Firms in Product Space

Luca Macedoni[a], John Morrow[b], Vladimir Tyazhelnikov[c]

August 2025

Abstract

Which products are efficiently produced together and *which* firms supply *which* products? Modeling multi-product firms under variable markups, we estimate firms' absolute advantages for produced products and develop an algorithm to predict them for *unproduced* products. Better advantages imply increased product adoption and explain which firms supply products when export demand induces domestic adoption. Predicted advantages and markups imply measures of Revenue and Competition Potential which explain firm sales and scope growth. If all firms produced all potential products, consumer welfare could increase by 16-30% under constant markups, rising to 46-86% under variable markups.

JEL Codes: F1, F6, D2, L1.
Keywords: Multi-product firms, firm capabilities, absolute advantage, classification, variable markups.

Acknowledgments. We thank Nick de Roos for early suggestions and Axel Roepstorff and Kristian Maretty Højen for excellent research assistance. We are grateful for comments from Kirill Borusyak, Magnus Tolum Buus, Wanyu Chung, Andreas Hafler, Luca Lodi, Guzman Ourens, Catherine Thomas, and participants at the 1st SICKCL, Economics of Global Interactions, KCL Seminars, ETSG, LSE Trade Workshop, Kent Macro Workshop, Australasian Trade Workshop, WEAI, Nottingham, Zurich ETH, Paris Trade & Macro Workshop, Asia-Pacific Trade Seminar, PSE/Sciences Po, InsTED, Lille, Organisational Economics Workshop, UC Louvain, Tilburg GIFT Conference, BCAM Macro, CEP Retreat, ESCOE, ITSG, Workshop on the Impact of Global Trade Shocks, Firms/Workers/Inst Linz, RES, SMU, INSE at Peking University, Universitat Pompeu Fabra, HKS Growth Lab, NYU Trade Day, CUHK Trade Workshop, Geneva Grad Institute, EUI Trade and Development, QMUL, ULB, Maastricht, University of Verona, the Competition and Markets Authority, SPRU, University of Pavia, University of Milan, UNSW, Aston Business School and University of Zurich.
[a]University of Milan. Financial support from the Carlsberg Foundation is gratefully acknowledged.
[b]Queen Mary University of London, CEP, and CEPