in case of corporate reorganization. Thomson Reuters M&A provides complete coverage of global mergers and acquisitions activity, including more than 300,000 US-target transactions, since 1970. The data covers mergers of equals, leveraged buyouts, tender offers, reverse takeovers, divestitures, stake purchases, spinoffs , and repurchases. It also provides detailed information about the target, the acquirer, and the terms of the deal. This comprehensiveness is particularly important given that firms that appear both in NielsenIQ data and USPTO are most likely large firms that undergo many corporate reorganizations.

Product-related and Process-related Patents.— Following Bena and Simintzi (2017), we create proxies for product-related patents and process-related patents based on the formal claims included in patent applications. Patent claims define the scope of a patent's protection and hence represent the essence of a patent application. On average, patents in USPTO have around 15 claims. Some of these are independent claims, while others derive from them. Claim texts are written in technical terms and often have a rigorous semantic structure.

The formulaic nature of claims gives us an opportunity to create the following simple classification. We say the claim is a process claim if the claim text starts with "method" phrases ("Method for", "Method of", "Method in", "Method define", and the like) or "process" phrases ("Process for", "Process according", "Process in", and the like). Then, as a baseline, we classify a patent as a process patent if the main (usually, the first) claim

[49] This step eliminates patents assigned to individuals as well as other patents that are missing assignee information, which mostly constitute pending patents.

of the patent is a process claim. The patent is a product patent if it is either a design patent or a non-process utility patent. In the latter case, claims often start with words like "Apparatus", "Device", and the like. According to this definition, up to 70% of patents are product-related patents. We also tested an alternative definition that defines process patents based on the criteria that the share of process claims is larger than 50%. These two measures are highly correlated (0.74) and our results based on the baseline variable are robust to this alternative definition.

A.3 Algorithm of *Firm Match*

Firm Name Cleaning Algorithm.— We assign each company name (from NielsenIQ or USPTO data) to a unique company identifier using the following procedure.

Step 1. In the first step, we run all company names through a name-standardization routine to generate unique company identifiers. Our routine is the following.

(1) After capitalizing all letters, we keep the first part of the company name before the first comma. (2) We remove leading and trailing instances of "THE", we replace different spellings of "AND" words with "&", and replace accented or acute letters with regular ones. (3) We remove special characters. (4) We standardize frequent abbreviations using dictionaries from the NBER Patent Data Project. For example "PUBLIC LIMITED" or "PUBLIC LIABILITY COMPANY" become "PLC";"ASSOCIATES" or "ASSOCIATE" become "ASSOC"; "CENTER" or "CENTRAL" become "CENT". (5) We delete trailing company identifiers. (6) If the resulting string is null, we protect it. (7) We repeat the previous steps on the original company names except for protected strings, for which we now keep the whole string and not just the first portion before the comma. (8) If the string is protected, we remove company identifiers in any place of the string (not just if trailing as in 5). (9) We remove spaces to further decrease misspellings. (10) We assign unique company identifiers based on the cleaned names.

Step 2. In addition to the extensive cleaning in Step 1, we take advantage of a "dictionary" that resulted from a large effort undertaken within the NBER Patent Data Project. After manual checks and searches of various company directories to identify name misspellings and various company reorganizations, the NBER files provide a mapping between patent assignee names and unique company identifiers (*pdpass*). Although this data is based on the assignees of granted patents before 2006, we use this mapping as a "dictionary" that we use in conjunction with our results from Step 1. This helps us leverage both our algorithm from Step 1 and the NBER *pdpass* information, combining the strengths of each method to create new unique company identifiers.

For example, Siemens appears in the data with many different name variations. "SIEMNES AG", "SIEMANS ATKIENGESELLSCHAFT", and "SIEKENS AG" are just a few of such variations that Step 1 does not capture, but the NBER files identify as names under the same *pdpass*. In such a case, we use *pdpass* identifiers to group the three firms together. On the other hand, the NBER file does not identify "SIEMENS CORP" "SIEMENS AG" and "SIEMENS" as the same company as the ones referenced by the first three name variations above. In such a case, we use the unique identifiers from Step 1 to group these firms together. Finally, after combining information from NBER files with our cleaning after Step 1, we pool all six variations into one new company code.

Our algorithm builds upon proven algorithms from Hall et al. (2001) and Akcigit et al. (2016). We also applied an extensive number of manual quality checks to our cleaning algorithm. For example, we identified the largest CPG firms, and for each firm we looked up the corresponding set of patents on *Google Patents* to verify that our matching algorithm was obtaining the same patents.

A.4 Algorithms of *Patents-to-Products Match*

A.4.1 *Summary of the Methods of Natural Language Processing*

For convenience, the following section summarizes general methods from natural language processing that we refer to throughout our description of the algorithms below.

i) Parsing Methods

We use 1-grams and 2-grams (single words and two-word phrases) as tokens. In general one could use n-grams, meaning distinct n-length phrases. For the types of documents we are interested in, however, meaningful and irreducible phrases having 3 or more words are quite rare. Also note that we will use the terms "word","term", and "token" interchangeably and these will refer to the set of 1-grams and 2-grams in all cases.

ii) Lemmatizer Methods

We use WordNetLemmatizer provided as part of the NLTK Python module (nltk.org), which utilizes the WordNet lexical database (`wordnet.princeton.edu`), to reduce words to their root forms by removing conjugations like plural suffixes (Fellbaum, 2010). For instance, the word "compounds" would be mapped to "compound".

iii) Word Vector Normalization

Patent (or product category) text documents are first converted into term vectors that indicate, for each term, how many times the term appears in a document. Each document vector is of length $\mathcal{M}$, which is the number of terms that we include in our vocabulary. The corpus of documents can then be represented by a very sparse matrix of term counts with elements c_{km}, where $k \in \{1, \ldots, K\} = \mathcal{K}$ represents the document (patent or a product category) and $m \in \{1, \ldots, M\} = \mathcal{M}$ represents the term.

We then use a word-based weighting scheme called total-frequency-inverse-document-frequency (tf-idf) to account for the fact that more common words tend to be less important and vice versa (Aizawa, 2003). A number of possible functional forms could be used here, but we choose the commonly used sublinear form

$$w_m = \log\left(\frac{K+1}{d_m+1}\right) + 1 \quad \text{where} \quad d_m = |\{k \in \mathcal{K} | c_{km} > 0\}|$$

Thus if a word appears in all documents, it is assigned a weight of one, while those appearing in fewer documents get larger weights, and this relationship is sublinear. For our weighting scheme, we use document frequencies from the patent data, as that corpus is considerably larger and less prone to noise.

Finally, we are left with a weighted, ℓ^2-normalized word frequency vector f_k for each document k, both on the patent and product side of our data, with elements

$$f_{km} = \frac{w_m c_{km}}{\sqrt{\sum_{m'} (w_m c_{km'})^2}}$$

A.4.2 *Step 1: Defining Product Categories*

We start by developing an intermediate categorization of NielsenIQ products into product categories that are more aggregated than product modules but less aggregated than product groups.

Step 1.a - Collect Representative Documents

For each low-level product classification from NielsenIQ (1,070 modules), we explored different sources of text that might allow us to characterize the modules. First, we studied sources of text within Nielsen. For example, we explored the use of product attributes from each UPC, and we found that while informative, some characteristics are shared and not sufficiently different. Second, we explored sources of data outside Nielsen, like dictionaries and various websites. After many manual checks, we decided to use Wikipedia pages, and based on module descriptions, we manually selected the closest Wikipedia articles for each product module.[50]

The main advantage of using Wikipedia entries is that they often include technical descriptions that use words that also appear in patent texts and are comprehensive enough to cover all modules. The use of Wikipedia text to encode textual knowledge is already common in the machine learning literature. For instance, two of the most advanced word embeddings currently available, BERT (Google, Devlin et al., 2018) and fastText (Facebook, Joulin et al., 2017), use the entire Wikipedia corpus for training purposes, in a addition to large corpora of text from books and websites. While there are a number of papers in the economics literature that study Wikipedia, we are unaware of any such usage as a direct input into a separate analysis.

For each module, we construct a representative document that includes the title of the module (repeated 10 times), the title of the Wikipedia article (10 times), the entire text (1 time), and the first 10% text of the Wikipedia article (10 times).

Step 1.b - Create Representative Word Vectors

To create the representative word vector for each module, we (i) concatenate all the text; (ii) apply the parsing and lemmatizing algorithms described above; (iii) exclude terms that appear in more than 80% of documents (to exclude words like "the" and "and"); (iv) and re-weight according to the tf-idf sublinear transformation described above.

Note that for modules that include multiple Wikipedia articles, we first vectorize each Wikipedia entry and then average these vectors together to avoid overweighting longer entries (in an ℓ^2-norm-preserving sense).

Step 1.c - Cluster Analysis

We aggregated these module vectors into clusters using the popular k-means clustering technique. k-means clustering (Lloyd, 1982) is used to find a partitioning of a vector space into clusters of similar vectors. This procedure allows one to specify the desired number of clusters K beforehand and yields a partitioning that minimizes the within-group vector variance, or the average squared distance from the cluster mean.

[50] To ensure the best selection of these articles, we cross-checked the results after assigning this task to five independent readers. Some examples of our article selection are: Wikipedia articles titled "Humidifier" and "Dehumidifier" correspond to the "Humidifier and vaporizer appliance" product module; an article "Artificial nails" is assigned to the "False nails and nail decorations" module; articles "Soft drinks" and "Carbonated water" are assigned to the "Soft drinks- carbonated" module.

Letting x be a given module vector and S_i^K be a cluster i of a cluster set S^K, we choose our partitioning S^K so as to minimize

$$\sum_{i=1}^{K} \sum_{x \in S_i^K} ||x - \mu_i||^2, \text{ where } \mu_i = \frac{1}{|S_i^K|} \sum_{x \in S_i^K} x$$

In our main analysis, we use $K = 400$ clusters. This choice is supported by extensive manual checks and experimentation with alternative partitions. We first explore k-means clustering for $K = 100, 200, ..., 900$. We find that our baseline k-means clustering partitions the product space quite well, striking a balance between minimizing the differences of vectors within a cluster while maximizing the differences across clusters.

Additionally, we show that our clustering of the product space is robust. By experimenting with various other state-of-the-art clustering techniques such as HDBSCAN (Campello et al., 2013)—a hierarchical clustering algorithm that does not need substantial tuning—we conclude that many product modules are grouped together independently of the clustering method used.

Finally, the implied clustering also accords well with the external classification scheme from NielsenIQ. By comparing our partitioning to the original 114 group aggregation from NielsenIQ (not used an input in our clustering algorithm), we see that products clustered into the same product categories also fall into same groups defined by Nielsen.

The final clustering into product categories groups together precisely those product modules that the patent matching algorithm would have trouble distinguishing between, and vice versa. For example, with this clustering, the separate product modules "Detergents – packaged", "Detergents – light duty", "Detergents – heavy duty", "Laundry treatment aids", and "Fabric washes – special" are grouped into one product category. The patent matching algorithm would struggle to accurately map a related patent to only one of these modules, especially given that the same patent could plausibly lead to innovations in all of these product modules at the same time.

Step 1.d - Creating Pseudo Product Categories

We create additional pseudo-product categories to describe products outside the consumer goods sector. These pseudo-categories are designed purely to improve the match to consumer products, as will be explained below, and are not used in our main analysis. We selected a sufficiently large and diverse set of pseudo-categories through an iterative process of experimentation and examination of patents held by firms in our sample that produce goods outside the coverage of the consumer goods data. We add 19 of these pseudo-categories to the existing 400 product categories in the data. Some examples include "computers," "car," "touchscreen", "heat engine," "software," and "airplane". As we did with the original modules, we create word vectors for each pseudo-module based on the associated set of Wikipedia articles that describe it.

Step 1.e - Word Vectors for Product Categories

The final word vector for product categories (including pseudo-product categories) simply combines the titles and Wikipedia word vectors (Step 1.b) of all modules that were clustered together to make a product category (Step 1.c).

A.4.3 *Step 2: Patent Vectors and Similarity Scores*

Step 2.1 - Collect Representative Documents for Patents

We use a variety of text fields to construct patent documents, including the title, abstract, international patent classification description, and the titles of cited patents. We upweight the title of the patent by a factor of 5 compared to the abstract, because the title has a much higher signal-to-noise ratio than the other patent text fields. Specifically, a patent's title tends to express the main application of the patent, whereas the abstract, description, and claims contain technical implementation details that are not as relevant for our purposes. For the same reasons we also upweight the patent classification description by a factor of 3.

Step 2.2 - Create Representative Vectors for Patents

To create the representative vector, we: (1) concatenate all the text; (2) apply parsing and lemmatizing algorithms (see description below); (3) exclude terms that appear in more than 80% of documents (excludes words like "the" and "and"); (4) and re-weight according to the tf-idf sublinear transformation (see description above). Constructing representative documents on the patent side consists of simply concatenating all of the available text into one document.

The patent corpus is on average shorter than the product category vectors. The average number of words per patent is 263 with standard deviation of 333 (in terms of unique words, we get mean 107 and standard deviation 93). The average number of words is about 7,200 per Wikipedia article, with a standard deviation of 6,500 (in terms of distinct words, the mean is 2,500 and standard deviation of about 2,000). We evaluated if there is a good overlap in the words used on longer documents to insure that there was not too much noise. About 50% of the words seen in our product category vectors show up in the patents somewhere.

Step 2.3 - Computing Similarity Scores Between Patents and Categories

At this point, we have the normalized word vectors for each product category j, f_{jm}, and the normalized word vectors for each patent p, f_{pm}. Multiplying any two such word vectors together yields the similarity score between two documents:

$$s_{jp} = \sum_{m \in \mathcal{M}} f_{jm} f_{pm},$$

where $\mathcal{M}$, as before, denotes size of a vector, which is the number of terms in the vocabulary. The similarity is guaranteed to lie in the range $[0, 1]$, with zero corresponding to zero word overlap and one corresponding to the case in which the documents are identical (or are multiples of one another). Notice that this vectorization approach (sometimes referred to as "bag of words") ignores any information about the order of words or phrases.

Thus, for each patent, we now have similarity metrics for each product category. The next section describes how we designate the matched product category for each patent.

A.4.4 *Step 3: Classifying Patents into Product Categories*

The final step of our patent-product matching algorithm consists in using the similarity scores to determine which pairs of patents and products are valid matches. Because some patents may correspond to certain general production processes—and not directly to products—or to products outside the consumer goods sector, we allow for the option that a patent is not assigned to any product category, or is a "non-match".

Step 3.1 - Threshold Similarity

We first adjust the algorithm to include a similarity score threshold below which we believe considering the two documents as similar would be too noisy. We tested different threshold levels and, in our baseline algorithm, we restrict the set of potential product categories for each patent p to product categories whose similarity score exceeds 0.025. For those patents that have less than five product categories satisfying this condition, we include the set of product categories that have the five highest similarity scores. For each patent, we denote the set of product categories satisfying these conditions as:

$$\Theta_p = \{j \in \Omega \mid s_{jp} > 0.025 \vee \text{rank}(s_{jp}) \leq 5\} \tag{8}$$

where Ω is the set of all product categories and s_{jp} is the similarity score between patent p and product category j.

Step 3.2 - Production Condition To further improve the match, we leverage firms' production information from Nielsen. For each patent, we define the set of potential matches, G_p, whose elements consist of all product categories in which the patenting firm ever sold a product, according to our product data.

$$G_p = \{j \in \Omega \mid p \text{ is patent of firm } i \ \wedge \ \sum_{t=2006}^{2015} \text{sales}_{ijt} > 0\}, \tag{9}$$

where $sales_{ijt}$ are the sales of firm i in product category j in year t. Note that this production condition, will exclude all pseudo-categories and product categories that the firm never produced from the set of potential matches.[51]

Step 3.3 - Select the Maximum

Together, the criteria above imply that patent p will be classified as a "non-match" if none of the product categories satisfy the thresholds and the production conditions:

$$\Theta_p \wedge G_p = \emptyset$$

For the patents that have at least one product category satisfying those conditions, we assign the final patent-product category match j_p^* to be a product category with the highest similarity score:

$$j_p^* = \max_{j \in \Theta_p \wedge G_p} s_{jp} \tag{10}$$

This defines the matching of a patent p to the set of products grouped in the category j_p^*.

[51]This makes it clear that having pseudo-categories helps to filter out many patents of the firms who heavily produce non-CPG products. For example, some firms like Toshiba or Samsung produce small electronics in our data, however they hold large portfolios of patents related to computer hardware or other high-tech technologies that are not relevant for the consumer products sector that we are analyzing. For such patents, the set Θ_p often consists only of pseudo-modules that then are easily filtered out by condition (9).

A.5 Robustness and Match Validation

A.5.1 Manual Checks of the Patent-Product Category Matches

We manually checked many patent-to-products matches; Table A1 lists some examples. The top 100 product categories sorted by their revenue and the largest firms selling in those categories are shown. For each firm, we then list an example of the highest-similarity patents in the corresponding product categories and their similarity scores. Comparing the titles of the patents and product categories, we see that product categories selected by our algorithm match the content of the patents well.

Table A1: Top Selling Firms by Categories and their Patents with the Highest Similarity Score

	Company	Product category	Application ID	Title of the Patent	Similarity
1	Philip Morris USA	Cigarette/smoking accessories	13912780	Cigarette and filter sub-assemblies with squeezable flavor capsule and method of manufacture	0.544838
2	Procter & Gamble	Diapers and baby powder	29396475	Absorbent article with a pattern	0.487175
3	Procter & Gamble	Laundry detergent	13905161	Laundry detergent composition	0.387514
4	Nikon	Camera	29385057	Projector equipped digital camera	0.33897
5	General Electric	Lamp	29283361	Lamp	0.427732
6	Coca-Cola USA	Soft drink	13816800	Phytase in ready-to-drink soft drink	0.307128
7	Procter & Gamble	Toilet	13585921	Method of reducing odor	0.191963
8	Procter & Gamble	Paper cup	11897767	Array of paper towel product	0.242879
9	Warner Home Video	Photographic film	10428440	Method of distributing multimedia presentation in different format on optical disc	0.08106
10	Procter & Gamble	Sanitary napkin	29465209	Absorbent article	0.204989
11	L'Oreal USA	Cosmetics	9987885	Anhydrous and water resistant cosmetic composition	0.305982
12	Procter & Gamble	Fabric softener	13070526	Method of making fabric softener	0.41355
13	Kimberly-Clark	Facial tissue	10034881	Method of making a high utility tissue	0.198823
14	Unilever USA	Soap	10320295	Soap wrapper	0.41769
15	L'Oreal USA	Hair coloring	14554789	Hair coloring appliance	0.455061
16	S.C. Johnson & Son	Air freshener	29438208	Dispenser	0.496183
17	Kraft Heinz Foods	Cheese	11618467	Method and system for making extruded portion of cheese	0.596449
18	Nestle Waters North America	Bottle	29434474	Water cooler	0.200115
19	The Hershey Company	Candy	9985948	Confectionary product low fat chocolate and chocolate like product and method for making them	0.282462
20	Procter & Gamble	Hair conditioner	12047712	Tool for separating a hair bundle	0.559868
21	Wm. Wrigley Jr.	Chewing gum	10453862	Method for making coated chewing gum product with a coating including an aldehyde flavor and a dipeptide sweetener	0.578689
22	Kimberly-Clark	Wet wipe	9965645	Wet wipe dispensing	0.506875
23	Procter & Gamble	Razor	29387316	Shaving razor package	0.54803
24	Activision Publishing	PC game	11967969	Video game forward compatibility including software patching	0.347854
25	Frito-Lay	Potato chip	11777839	Method for reducing the oil content of potato chip	0.521346
26	General Mills	Breakfast cereal	29183322	Layered cereal bar having cereal piece included thereon	0.28897
27	Abbott Laboratories	Milk	9910094	Powdered human milk fortifier	0.492503
28	Procter & Gamble	Toothpaste	11240284	Toothpaste dispenser toothpaste dispensing system and kit	0.388327
29	Procter & Gamble	Deodorant	12047430	Deodorant composition and method for making same	0.290906
30	The Minute Maid Company	Juice	12940252	Method of juice production apparatus and system	0.31321
31	Colgate-Palmolive	Toothbrush	11011605	Oral care implement	0.425624
32	Driscoll Strawberry Associates	Fruit	10722055	Strawberry plant named driscoll lanai	0.298149
33	The Duracell Company	Battery charger	10042750	Battery cathode	0.262253

Notes: The table presents information on the top 100 product categories according to their revenue. Each row reports the name of the highest-selling firm in a category together with an application ID and title of the firm's patent with the highest similarity score in the corresponding product category. The last column reports a similarity score from matching the patent to the category.

	Company	Product category	Application ID	Title of the Patent	Similarity
34	Alcon Laboratories	Disinfectant	9765234	Conditioning solution for contact lens care	0.362715
35	Pennzoil-Quaker State	Motor oil	10253126	Environmentally friendly lubricant	0.218752
36	Procter & Gamble	Oral hygiene	13150392	Method for whitening teeth	0.361255
37	Abbott Laboratories	Nutrition	10004360	Pediatric formula and method for providing nutrition and improving tolerance	0.108124
38	Anheuser-Busch InBev	Beer	12734356	Process for preparing a fermented beverage	0.419399
39	Procter & Gamble	Shampoo	12040980	Shampoo containing a gel network	0.386299
40	Nabisco Biscuit	Cookie	9761322	Novelty cookie product	0.155735
41	Kraft Heinz Foods	Coffee	13810612	Coffee product and related process	0.497631
42	Royal Appliance Mfg. Co.	Vacuum cleaner	10224483	Vacuum cleaner having hose detachable at nozzle	0.503479
43	Uniden Corp. of America	Mobile phone accessories	10268080	Rotating detachable belt clip	0.052147
44	Lexmark International	Ink cartridge	9766363	Ink cartridge and method for determining ink volume in said ink cartridge	0.505055
45	Gerber Products	Baby food	10295283	Blended baby food	0.24046
46	The Clorox Company	Hard-surface cleaner	12141583	Low residue cleaning solution comprising a c-to-c alkylpolyglucoside and glycerol	0.195491
47	The Clorox Company	Bleach	14724349	Intercalated bleach composition related method of manufacture and use	0.390043
48	L'Oreal USA	Cosmetic mascara	10759614	Two step mascara	0.359273
49	Lifescan	Stool test	10179064	Reagent test strip with alignment notch	0.123588
50	Playtex Products	Tampon	10834386	Tampon assembly having shaped pledget	0.558883
51	Kimberly-Clark	Urinary tract infection	12680575	Management of urinary incontinence in female	0.400734
52	Procter & Gamble	Microfiber	11016522	Rotary spinning process for forming hydroxyl polymercontaining fiber	0.113136
53	Sandisk Corporation	Floppy disk	10772789	Disk acceleration using first and second storage device	0.232516
54	Procter & Gamble	Acne	10633742	hptp-beta a target in treatment of angiogenesis mediated disorder	0.026864
55	Kraft Heinz Foods	Pasta	29220156	Spider shaped pasta	0.643155
56	L'Oreal USA	Eye liner	14368230	Method for delivering cosmetic advice	0.200779
57	Lexmark International	Printer (computing)	11766807	Hand held printer configuration	0.431107
58	Dreyer's Grand Ice Cream	Ice cream	10213212	Apparatus for forming an extruded ice cream dessert with inclusion	0.411786
59	Imation Corp.	Compact cassette	9882669	High speed tape packing	0.240291
60	Conagra Brands	Canning	12814296	Method and apparatus for smoking food product	0.144703
61	Nestle Purina PetCare	Dog food	29212029	Pet food	0.313367
62	Fort James Corporation	Disposable food packaging	29178752	Disposable plate	0.173866
63	L'Oreal USA	Face powder	9847388	Use of fiber in a care composition or a makeup composition to make the skin matte	0.139978
64	Conair Corporation	Hair styling tool	29285527	Curling iron	0.124045
65	Johnson & Johnson	Adhesive bandage	11877794	Adhesive bandage and a process for manufacturing an adhesive bandage	0.229017
66	Unilever USA	Shower gel	10242390	Viscoelastic cleansing gel with micellar surfactant solution	0.121894

	Company	Product category	Application ID	Title of the Patent	Similarity
67	Procter & Gamble	Dishwasher	11348667	Method of cleaning a washing machine or a dishwasher	0.296332
68	Pepsi-Cola North America	Tea	12147245	Coumalic acid to inhibit nonenzymatic browning in tea	0.483404
69	General Mills	Sweet roll	14340046	Method of forming dough composition	0.471588
70	Alcon Laboratories	Eye drop	9919301	Use of certain isoquinolinesulfonyl compound for the treatment of glaucoma and ocular ischemia	0.030695
71	Tyson Foods	Frozen food	13245589	Big poultry cutup method	0.311296
72	Pactiv Corp	Zipper storage bag	10289641	Reclosable bag having tamperevident member removable from the bag along a line of weakness located below the bag zipper	0.224593
73	Lipton	Margarine	9880200	Preparation of a blend of triglyceride	0.317454
74	Handi-Foil Corporation	Kitchen utensil	29418653	Pan with handle	0.167337
75	Hartz Mountain	Pet	10647660	Pet chew and method of providing dental care to pet	0.345158
76	Acco Brands USA	Notebook	11454292	Notebook computer folding ergonomic pad	0.130091
77	Johnson & Johnson	Lotion	12340858	Structured lotion	0.230563
78	Glaxosmithkline	Anti-inflammatory drug	11355808	Use of Immune cell specific conjugate for treatment of inflammatory disease of gastrointestinal tract	0.108521
79	Kraft Heinz Foods	Processed cheese	10207591	Processed cheese made with soy	0.43164
80	Fort James Corporation	Napkin	29215802	Tabletop napkin dispenser	0.263922
81	Omron Healthcare	Sphygmomanometer	29344018	Sphygmomanometer	0.463227
82	General Mills	Cracker (food)	10172401	Advertising quadrate carrier assembly with premium cradle	0.02869
83	BIC USA	Pen	29138586	Writing instrument	0.314765
84	The Libman Company	Mop	29298481	Mop	0.426008
85	Frito-Lay	Snack	10893425	Method and apparatus for layering seasoning	0.12532
86	Fresh Express Incorporated	Salad	29362982	Paper bag with a transparent vertical window for salad ingredient	0.241787
87	Procter & Gamble	Shaving cream	11110034	Shaving system with energy imparting device	0.322912
88	Nestle Purina PetCare	Litter box	29228923	Cat litter box	0.567078
89	Frito-Lay	Corn chip	9998661	Apparatus and method for making stackable tortilla chip	0.15851
90	Elizabeth Arden	Eau de toilette	29414481	Perfume bottle	0.241875
91	Bimbo Bakeries USA	Bread	13618124	Method and system for the preservation and regeneration of pre-baked bread	0.263577
92	E & J Gallo Winery	Wine	10970490	Method and apparatus for managing product planning and marketing	0.215571
93	BIC USA	Lighter	11221295	Multi-mode lighter	0.369379
94	Sara Lee Foods	Sausage	10014160	Split sausage and method and apparatus for producing split sausage	0.520147
95	Frito-Lay	Mixed nuts	11553694	Method for making a cubed nut cluster	0.158285
96	Kiss Nail Products	Manicure	12924589	Artificial nail and method of forming same	0.361627
97	Frito-Lay	Dipping sauce	10109398	Apparatus and method for improving the dimensional quality of direct expanded food product having complex shape	0.1273
98	Kraft Heinz Foods	Bacon	9799985	Bacon chip and patty	0.556922
99	Emerson Radio Corp.	Microwave oven	29149130	Protective cage and radio combination	0.0148
100	Procter & Gamble	Dentures	13043649	Denture adhesive composition	0.467318

A.5.2 External Validation. Virtual Patent Markings

One of the important validation exercises for the patent-to-products match relies on external information. We use information from virtual patent markings which were introduced with the 2011 Leahy-Smith America Invents Act. Under that act, firms may give notice to the public that their product is patented. Recently, de Rassenfosse (2018) provides estimates of the adoption rate of virtual markings and studies factors that account for the likelihood of adoption. Overall, the adoption rate is relatively small and varies systematically with firm size. Indeed, our online searches showed that only a handful of the CPG firms in our sample used virtual patent markings.[52] This means that we cannot use patent markings to match patents to products for all firms in our data set. We can, however, use them as a useful validation exercise to compare the marking's product-patent matches with our algorithm.

To this end, we selected Procter & Gamble (P&G) and Kimberly-Clark (KC) for our validation exercise, as these are among the largest firms in our sample.[53] We start by collecting the product-patent links from the websites. In most cases, the markings are associated with brands and not particular products. Hence, an important challenge lies in linking the listed brands on the websites with the brands in Nielsen. We use exact name matches, non-exact name matching, and extensive manual matching to determine the closest NielsenIQ brand equivalents. We then proceed to identify the product categories that include products of those brands. This process allows us to obtain a mapping between patents and product categories that solely comes from the markings listed by P&G and KC.(311 and 87, respectively).[54]

For each patent, we then compare the matched product categories in our patents-to-product data set with the product categories obtained from the virtual markings listed by P&G and KC. We begin by exploring similarity scores. For each patent-product category pair from the virtual markings, we obtain a similarity rank that our algorithm assigns to that product category. For example, when the rank value is one, the product category in the virtual markings corresponds to our algorithm's highest top-1 similarity category. When it is two, the match was very close to the category from the markings, and so on, thus providing a notion of closeness between the algorithm-based and marking-based matches. The first plot in Figure A.5.2 plots the distribution of these ranks. The algorithm-based preferred (highest-similarity) product categories coincide most of the time with the patent-product category mapping from the virtual markings. For 69% of patents, and 79% of patents conditional on a match, the virtual marking product categories are ranked as one or two based on the similarity scores.[55]

[52] Even if firms use virtual patent markings, they report only a selected set of products and just a small fraction of patent portfolio they hold.

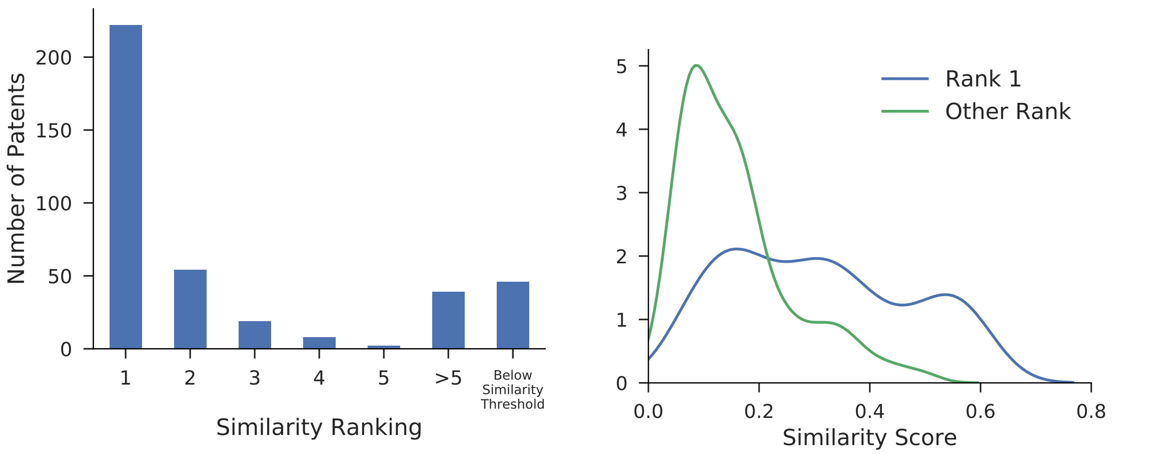
[53] We also found virtual markings are Clorox and Smuckers. However, because the products reported on their websites could not be mapped cleanly to our product categories, we did not analyze them.

[54] P&G and KC hold many more patents that are not included in the virtual markings. We also had to exclude patents listed under brands we could not match cleanly to the NielsenIQ data.

[55] Note that we cannot compare these numbers to 100% given that the ranking is unavoidably affected by some noise that comes from our manual mapping of the product listings on the websites to the notion of product categories in our data.

Figure A1: Virtual Patent Markings. P&G and KC Case Study

Distribution of similarity ranks for virtual markings — Distribution of similarity scores



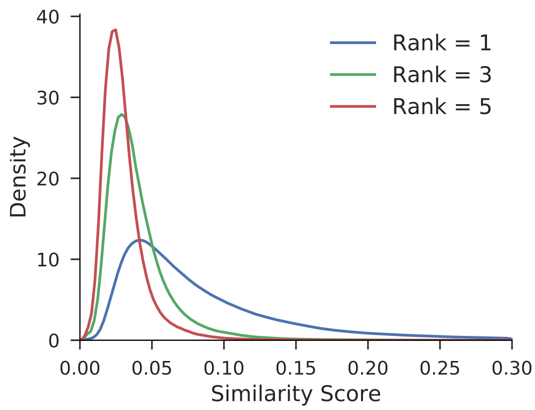
Notes: We use patent markings from P&G and KC. For each patent-product category pair from the virtual markings, we obtain a similarity rank that our algorithm assigned to this product category and show the distribution of ranks in the first graph. When the rank is one, the product category in the virtual marking corresponds to our algorithm's highest top-1 similarity category. The second graph shows the distribution of similarity scores for rank-1 and higher-rank product categories.

Another way to visualize the accuracy of the match is to examine the distribution of similarities conditioning on whether the match was the top rank-1 (coinciding with the category from virtual markings) or a lower rank. If these two distributions were very similar, this would mean that even if the match is accurate, it is not very robust, as small elements of noise or bias could change the results of the match. In fact, as shown in the second plot of Figure A.5.2, these two distributions are quite distinct, with the rank-1-match distribution weighted towards the right, meaning the results of the match should be rather robust.

A.5.3 Robustness of the Match. Patent Similarity with Top vs Lower-Rank Categories

As discussed, for our match, we pick product categories which have the highest similarity scores with patents. That is, we first pick the top five categories that have the highest similarity values with patents, and then we assign the top-similarity category conditional on a firm producing a product in that category. However, if the similarity scores for different categories are too close (either because the algorithm is not able to pick up the distinctions between documents or the categories are too finely defined) so that the algorithm cannot clearly differentiate between them, our choice of the top-rank match would not be robust to small perturbations of the algorithm or category clustering. To explore this issue, we plot the distribution of similarity scores of patents with different-rank product categories in Figure A2. The rank-1 category is the category with the highest similarity score for a patent, and so on. We find that top-ranked categories have substantially different (shifted to the right) distributions than slightly lower-ranked categories, thus providing evidence of the robustness of the match. The patents' mean similarity score for rank-1 categories is 3 times higher than the mean similarity score for rank-5 categories.

Figure A2: Similarity Distribution by Rank

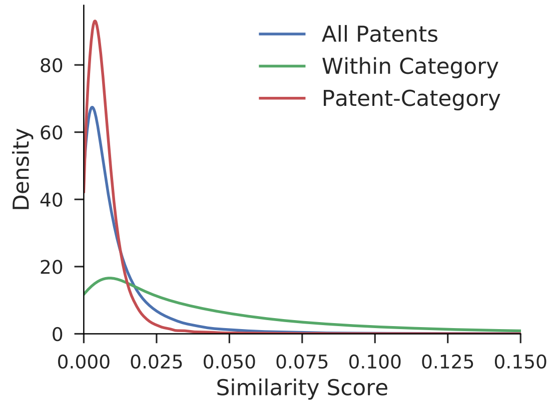


Notes: The figure shows similarity scores distribution of patents for different-rank product categories. Rank-1, Rank-3, and Rank-5 show similarities with categories ranked as the highest, rank-3, and rank-5 similarity categories.

A.5.4 Actual vs Placebo Match of Patents to Product Categories

We next verify that by grouping patents into distinct categories, we are indeed carving out well-defined neighborhoods in the technological space. We again employ word vectors to assess document similarity, but this time between pairs of patent texts. Specifically, we look at the distribution of similarity scores between pairs of patents classified into the same product category and compare this distribution to that of pairs of patents selected at random from the entire set of patents held by CPG firms. The similarity distribution based on this match looks very different from our placebo distribution as seen in Figure A3. The patents' mean similarity score is 5.6 times higher if patents are assigned to the same product categories. In ordinal terms, the median within-category similarity lies at the 93rd percentile in the overall distribution.

Figure A3: Distribution of Pairwise Patent Similarities



Notes: The green density curve shows the distribution of similarities between pairs of patents classified into the same product category. The blue curve shows the distribution of similarities between randomly drawn pairs of patents amongst all those owned by NielsenIQ firms.

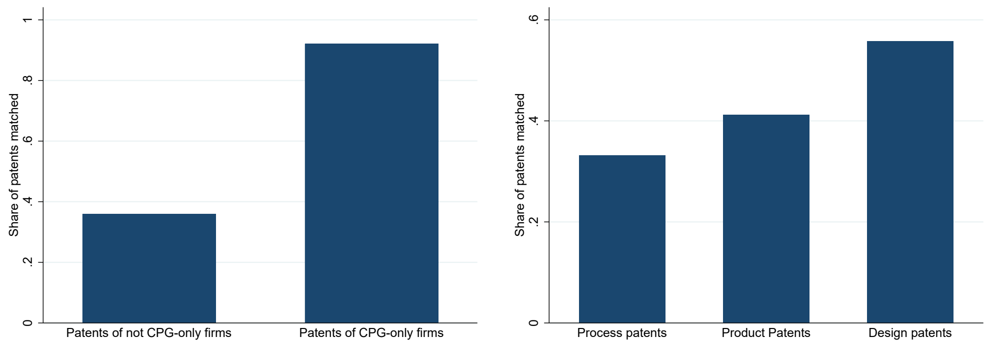
A.5.5 Validating Non-matches. CPG-only Firms and Product-Related Patents

Our patents-to-products dataset at the firm × product category level would ideally filter out patents that are not related to the products in our data. Hence, correct non-matches would arise for the following two main reasons. First, a patent may relate to other non-CPG goods that the firm may be producing, which are not covered in our sample; and second, a patent may be a general process/method patent that does not relate to the products directly. We examine these possibilities.

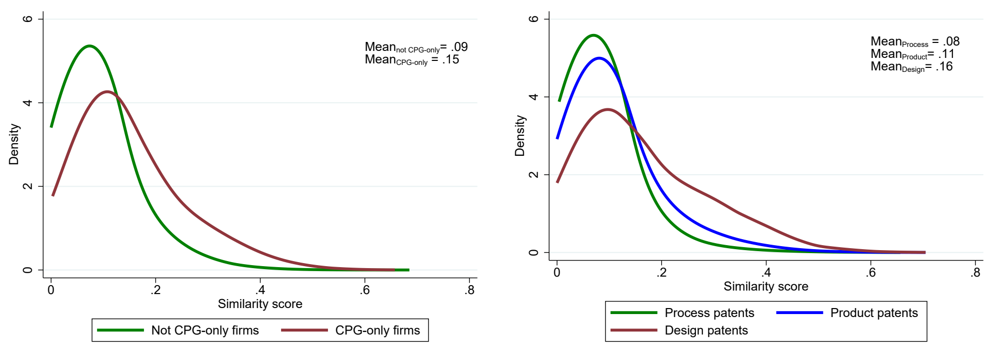
Panel (a) in Figure A.5.5 shows the share of patents that match to firms' product categories for a sample of firms that we can accurately identify as CPG-only firms and not CPG-only firms (see Appendix A.1 for details). Indeed, 92% of patents held by CPG-only firms match, while only 36% of patents of not CPG-only firms match to our product categories. This result reassures us that our algorithm indeed picks the correct matches. As seen from Panel (b), the similarity scores for CPG-only firm patents are also significantly higher.

Figure A4: Match Validation. CPG-only Firms and Product-Related Patents

(a) Share of patents matching to firms' product categories



(b) Rank-1 similarity distribution for patents

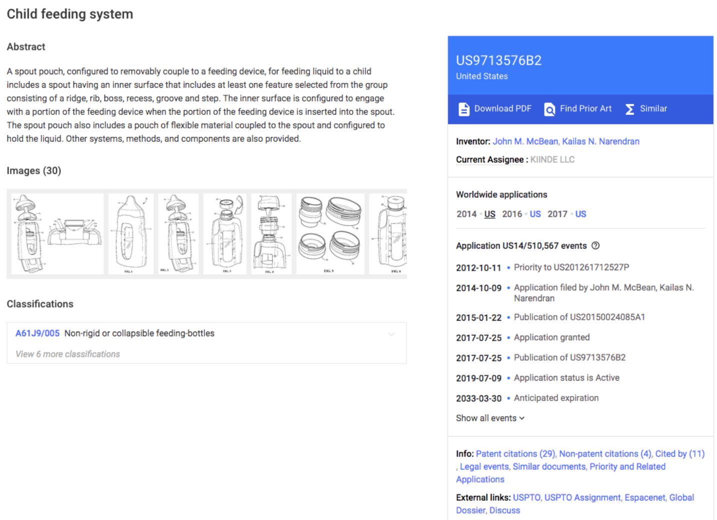


Notes: Panel (a) shows the share of patents that match with product categories in which firms ever sell a product. The left figure compares patents of the CPG-only and non-CPG-only firms, while the right figure compares process, product-related, and design patents. CPG-only firms and non-CPG-only firms refer to the sample of firms defined in Appendix Section A.1. Process and product-related patents are defined in Appendix Section A.2. Panel (b) displays the similarity score distribution for patents of CPG-only and non-CPG-only firms on the left and of process, product-related, and design patents on the right.

Panel (a) also demonstrates that the share of patents that are matched is higher if the patent is more likely to be directly related to products. Using our proxies for process- and product-related patents (see Appendix A.2 for details) and considering design patents as most directly related to products, we plot the share of all process, product, and design patents that are matched. The probability of a match increases along with the likelihood of a patent being related to a product, which is reassuring. Panel (b) also confirms that the similarity scores of product-related patents are much higher than the similarity scores of process patents.

A.6 Patents and Products in CPG. Examples.

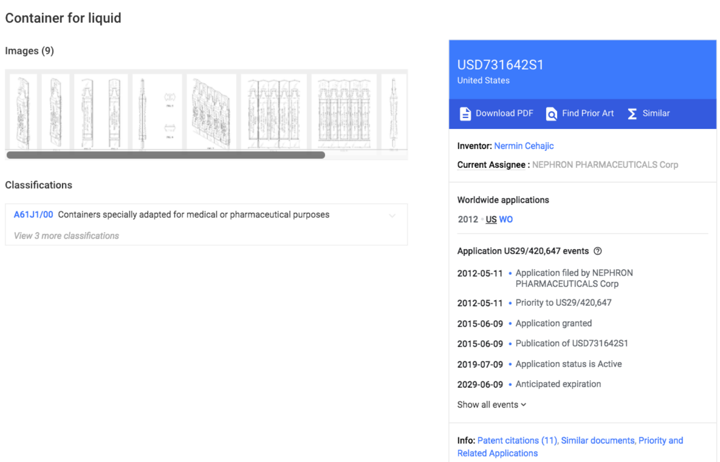
Figure A5: Example: Kiinde LLC



(a) Patent application in 2013

(b) Direct pump adapters introduced in 2014

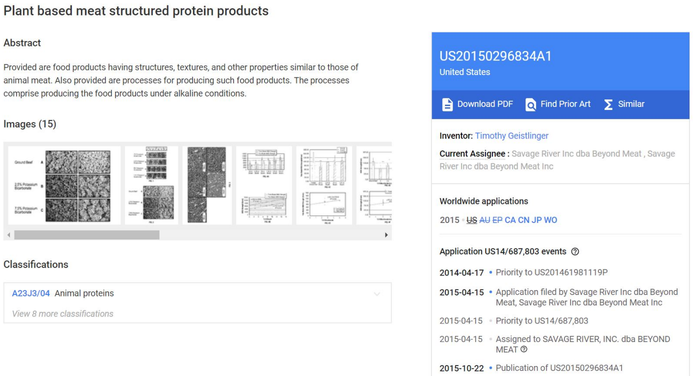
Figure A6: Example: Nephron Pharma



(a) Patent application in 2012

(b) Refill vials in 2012

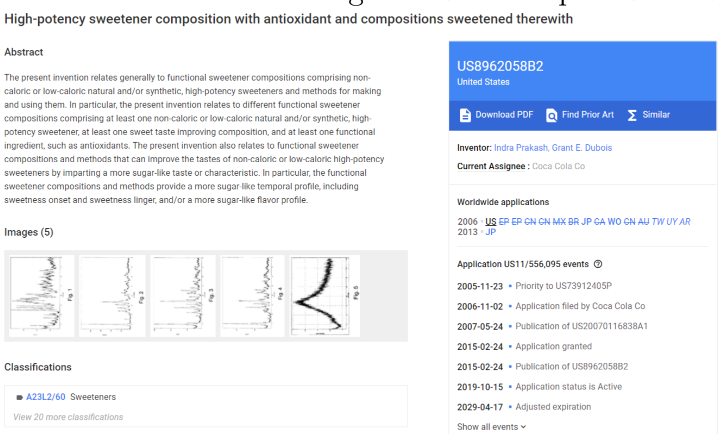
Figure A7: Example: Beyond Meat Inc.



(a) Patent application in 2014

(b) The first simulated beef product in 2014

Figure A8: Example: Coca Cola Inc.



(a) Patent application in 2005

(b) The first sugar-free Coke Zero in 2005

A.7 Measuring Product Innovation

We use four measures of quality improvements brought by new products: a novelty index whose weights are the contributions of each attribute to the product price (baseline q); a novelty index that equally weights each attribute ($q1$); a novelty index whose weights reflect the total sales accounted by each attribute ($q2$); and a quality measure that weights each product by its residual demand ($q3$). These measures capture different dimensions of quality. The first three measures (baseline q, $q1$ and $q2$) explicitly capture the novelty of a new product by using information about its attributes. The second type of measure ($q3$) captures any residual demand (or appeal), which can arise from vertical quality differentiation or subjective differences in consumer taste. We next describe the construction of the novelty-based measures and residual demand in detail, followed by a discussion of the descriptive statistics for these measures.

A.7.1 Novelty-Based Measures

Overview. — We define a product u in product category j as a vector of characteristics $V_u^j = \left[v_{u1}^j, v_{u2}^j, .., v_{uA^j}^j\right]$, where A^j denotes the number of attributes (e.g. color, formula, size) observed in product category j and v_{ia}^j represents a characteristic within an attribute (e.g. blue, red, green).[56] Let Ω_t^j contain the set of product characteristics for each product ever sold in product category j at time t, then the *novelty index* of product u in product category j, launched at time t is defined as follows:

$$q_u \equiv \text{Novelty}_{u(t)}^{(j)} = \sum_{a=1}^{A^j} \omega_a^j \mathbb{1}[v_{ua}^j \notin \Omega_t^j],$$

where ω_a^j represents the category-specific weight given to new characteristics within attribute a. The measures q, $q1$ and $q2$ only differ in the way we compute their ω_a^j.

For q, we estimate ω_a^j using hedonic price regressions in order to be able to quantify the importance of each attribute within a product category. The section below provides the details on the hedonic methods used.

The simplest measure $q1$, simply weights each attribute equally. For example, if a new product within the "pain remedies-headache" category enters the market with a flavor and formula that has never been sold before, its novelty index is $(1+1)/A^{\text{soft drinks}} = 2/10$. Note that comparing the novelty index of different products across distinct categories depends not only on the number of new attributes of each product, but also on the total amount of observable characteristics the NielsenIQ data provides for each category.

Measure $q2$ is very similar to q. We use weights generated by hedonic regressions and scale them by observed quantities to get to the sales-based weights for each attribute. In this case, we also normalize the weights so that all weights within a product category add up to one.

[56] We refer to product categories for simplicity of notation. Our analysis is conducted first at the product module level (as defined by NielsenIQ RMS data) and then aggregated at the firm level (firm match) or firm $\times$ product category level (patents-to=products match).

Hedonic Regression Weights. — We estimate product category weights ω_a^j for our measure q using hedonic methods. In particular, we estimate a linear characteristics model using the time-dummy method. The time-dummy method works by pooling data across products and periods and regressing prices on a set of product characteristics and a sequence of time-dummies. Since the regression is run over data which is pooled across time periods, any product characteristic which is held by at least one product in some period can be included even if it is not present in all periods. The estimated regression coefficients represent the shadow price for each of the included characteristics. To implement this method, we estimate the following equation by non-negative least squares:

$$p_{ut} = \sum_c \pi^c a_u^c + \lambda_t + \epsilon_{ut}, \tag{11}$$

where u denotes the product, c is the characteristic, and t is the time period (years). a_u^c is an indicator that equals one if a given characteristic c is present in product u. Recall that each attribute a (e.g color) has distinct characteristics c (e.g. blue, red). The estimated regression coefficients, π^c, represent the shadow price for each of the included characteristics. We use non-negative least squares so that the shadow prices are weakly positive. Lastly, λ_t represents time effects.

Using this method, we obtain a correlation of approximately 0.91 between the actual price and $\sum_c \pi^c$.[57] The weight ω_a^j is the average contribution of the characteristics within each attribute to the price normalized so that $\sum_a^{A^j} \omega_a^j = 1$; these are the weights used in our baseline novelty index.

A.7.2 Residual Demand Measure

An alternative way of measuring the degree of product innovation brought by new products to the market is to weight them by their implied quality (or residual demand) using a structural specification of their demand function. To derive an implied quality for each product, we follow Hottman et al. (2016) and Argente et al. (2018) and use a nested constant elasticity of substitution (CES) utility system that allows the elasticity of substitution between varieties within a firm to differ from the elasticity of substitution between varieties supplied by different firms. The model features oligopolistic competition with a finite number of heterogeneous multi-product firms, where the output of each category is described by a nested CES structure over a finite number of products within a finite number of firms (j is omitted for simplicity of notation)

$$y = \left(\sum_{i=1}^{M} \left(\sum_{u=1}^{N_i} \left(\gamma_{ui} y_{ui} \right)^{\frac{\sigma-1}{\sigma}} \right)^{\frac{\sigma}{\sigma-1}\frac{\eta-1}{\eta}} \right)^{\frac{\eta}{\eta-1}}$$

[57]These dummies for characteristics seem to explain differences in prices well. The variance of linear combination of the fixed effects of the attributes (excluding time fixed-effects) relative to the variance of the prices is 0.827.

where σ is the elasticity of substitution across products within the same firm, η is the elasticity of substitution across firms, and γ_{ui} and y_{ui} are the implied quality and quantity of product u produced by firm i, respectively. Using the first order conditions of the consumer we can write the demand for product u produced by firm i as follows:

$$y_{ui} = (\gamma_{ui})^{\sigma-1} \left(\frac{p_{ui}}{p_i}\right)^{-\sigma} \left(\frac{p_i}{p}\right)^{-\eta} \frac{Y}{p} \quad , \qquad p_i = \left(\sum_{u=1}^{N_i} \left(\frac{p_{ui}}{\gamma_{ui}}\right)^{1-\sigma}\right)^{\frac{1}{1-\sigma}} , \tag{12}$$

where the demand for the product depends on the implied quality γ_{ui} and price p_{ui} of the product, as well as the firm's price index p_i, the category's price index p, and the size of the category Y. Conditional on observing the prices and quantities from the data and obtaining estimates for σ and η, we recover γ_{uijt} as a structural residual that ensures that the model replicates the observed data up to a normalization.[58] We normalize the implied quality so that its geometric mean within each category and time period equals one. The key advantage of this normalization is that we can compare a product's implied quality within the firm and across firms within a category and time period. Using this normalization and equation 12, we obtain the product implied quality as:

$$\gamma_{ui} = \left(\frac{s_{ui} \times s_i}{\prod_{u,j}(s_{ui} \times s_i)^{\frac{1}{M}}}\right)^{\frac{1}{\sigma-1}} \left(\frac{s_i}{\prod_{u,j}(s_i)^{\frac{1}{M}}}\right)^{\frac{\sigma-\eta}{(1-\eta)(1-\sigma)}} \left(\frac{p_{ui}}{\prod_{u,j}(p_{ui})^{\frac{1}{M}}}\right),$$

where s_{ui} and s_i are the share of sales of product u and the share of sales of firm i, respectively, and M denotes the total number of products sold in a category. The estimation procedure for σ and η follows Broda and Weinstein (2010) and Feenstra (1994). The estimation has two steps. In the first step, we estimate the elasticity of substitution across products within firms using product shares, product prices, and firm shares using a GMM procedure. The key identification assumption is that demand and supply shocks at the product level are uncorrelated once we control for firm-time specific effects. In the second step, we use these estimates for products to estimate the elasticity of substitution across firms for each category using the procedure developed by Hottman et al. (2016). We use the estimates from Argente et al. (2018).

To capture the incremental effect of new products on the residual demand of the firms, our measure of quality improvement $q3$ is the geometric average of the implied quality of the new products relative to the geometric average of all products sold by the firm.

A.7.3 Descriptive Statistics

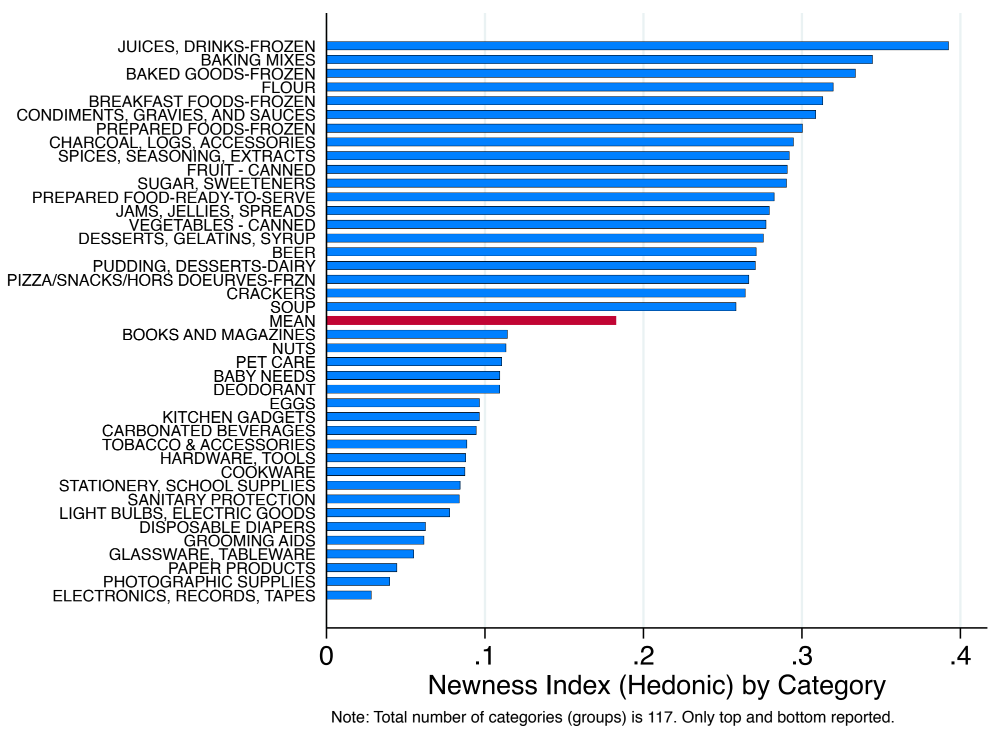
Novelty Indices Across Product Categories. — Although in our main analyses we only consider within category variation in novelty, Figure A9 shows the degree of heterogeneity in novelty index q across different product categories. The quality measure q has a correlation of 0.93 with the equal-weights measure $q1$. Conditional on having an equal-weights index larger than zero, the correlation is 0.79. "Juices, Drinks-frozen" has a high novelty index mainly due to the prevalence of new brands and new flavors over our sample

[58] Normalization is required because the utility function is homogeneous of degree 1 in the implied quality.

period. Over our sample period, there are more than 50 new brands and 67 new flavors in this category, which can be explained by recent trends in this category to increase the nutrients, reduce the sugar content, and to create products according to the consumer's lifestyles. The novelty index for "Baking Mixes" and "Flour" can be explained by the surge in home-based baking observed in recent years, which led to more than one thousand new brands in these categories. Only in "Baking Mixes" we observed more than 600 new flavors during our sample period. These categories have also seen significant innovations in packaging. An example is stand-up pouches, which use less plastic, increase the shelf life of products, and reduce the likelihood they are damaged during shipping.

Figure A10 shows some examples of products with high and low equal-weights novelty $q1$ in our data. For example, the product Asthmanefrin Inhalation Solution - Liquid Refill is part of the group Medications/Remedies/Health Aids. When it was introduced in the market, this product had six of the eight attributes that we observe in our data for that product group, it was a new brand, launched by a new firm, it was a liquid, bronchilator refill. As a results, its equal-weights novelty index is 6/8=0.75.

Figure A9: Novelty Index (Baseline q)



Notes: The figure presents the average novelty index for a sample of product groups in our data. In particular, it shows the mean novelty index by groups along with the top and bottom groups as ranked by this measure. We compute the novelty index for each product using equation A.7.1. We average across products and product modules to the category level. We focus on cohorts from 2006Q3 to 2014Q4 and on modules with at least 20 barcodes.

Figure A10: Novelty Index: Examples

(a) High-Novelty Products (b) Low-Novelty Products

Correlation with Product and Firm Performance. — Our baseline measure of quality q explicitly captures the novelty of a new product by using information about its attributes. This use of product attributes offers important advantages in the context of our paper. Patents are granted on the basis of novelty, and thus using a quality-adjusted measure of product introduction that explicitly accounts for new features of the product should help to align the notion of innovation on patents and products side. However, new features of the product may not affect the market at all if they are not valued by customers. Our baseline measure q partially accounts for this potential concern by weighting any new characteristic according to its shadow price using hedonic regressions. In addition, Table A2 shows that our baseline measure is correlated with product and firm outcomes, and thus may be capturing some vertical quality differentiation or subjective differences in consumer taste.

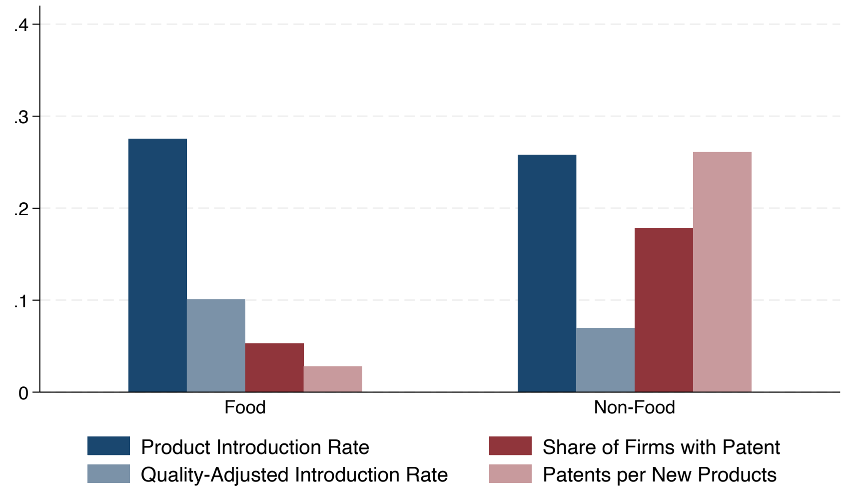
Table A2: Novelty Measure: Correlation with Firm Outcomes

	(1) Growth Rate	(2) Growth Rate New	(3) Duration 4q	(4) Duration 16q
Novelty(t)	0.0815**	1.6463***	0.1762***	0.1879***
	(0.032)	(0.063)	(0.014)	(0.024)
Product Introduction(t)	1.0351***	2.0674***	0.0855***	0.0964***
	(0.011)	(0.019)	(0.004)	(0.007)
Observations	408,161	241,540	96,942	53,611
R-squared	0.367	0.363	0.478	0.573
Time-Category	Y	Y	Y	Y
Firm-Category	Y	Y	Y	Y

Notes: The table shows the correlation between our measure of novelty q and several firm outcomes. $Growth\,rate\,(DH)$ is the revenue growth of the firm estimated as in Davis and Haltiwanger (1992), i.e. $2(y_t - y_{t-1}/(y_t + y_{t-1})$. $Growth\,rate\,New\,(DH)$ is the revenue growth of new products. $Duration\,4q$ and $Duration\,16q$ are the share of products introduced a time t that last in the market more than 4 or 16 quarters respectively. $log\ N$ is log number of products introduced using the inverse hyperbolic sine transformation. Standard errors are clustered at firm $\times$ category.

B Additional Empirical Results

Figure B1: Main Summary Statistics for Food and Non-Food Categories



Notes: The figure presents summary statistics for food and non-food product categories in Nielsen. Using the patents-to-products match, we compute the average product introduction rate, quality-adjusted introduction rate, number of new products, and patent applications at the firm × category level (patent statistics are winsorized at the top 5%). With these, we compute average product introduction rate, quality-adjusted introduction rate, patents per new products and share of firms with patents at the product category-level. The plot shows these statistics by aggregating them within food and non-food product category (weighting by the total revenue of each product category).

Table B1: Product Introduction and Patenting, Logs

	Log Product Introduction			Log Product Introduction Quality-Adjusted		
	(1)	(2)	(3)	(4)	(5)	(6)
Log patents(t-1)	0.0380***			0.0189***		
	(0.010)			(0.005)		
Log patents granted(t-1)		0.0405***			0.0192***	
		(0.011)			(0.005)	
Log patents citations adj.(t-1)			0.0249***			0.0139***
			(0.007)			(0.003)
Observations	409,641	409,641	407,891	409,641	409,641	407,891
R-squared	0.692	0.692	0.690	0.623	0.623	0.619
Time-Category	Y	Y	Y	Y	Y	Y
Firm-Category	Y	Y	Y	Y	Y	Y

Notes: The table shows regressions of the product introduction rates and quality-adjusted product introduction rates as a function of the patenting rates, using firm × category × year data. Product and quality-adjusted product introduction rates are defined as the number of new products or quality-adjusted new products over a total number of products in firm × category × year. Product quality measures are defined in Section 2.3.1. *Patents* is the ratio of the firm's number of patent applications in a particular category-year over the total number of cumulative patents in that category-year; *Patents granted* is the ratio of the firm's number of granted patent applications in a particular category-year over the total number of cumulative patents in that category-year; *Patents citation-adjusted* is the ratio of the firm's number of citations-weighted granted patents in a particular category-year over the total number of citation-weighted granted patents in that category-year. Observations at the firm × category × year level with zero patents are included in the regression. Standard errors are clustered at the firm × category level.

Table B2: Product Innovation and Patenting: Food and Non-Food Categories

	Product Introduction			Product Introduction Quality-Adjusted		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel 1 - Food						
Patents(t-1)	0.0293**			0.0089**		
	(0.014)			(0.004)		
Patents granted(t-1)		0.0416**			0.0100	
		(0.020)			(0.006)	
Patents citations adj.(t-1)			0.0402			0.0108
			(0.027)			(0.008)
Observations	259,799	259,799	259,468	259,799	259,799	259,468
R-squared	0.328	0.328	0.328	0.307	0.307	0.307
Time-Category	Y	Y	Y	Y	Y	Y
Firm-Category	Y	Y	Y	Y	Y	Y
Panel 2 - Non-Food						
Patents(t-1)	0.0506***			0.0208***		
	(0.009)			(0.003)		
Patents granted(t-1)		0.0526***			0.0263***	
		(0.012)			(0.005)	
Patents citations adj.(t-1)			0.0608***			0.0309***
			(0.015)			(0.006)
Observations	149,842	149,842	149,025	149,842	149,842	149,025
R-squared	0.382	0.382	0.383	0.295	0.295	0.296
Time-Category	Y	Y	Y	Y	Y	Y
Firm-Category	Y	Y	Y	Y	Y	Y

Notes: The table shows regressions of the product introduction rates and quality-adjusted product introduction rates as a function of the patenting rates, using firm × category × year data. Product and quality-adjusted product introduction rates are defined as the number of new products or quality-adjusted new products over a total number of products in firm × category × year. Product quality measures are defined in Section 2.3.1. *Patents* is the ratio of the firm's number of patent applications in a particular category-year over the total number of cumulative patents in that category-year; *Patents granted* is the ratio of the firm's number of granted patent applications in a particular category-year over the total number of cumulative patents in that category-year; *Patents citation-adjusted* is the ratio of the firm's number of citations-weighted granted patents in a particular category-year over the total number of citation-weighted granted patents in that category-year. The results for food include the departments of dry grocery, frozen foods, dairy, deli, packaged meat, fresh produce and alcoholic beverages; and for non-food include health and beauty, non-food grocery, and general merchandise. Standard errors are clustered at the firm × category level.

Table B3: Product Introduction and Patenting by Size: Quality Measures

	(1)	(2)	(3)	(4)	(5)	(6)
	Product Introduction Quality-Adjusted					
	q1 Novelty Simple		*q2* Novelty Sales		*q3* Residual Demand	
Patents(t-1)	0.0117***	0.0356***	0.0139***	0.0362***	0.0115	0.1680*
	(0.002)	(0.006)	(0.002)	(0.008)	(0.016)	(0.096)
Size(t)		0.0002***		0.0004***		0.0073***
		(0.000)		(0.000)		(0.002)
Patents(t-1) x Size(t)		-0.0021***		-0.0020***		-0.0109*
		(0.000)		(0.001)		(0.006)
Observations	409,641	409,641	409,641	409,641	83,329	83,329
R-squared	0.263	0.264	0.262	0.262	0.324	0.324
Time-Category	Y	Y	Y	Y	Y	Y
Firm-Category	Y	Y	Y	Y	Y	Y

Notes: The table shows regressions of the quality-adjusted product introduction rates as a function of the patenting rates, using firm $\times$ category $\times$ year data, similar to Table 4 but introducing size (firm sales in category-year) and size interaction with patenting. Quality-adjusted product introduction rates are defined as the number of new products or quality-adjusted new products over a total number of products in a firm $\times$ category $\times$ year. Quality is defined as follows: *q1* is a quality measure that weighs each attribute equally, *q2* is a weighted quality measure using weights that reflect "shadow sales", and *q3* is a measure of residual demand taken from Hottman et al. (2016). This measure does not use information about the degree of novelty of a product and instead captures the appeal of new products relative to other products sold in the market, under some functional-form assumptions. *Patents* is the ratio of the firm's number of patent applications in a particular category-year over the total number of cumulative patents in that category-year. Observations at the firm $\times$ category $\times$ year level with zero patents are included in the regression. Standard errors are clustered at the firm $\times$ category level.

Table B4: Product Introduction and Patenting: International Patents

	(1)	(2)	(3)	(4)	(5)	(6)
	Product Introduction			Product Introduction Quality-Adjusted		
Patent Rate(t-1)	0.0446***	0.0864***	0.0867***	0.0174***	0.0460***	0.0460***
	(0.008)	(0.023)	(0.023)	(0.003)	(0.008)	(0.008)
Size(t)		0.0102***	0.0102***		0.0012***	0.0012***
		(0.000)	(0.000)		(0.000)	(0.000)
Patent Rate(t-1) x Size(t)		-0.0037**	-0.0038**		-0.0025***	-0.0025***
		(0.002)	(0.002)		(0.001)	(0.001)
Share of Int. Patent	0.0068	0.0063	-0.0119	0.0013	0.0014	0.0025
	(0.006)	(0.006)	(0.019)	(0.002)	(0.002)	(0.006)
Share of Int. x Size(t)			0.0014			-0.0001
			(0.001)			(0.000)
Observations	409,641	409,641	409,641	409,641	409,641	409,641
R-squared	0.357	0.362	0.362	0.302	0.303	0.303
Time-Category	Y	Y	Y	Y	Y	Y
Firm-Category	Y	Y	Y	Y	Y	Y

Notes: The table shows regressions of the product introduction rates and quality-adjusted product introduction rates as a function of the patenting rates, using firm × category × year data, similar to Table 4 but introducing size (firm sales in category-year), size interaction with patenting, and controlling for the share of international patents. Product and quality-adjusted product introduction rates are defined as the number of new products or quality-adjusted new products over the total number of products in the firm × category × year. Product quality measures are defined in Section 2.3.1. *Patents* is the ratio of the firm's number of patent applications in a particular category-year over the total number of cumulative patents in that category-year. A USPTO patent is classified as international if the firm files for protection in at least one country outside the United States. Observations at the firm × category × year level with zero patents are included in the regression. Standard errors are clustered at the firm × category level.

Table B5: Sales, Patents, and Product Introduction: Quantity and Prices

	(1) Log Quantity	(2) Log Quantity	(3) Log Price	(4) Log Price
Log Cum Patents(t-1)	0.069***	0.072***	0.016**	0.017***
	(0.022)	(0.022)	(0.007)	(0.007)
Log New Products	0.073***		0.022***	
	(0.004)		(0.001)	
Log Q-new Products		0.099***		0.035***
		(0.009)		(0.002)
Observations	407,937	407,937	407,937	407,937
R-squared	0.855	0.855	0.921	0.921
Time-Category	Y	Y	Y	Y
Firm-Category	Y	Y	Y	Y
Controls	Y	Y	Y	Y

Notes: The table repeats columns (3) and (4) of Table 7 for quantities and prices. Columns (1)-(2) report regressions for log quantity at time t, controlling for lagged log quantity; columns (3)-(4) do the same for log price. Since the baseline unit of analysis is a product category, which aggregates multiple NielsenIQ modules, and often the quantities and prices are not directly comparable across modules within a category, we normalize each product's quantity and price relative to the median within its module, then compute the firm-level log quantity and log price as the average of these normalized (logged) values within each category.

Table B6: Sales, Patents, and Product Introduction: Flow of Patents

	(1)	(2)	(3)	(4)	(5)	(6)
	Log Sales (t)				Log Sales (t)	
Log Patents(t-1)	0.044**		0.029	0.033*	-0.136**	-0.074
	(0.020)		(0.019)	(0.019)	(0.069)	(0.074)
Log New Products		0.517***	0.517***		2.024***	
		(0.006)	(0.006)		(0.023)	
Log Q-new Products				0.760***		3.452***
				(0.013)		(0.054)
Log Patents(t-1) × Size(t-1)					0.013***	0.009*
					(0.005)	(0.005)
Log New Products × Size(t-1)					-0.130***	
					(0.002)	
Log Q-new Products × Size(t-1)						-0.219***
						(0.004)
Observations	408,161	408,161	408,161	408,161	408,161	408,161
R-squared	0.900	0.905	0.905	0.902	0.910	0.905
Time-Category	Y	Y	Y	Y	Y	Y
Firm-Category	Y	Y	Y	Y	Y	Y
Log Sales(t-1)	Y	Y	Y	Y	Y	Y

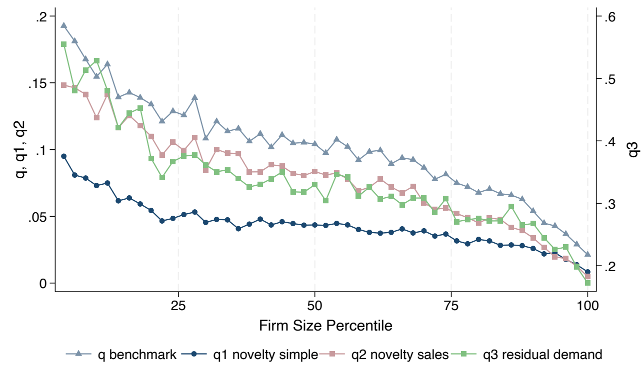
Notes: The table replicates the Table 7, but using *Log Patents(t-1)*—log number of patent applications at t-1, instead of the stock of patents. The table shows regressions of log sales on patent stock in the previous period and the product introduction, conditional on firm sales in the previous period, using firm × category × year data. Baseline regression specifications in (3) and (4) are based on equation (3). All regressions additionally control for time-category and firm-category fixed effects. Standard errors are clustered at the firm × category level.

Table B7: Sales on Patents, and Product Introduction over Time

	Log Sales (t)			
	2006-2010		2011-2015	
Log Cum Patents(t-1)	0.026	0.132	-0.188**	-0.092
	(0.139)	(0.145)	(0.083)	(0.090)
Log New Products	1.692***		1.874***	
	(0.036)		(0.031)	
Log Q-new Products		2.606***		2.997***
		(0.080)		(0.068)
Log Cum Patents(t-1) × Size(t-1)	0.012	0.006	0.022***	0.016***
	(0.009)	(0.010)	(0.005)	(0.006)
Log New Products × Size(t-1)	-0.114***		-0.123***	
	(0.003)		(0.002)	
Log Q-new Products × Size(t-1)		-0.171***		-0.192***
		(0.006)		(0.005)
Observations	175,566	175,566	226,843	226,843
R-squared	0.931	0.929	0.928	0.925
Time-Category	Y	Y	Y	Y
Firm-Category	Y	Y	Y	Y
Log Sales(t-1)	Y	Y	Y	Y

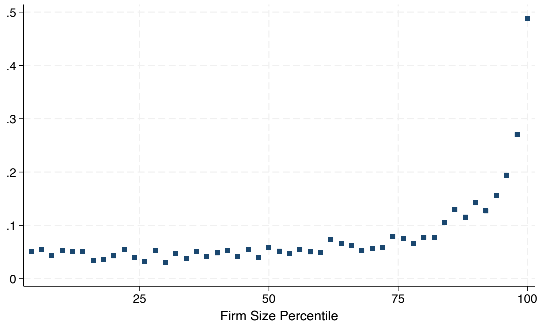
Notes: The table shows regressions of log sales on patent stock in the previous period and the product introduction, conditional on firm sales in the previous period, using firm × category × year data. *Log Cum Patents(t-1)* is the log number of patent applications by time t-1–net of depreciation, and the *Log New Products (Q-new Products)* is the number of new products (quality-adjusted new products) introduced at time t. To account for zeros, logs are calculated using the hyperbolic sine transformation. Baseline regression specifications in (3) and (4) are based on equation (3). The last two columns interact patent stock and product introduction variables with firm log size (sales) in the previous period. All regressions additionally control for time-category and firm-category fixed effects. Standard errors are clustered at the firm × category level.

Figure B2: Product Innovation Rate by Size: Alternative Quality Adjustments

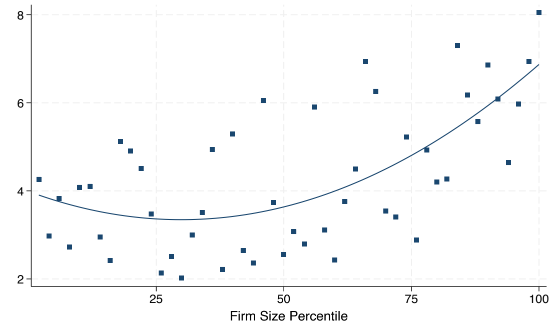


Notes: The figure plots the quality-adjusted innovation rates over firm size percentiles for different measures of quality. For each firm × product category, we compute average sales and define firm size percentiles based on the average sales distribution in that product category. Each panel plots the average value of the respective variables in each percentile. For each firm × product category, we compute their average sales and quality-adjusted product entry rates using our benchmark and three alternative quality measures—$q1, q2, q3$. By construction, q, $q1$, and $q3$ lie between 0 and 1. $q3$ is plotted on a different axis because its mean is equal to 1, as it is the geometric average of the implied quality of the new products relative to the geometric average of all products sold by the firm. Each dot/triangle plots the averages after weighting different product categories by their importance in the whole sector, as measured by their sales share.

Figure B3: Patenting and Firm Size



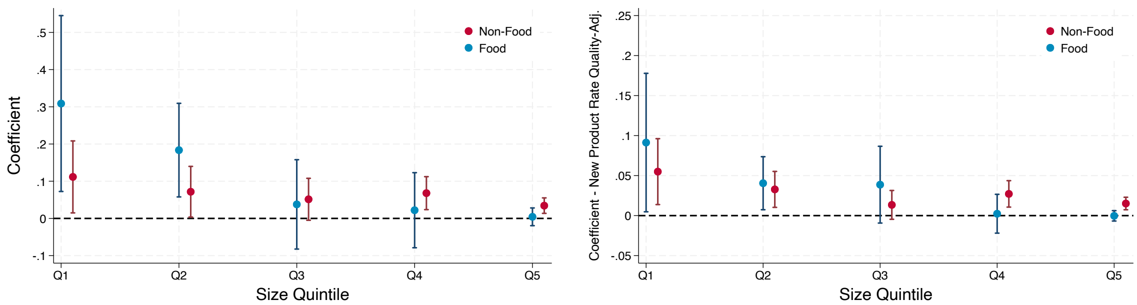
(a) Probability of patent application



(b) Number of patent applications (log)

Notes: This figure plots the relationship between patenting and firm size, defined by sales. For each firm × product category, we compute average sales and define firm size percentiles based on the average sales distribution in that product category. Each panel plots the average value of the respective variables in each percentile. For each firm × product category, we compute the probability of having filed a patent and the average number of patent applications on file. Within each product category, we compute the average probability and number of patents ×1000 (log) for each bin. Each dot/triangle plots averages after weighting different product categories by their importance in the whole sector, as measured by their sales share.

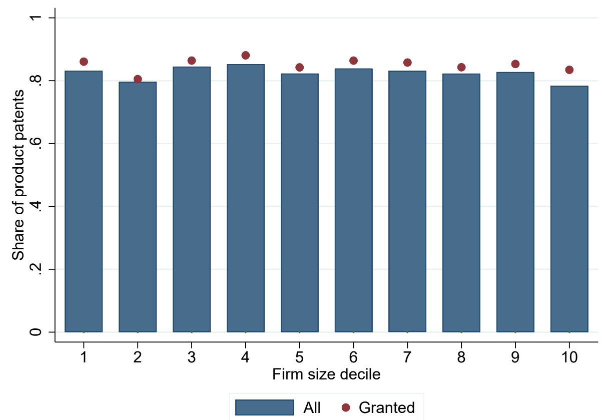
Figure B4: Product Innovation and Patenting by Firm Size. Different Categories



(a) Product Introduction Rate (b) Quality-Adj. Product Introduction Rate

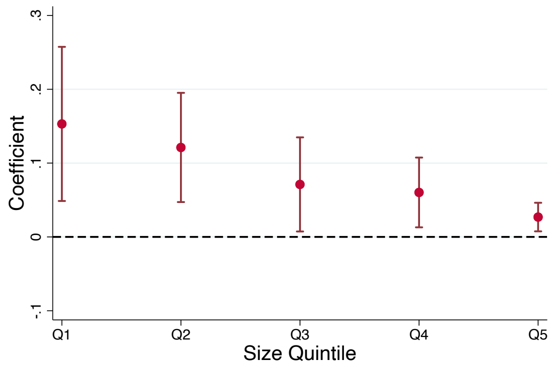
Notes: The figure plots the β_k coefficients from estimating the regressions $Y_{ijt} = \sum_{k=1}^{5} \beta_k \, P_{ijt-1} \times Q_{ijk} + \alpha_{ij} + \gamma_{jt} + u_{ijt}$, where Y_{ijt} is product introduction rate (Panel (a)) and quality-adjusted product introduction rate (Panel (b)) of firm i in product category j in year t; P_{ijt-1} is the patenting rate of firm i in product category j in year $t-1$; Q_{ijk} are dummies equal to one if the firm i's average sales in product category j are in the k^{th} quintile of firm sales distribution in j. The regression includes firm $\times$ category and category $\times$ year fixed effects. Standard errors are clustered at firm, category, and year levels. We aggregate product categories into food (dry grocery, frozen foods, dairy, deli, packaged meat, fresh produce and alcoholic beverages) and non-food (health and beauty, non-food grocery, and general merchandise. Standard errors are clustered at the firm $\times$ category level.

Figure B5: Share of Product-Related Patents by Firm Size Percentile

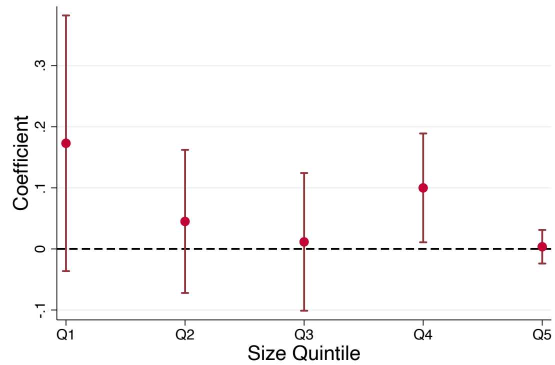


Notes: This figure plots the shares of product-related patents held by firms across deciles of firm size, defined in terms of sales. We classify patents into product-related patents based on the claims of patent documents (Appendix A.2).

Figure B6: Product Introduction and Patenting by Size: Process and Product Patents



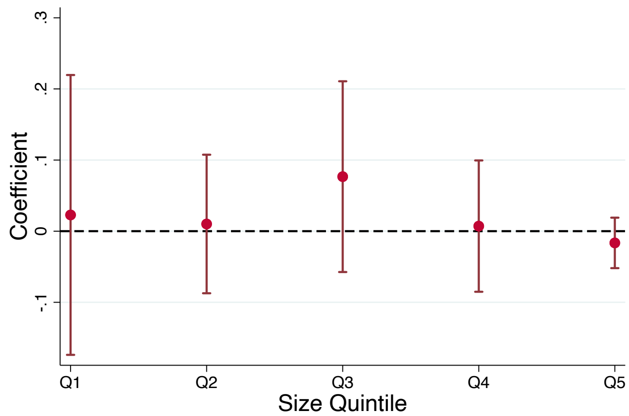
(a) Product Patents



(b) Process Patents

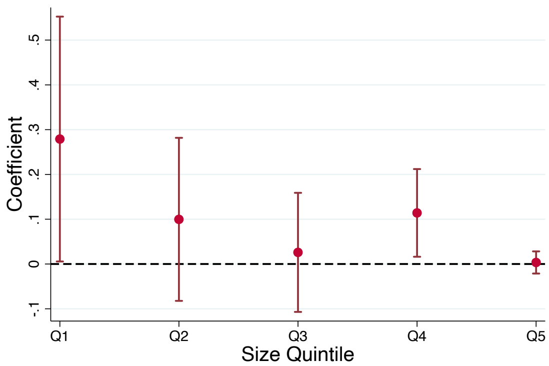
Notes: The figures show results similar to Figure 7, for product introduction rates, splitting patents into process and product patents based on classification in Bena and Simintzi (2017). The left panel plots regression coefficients including only product-related patents. The right panel plots regression coefficients, including only (matched) process-related patents. Standard errors are clustered at the firm × category level.

Figure B7: Prices and Process Patents

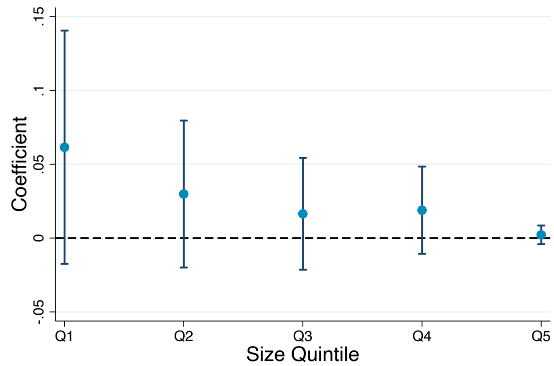


Notes: This figure plots the coefficients from a regression of prices on the interaction between rates of process patenting and quintiles of firm size, based on average sales. The figure plots the β_k coefficients from estimating the regressions $Y_{ijt} = \sum_{k=1}^{5} \beta_k \, P_{ijt-1} \times Q_{ijk} + \alpha_{ij} + \gamma_{jt} + u_{ijt}$, where Y_{ijt} are prices and P_{ijt-1} is the process patenting rate of firm i in product category j in year $t-1$; Q_{ijk} are dummies equal to one if the firm i's average sales in product category j are in the k^{th} quintile of firm sales distribution in j. The regression includes firm $\times$ category and category $\times$ year fixed effects. Standard errors are clustered at the firm $\times$ category level.

Figure B8: Product Introduction and Patenting by Size: CPG-Only Firms



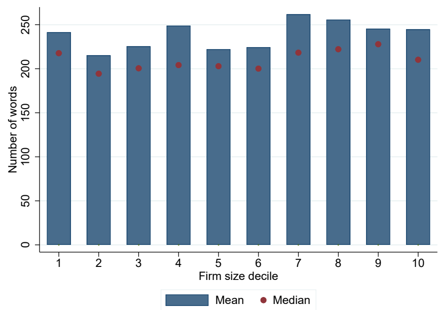
(a) Product Introduction Rate



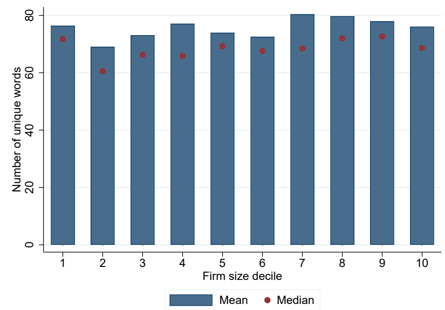
(b) Quality-Adj. Product Introduction Rate

Notes: The figure plots the β_k coefficients from estimating the regressions $Y_{ijt} = \sum_{k=1}^{5} \beta_k \, P_{ijt-1} \times Q_{ijk} + \alpha_{ij} + \gamma_{jt} + u_{ijt}$, where Y_{ijt} is product introduction rate (Panel (a)) and quality-adjusted product introduction rate (Panel (b)) of firm i in product category j in year t; P_{ijt-1} is the patenting rate of firm i in product category j in year $t-1$; Q_{ijk} are dummies equal to one if the firm i's average sales in product category j are in the k^{th} quintile of firm sales distribution in j. "CPG-only" firms are defined using Compustat and NETS, as described in Section A.1. The regression includes firm $\times$ category and category $\times$ year fixed effects. Standard errors are clustered at the firm $\times$ category level.

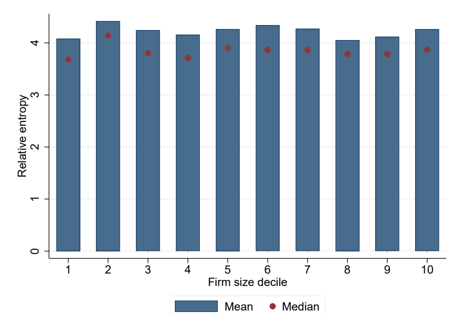
Figure B9: Patent Text and Match Properties by Firm Size Percentile



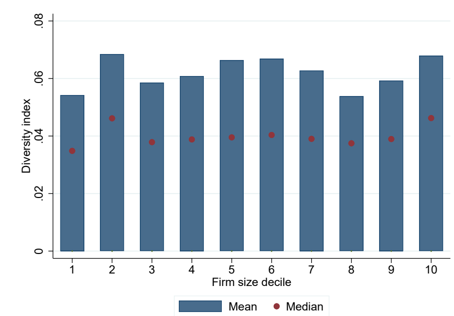
(a) Number of words in patent documents



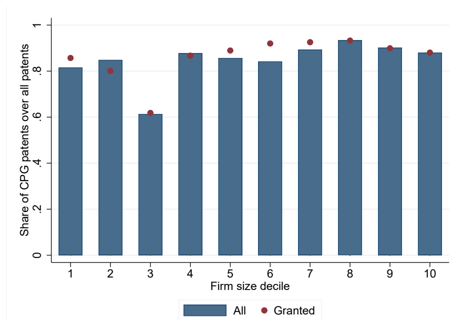
(b) Unique words in patent documents



(c) Relative entropy of patent word dist.



(d) Word diversity index of patent word dist.



(e) Share of matched patents

Notes: The figure plots various text and match characteristics of patents held by firms in different size (sales) deciles. All size deciles are constructed within product categories, except for panel (e) that is based on firm-level deciles. The first panels plot means and medians of the average number of words (a), the number of unique words (b), the relative entropy between the patent's word distribution and the word distribution of all patents (c), and the Simpson's diversity index of the patent's word distribution (d) of firms' patents. Panel (e) looks at the share of matched patents in the firms' whole patent portfolio (that is, the number of patents from patents-to-products dataset divided by the number of patents from firm-level match) for the sample of CPG-only firms (see Section A.1 for the definition) for which the non-matches are less likely to be due to the firm's operations outside the CPG sector. Panel (f) plots the similarity scores of the matched patents.

B.1 Results with Alternative Matching

In this section, we present our main results under different specifications of our matching algorithm. In our baseline specification, we classify product modules into 400 clusters, which we refer to as product categories. We do so since we believe this partition strikes a balance between aggregating very similar products while maximizing the difference between products across categories. Nonetheless, in this section, we show that our main findings are similar if we use the product classification scheme developed by Nielsen: 1,070 detailed product modules aggregated into a set of 114 broad product groups.

We also present our main results using a higher matching similarity threshold. Recall that patents with low text similarity are deemed unrelated to the product categories that we consider. In our baseline specification, we restrict the set of potential categories for each patent to the product categories whose similarity score exceeds 0.025. This section shows that our main results hold when using a higher similarity score of 0.05. We construct two versions of this exercise: version 1 assigns a patent to the highest-similarity product category, conditional on the similarity being above 0.05 and the rank being below 5; version 2 assigns a patent to the highest-similarity product category, conditional on the similarity being above 0.05. As before, if these conditions are not satisfied for any product category, then the patent is classified as a "non-match."

Table B8: Product Introduction and Patenting Rates (using the NielsenIQ product group aggregation)

	Product Introduction			Product Introduction Quality-Adjusted		
	(1)	(2)	(3)	(4)	(5)	(6)
Patents(t-1)	0.0407***			0.0187***		
	(0.008)			(0.003)		
Patents granted(t-1)		0.0527***			0.0223***	
		(0.010)			(0.004)	
Patents citations adj.(t-1)			0.0593***			0.0266***
			(0.014)			(0.006)
Observations	309,718	309,718	308,630	309,718	309,718	308,630
R-squared	0.363	0.363	0.364	0.303	0.303	0.303
Time-Category	Y	Y	Y	Y	Y	Y
Firm-Category	Y	Y	Y	Y	Y	Y

Notes: Notes: The table shows regressions of the product introduction rates and quality-adjusted product introduction rates as a function of the patenting rates, using firm × category × year data. Product and quality-adjusted product introduction rates are defined as the number of new products or quality-adjusted new products over a total number of products in firm × category × year. Product quality measures are defined in Section 2.3.1. *Patents* is the ratio of the firm's number of patent applications in a particular category-year over the total number of cumulative patents in that category-year; *Patents granted* is the ratio of the firm's number of granted patent applications in a particular category-year over the total number of cumulative patents in that category-year; *Patents citation-adjusted* is the ratio of the firm's number of citations-weighted granted patents in a particular category-year over the total number of citation-weighted granted patents in that category-year. Observations at the firm × category × year level with zero patents are included in the regression. Standard errors are clustered at firm × category level. We use an aggregation of modules into product groups as defined by Nielsen.

Table B9: Product Introduction and Patenting: by Size (using NielsenIQ product group aggregation)

	(1)	(2)	(3)	(4)
	Product Introduction		Product Introduction Quality-Adjusted	
Patents(t-1)	0.0407***	0.0920***	0.0187***	0.0507***
	(0.008)	(0.024)	(0.003)	(0.010)
Size(t)		0.0102***		0.0013***
		(0.000)		(0.000)
Patents(t-1) x Size(t)		-0.0046***		-0.0028***
		(0.002)		(0.001)
Observations	309,718	309,718	309,718	309,718
R-squared	0.363	0.367	0.303	0.303
Time-Category	Y	Y	Y	Y
Firm-Category	Y	Y	Y	Y

Notes: The table shows regressions of the product introduction rates and quality-adjusted product introduction rates as a function of the patenting rates, using firm × category × year data, similar to Table 4 but introducing size (firm sales in category-year) and size interaction with patenting. Product and quality-adjusted product introduction rates are defined as the number of new products or quality-adjusted new products over a total number of products in firm × category × year. Product quality measures are defined in Section 2.3.1. *Patents* is the ratio of the firm's number of patent applications in a particular category-year over the total number of cumulative patents in that category-year. Observations at the firm × category × year level with zero patents are included in the regression. Standard errors are clustered at the firm × category level. We use an aggregation of modules into product groups as defined by Nielsen.

Table B10: Product Introduction and Patenting: by Size (Higher Similarity Threshold)

	Threshold 0.05, version1				Threshold 0.05, version 2			
	Product Intro.		Product Intro. Quality-adjusted		Product Intro.		Product Intro. Quality-adjusted	
Patents(t-1)	0.0384***	0.0766***	0.0143***	0.0421***	0.0420***	0.1008***	0.0147***	0.0461***
	(0.008)	(0.023)	(0.003)	(0.009)	(0.008)	(0.026)	(0.003)	(0.009)
Size(t)		0.0102***		0.0012***		0.0103***		0.0012***
		(0.000)		(0.000)		(0.000)		(0.000)
Patents(t-1)		-0.0033**		-0.0024***		-0.0049***		-0.0026***
× Size(t)		(0.002)		(0.001)		(0.002)		(0.001)
Observations	409,434	409,434	409,434	409,434	409,131	409,131	409,131	409,131
R-squared	0.357	0.362	0.302	0.303	0.357	0.362	0.301	0.302
Time-Category	Y	Y	Y	Y	Y	Y	Y	Y
Firm	N	N	N	N	N	N	N	N
Firm-Category	Y	Y	Y	Y	Y	Y	Y	Y

Notes: The table shows regressions of the product introduction rates and quality-adjusted product introduction rates as a function of the patenting rates, using firm × category × year data, but introducing size (firm sales in category-year) and size interaction with patenting for higher similarity thresholds. To find the match, Version 1 keeps only categories with similarity above 0.05 and rank below 5; Version 2 keeps all categories with similarity above 0.05 regardless of rank. Product and quality-adjusted product introduction rates are defined as the number of new products or quality-adjusted new products over the total number of products in the firm × category × year. Product quality measures are defined in Section 2.3.1. *Patents* is the ratio of the firm's number of patent applications in a particular category-year over the total number of cumulative patents in that category-year. Observations at the firm × category × year level with zero patents are included in the regression. Standard errors are clustered at the firm × category level.

C Theoretical Appendix

C.1 Microfounding Profit Function

Consider a partial equilibrium framework that depicts the innovation process in a single product category. There are J potential producers, and aggregate output is produced using a combination of their quality-weighted varieties:

$$Y = \frac{1}{1-\beta}\left[\sum_{j=1}^{J} q_j^{\frac{\alpha}{1-\beta}} y_j\right]^{1-\beta}, \tag{13}$$

where y_j denotes the quantity, and q_j is the quality level of variety j. This specification implies that products from different producers are perfect substitutes after adjusting for their qualities. The parameter α captures the consumer's satiation with respect to additional quality. Labor is the only factor of production. Producers use labor to make intermediates by hiring workers at the common wage w. Output of variety j is then given by $y_j = l_j$, where l_j is the amount of labor used to produce variety j.[59] We assume that an overhead cost of production ϵ must be paid before choosing prices and output. Since producers' marginal costs are the same and qualities are different, under Bertrand competition, even a small overhead cost allows the highest-quality firm to act as the sole producer.[60]

The incumbent producer maximizes profits by choosing the price of its product subject to demand curve: $p = q^{\alpha} y^{-\beta}$ that follows from (13) (normalizing price of Y to one). We obtain the following equilibrium values for output (y), sales (R), and profits (Π):

$$y = \frac{1-\beta}{\beta}\frac{\pi}{w} q^{\gamma}, \qquad R = \frac{\pi}{\beta} q^{\gamma}, \qquad \Pi = \pi q^{\gamma}, \tag{14}$$

where $\pi \equiv \beta\left(\frac{1-\beta}{w}\right)^{\frac{1-\beta}{\beta}}$ and $\gamma \equiv \frac{\alpha}{\beta}$. Hence, an incumbent with a higher-quality product is larger and earns higher sales and profits. We assume $0 < \alpha < \beta$ (i.e., $0 < \gamma < 1$), implying that marginal quantity, sales, and profits decline with quality–a case supported by our calibration, which estimates $\gamma < 1$.[61]

[59] The model abstracts from production efficiency and cost-reducing process innovations, focusing instead on product innovations to align with our empirical analysis.

[60] This assumption, standard in this class of models, simplifies the setup. Alternatively, we could work with limit pricing, where the highest-quality firm serves the whole market but sets its price based on the second-highest-quality producer.

[61] Alternatively, a declining size-innovation relationship can arise from weaker scalability of R&D with firm size (Akcigit and Kerr, 2018) or from an innovation-advertising trade-off (Cavenaile and Roldan-Blanco, 2021).

C.2 Deriving Rates of Creative Destruction

Depending on the actions of the incumbent firm, our model delivers the following rates of creative destruction.

- If the incumbent firm neither patents the idea nor introduces a new product, creative destruction happens at a rate

$$p \times \Pr\left(q + \lambda^e > q\right) = p.$$

Hence, any product of higher quality introduced by an entrant will capture the full market.

- If the incumbent firm does not patent but successfully commercializes the product, creative destruction happens at a rate

$$p \times \Pr\left(q + \lambda + \lambda^e > q + \lambda\right) = p.$$

Again, any product of higher quality introduced by an entrant will capture the market.

- If the incumbent firm patents but does not introduce new products:

$$p \times \Pr\left(q + \lambda + \lambda^e > q + \lambda + \varepsilon\right) = p(1 - \varepsilon).$$

Although higher quality products by entrants can still win the market, now entrants' innovation needs to be sufficiently large to also withstand the legal protection from the incumbent's patent.

- Similarly, if the incumbent firm patents and also introduces new products:

$$p \times \Pr\left(q + \lambda + \lambda^e > q + \lambda + \varepsilon\right) = p(1 - \varepsilon).$$

C.3 Model Quantification and Counterfactuals: Details

The model features seven structural parameters: $r, p, \varepsilon, \lambda, \gamma, \tilde{c}_m$, and $\tilde{c}_p$, where $\tilde{c}_m$ and $\tilde{c}_p$ denote the respective cost parameters normalized by π. We set the interest rate to $r = 0.04$ and calibrate the parameters governing creative destruction, the entrant arrival rate (p) and the patent protection step (ε), using data on firm sales growth by patenting and product introduction status. In the model, firms that neither innovate nor patent face an expected sales decline from creative destruction equal to $\log(1-p)$. Using our baseline firm $\times$ category $\times$ year data for 2007-2015, we compute the median revenue growth for firm-category pairs without patents or new products, which implies $p = 0.095$. This decline is mitigated if a firm holds a patent. Conditional on not innovating, the impact of patenting on growth is $\log \frac{1-p(1-\varepsilon)}{1-p}$. To quantify this, we regress revenue growth on log patents with firm-category and category-year fixed effects, restricting to observations without product introduction. Multiplying the estimated coefficient by 0.88, corresponding to the shift from zero to one patent under the inverse hyperbolic sine transformation, yields the implied value of $\varepsilon = 0.23$.

Table C1: Model Parameters

Parameter		Identification	Value
r	Interest rate	External calibration	0.04
p	Arrival rate of entrants	Direct data	0.095
ε	Patent protection	Direct data	0.23
λ	Innovation step	Internal calibration	0.0751
γ	Elasticity of revenue to quality	Internal calibration	0.89
$\tilde{c}_m$	Cost of commercialization	Internal calibration	2.74
$\tilde{c}_p$	Cost of patenting	Internal calibration	42.13

Notes: Table presents all calibrated parameters, and the procedure to parameterize its value.

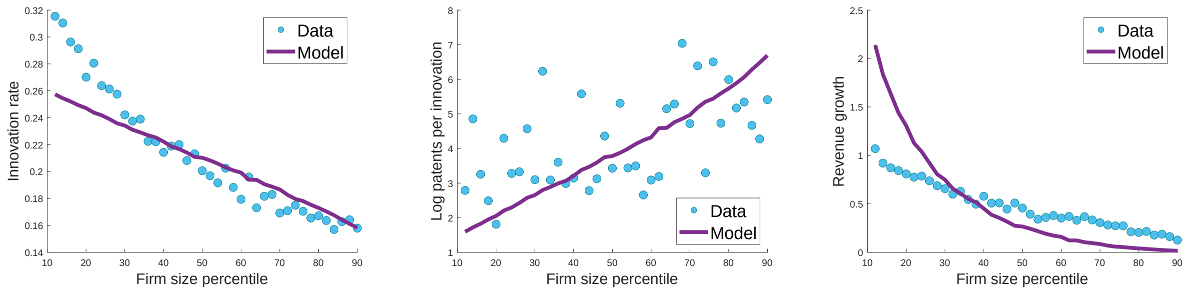
For the remaining four parameters, we calibrate the model to match the innovation rate, patents per innovation, and sales growth of firms between the 10th and 90th size percentiles.[62] Intuitively, $\tilde{c}_m$ and $\tilde{c}_p$ determine the levels of innovation and patenting, λ governs average growth conditional on innovation, and the curvature parameter γ shapes how this growth varies with firm size.

For each firm $\times$ product category, we compute: (i) the innovation rate (new products relative to existing ones), (ii) patents per innovation (log patent applications over new products), and (iii) revenue growth $(2(y_t - y_{t-1})/(y_t + y_{t-1}))$, comparing innovators to non-innovators. We then aggregate these measures to the average within size bins.

To map data percentiles to quality levels in the model, we first normalize the quality q of the average firm in each product category to one. Then, using profit equation $\Pi = \pi q^\gamma$, we obtain $q = (\frac{Rev}{\overline{Rev}})^{1/\gamma}$, where we measure Rev as the firm's revenue per products in a product category, and $\overline{Rev}$ denotes the average revenue in the product category. This gives us a mapping between the average normalized revenue in various percentiles to their corresponding levels of q in the model.

[62]Because of the model's parsimony, it has difficulty closely fitting the extremes of the size distribution, especially growth rates, so we estimate parameters outside this range.

Figure C1: Model vs Data



(a) Innovation (b) Patents per Innovation (c) Revenue Growth

We minimize the distance between the model's innovation rate $z_m(q)$, the log patent-to-innovation ratio $\log \frac{z_p(q_i)}{z_m(q_i)}$, and log revenue growth $\Delta \log R$ and their corresponding data moments. The objective function is

$$\min_{\lambda,\gamma,\tilde{c}_m,\tilde{c}_p} \sum_{i=10}^{90} \left[5(z_m(q_i) - m_i^1)^2 + (\log \frac{z_p(q_i)}{z_m(q_i)} - m_i^2)^2 + (\Delta \log R(q_i) - m_i^3)^2\right]^{1/2},$$

where m^1 denotes the product introduction rate, m^2 the log number of patent applications per new product, and m^3 revenue growth. We place a higher weight on the innovation moment, the cleanest measure and less prone to noise from the patent-matching procedure.

Table C1 reports the calibrated parameters, and Figure C1 illustrates the model fit. Despite its stylized nature and few parameters, the model matches firm innovation rates, patenting intensity, and growth across the size distribution quite well.

C.4 Size-Dependent Patenting Cost

To highlight firms' incentives for product innovation and patenting, and to clarify how motives for strategic patenting arise, especially among larger firms, the main model abstracted from size heterogeneity in patenting costs by assuming that c_p was independent of firm size ($\partial c_p / \partial q = 0$). In reality, large firms may patent more frequently also because their effective patenting costs are lower. Larger firms tend to be more experienced with filings, have in-house legal teams, and possess greater resources for litigation.[63] Such technological differences would naturally yield more patents with weaker links to product introduction for larger firms.

We next illustrate, through counterfactuals, that even if patenting costs decline with firm size, incentives for strategic patenting remain essentially unchanged. Lower costs provide large firms with additional technological reasons to patent more than small firms, but the strategic motive from the main model is unaffected.

To show this, we compare two economies with size-dependent patent costs to the baseline economy with uniform costs. Because the literature offers no reliable estimates of size-dependent patenting costs, we consider two illustrative cases: patent costs for large firms are set at one-half and one-fifth of those for small firms (and relative to the uniform baseline). Table C2 reports the results. The left panel repeats the uniform-cost economy from Table 9, while the middle and right panels show the size-dependent economies.

As expected, lower costs increase large firms' patenting, both relative to the baseline and to smaller firms. However, when we repeat our first counterfactual exercise—comparing benchmark outcomes to those in an economy without strategic patents while holding innovation fixed—the share of "excess" strategic patents is virtually identical: 84% in the baseline versus 84% and 83% in the size-dependent cases. Thus, even when large firms face lower patenting costs, their reliance on strategic patenting persists.

Table C2: Lower Patenting Cost for Large Firms. Comparison

	$c_p^{Large} = c_p$		$c_p^{Large} = c_p/2$		$c_p^{Large} = c_p/5$	
	Benchmark uniform	No strategic benchm. z_m	Benchmark size-dependent	No strategic benchm. z_m	Benchmark size-dependent	No strategic benchm. z_m
Innovation (z_m)	0.1550	0.1550	0.1597	0.1597	0.1737	0.1737
Creative destr. (τ)	0.0915	0.0944	0.0879	0.0939	0.0774	0.0919
Patenting (z_p)	0.1614	0.0252	0.3228	0.0520	0.8071	0.1416

Notes: The table compares three economies. The first economy (left panel) repeats the uniform-cost economy from Table 9, while the middle and right panels show the size-dependent economies with patent costs for large firms set at one-half and one-fifth of those for small firms (and relative to the uniform baseline). Within each panel, we report the benchmark and the counterfactual economy, where the latter shuts down strategic patenting while holding the product innovation rate fixed at the benchmark.

[63] Graham et al. (2009) report that application and enforcement costs are the most common reasons for forgoing patenting. Lerner (1995) documents that litigation costs in particular deter smaller firms from patenting.

Appendix References

Aizawa, Akiko, "An information theoretic perspective of TF-IDF measures," *Information Processing and Management*, 01 2003, *39*, 45–65.

Akcigit, Ufuk, Murat Celik, and Jeremy Greenwood, "Buy, Keep, or Sell: Economic Growth and the Market for Ideas," *Econometrica*, 2016, *84*, 943–984.

Argente, David, Munseob Lee, and Sara Moreira, "Innovation and Product Reallocation in the Great Recession," *Journal of Monetary Economics*, 2018, *93*, 1–20.

Bena, Jan and Elena Simintzi, "Globalization of work and innovation: Evidence from doing business in china," Technical Report, Discussion paper 2017.

Broda, Christian and David E Weinstein, "Product Creation and Destruction: Evidence and Price Implications," *American Economic Review*, 2010, *100* (3), 691–723.

Campello, Ricardo JGB, Davoud Moulavi, and Jörg Sander, "Density-based clustering based on hierarchical density estimates," in "Pacific-Asia conference on knowledge discovery and data mining" Springer 2013, pp. 160–172.

de Rassenfosse, Gaétan, "Notice Failure Revisited: Evidence on the Use of Virtual Patent Marking," NBER Working Paper 2018.

Feenstra, Robert C, "New Product Varieties and the Measurement of International Prices," *The American Economic Review*, 1994, pp. 157–177.

Fellbaum, Christiane, "Princeton university: About WordNet," 2010.

Goolsbee, Austan D. and Peter J. Klenow, "Internet Rising, Prices Falling: Measuring Inflation in a World of E-Commerce," *AEA Papers and Proceedings*, May 2018, *108*, 488–92.

Hall, Bronwyn, Adam Jaffe, and Manuel Trajtenberg, "The NBER Patent Citations Data File: Lessons, Insights and Methodological Tools," NBER Working Paper 8498 2001.

Hoberg, Gerard and Gordon Phillips, "Text-based Network Industries and Endogenous Product Differentiation," *Journal of Political Economy*, 2016, *124* (5), 1423–1465.

Hortaçsu, Ali and Chad Syverson, "The Ongoing Evolution of US Retail: A Format Tug-of-war," *Journal of Economic Perspectives*, 2015, *29* (4), 89–112.

Hottman, Colin J, Stephen J Redding, and David E Weinstein, "Quantifying the Sources of Firm Heterogeneity," *The Quarterly Journal of Economics*, 2016, *131* (3), 1291–1364.

Jaravel, Xavier, "The Unequal Gains from Product Innovations: Evidence from the U.S. Retail Sector," *The Quarterly Journal of Economics*, 2019, *134* (2), 715–783.

Lai, Ronald, Alexander D'Amour, David M. Doolin, Guan-Cheng Li, Ye Sun, Vetle Torvik, Amy Yu, and Lee Fleming, "Disambiguation and Co-authorship Networks of the U.S. Patent Inventor Database," *Research Policy*, 2014, *43* (6), 941–955.

Lloyd, Stuart, "Least Squares Quantization in PCM," *IEEE transactions on information theory*, 1982, *28* (2), 129–137.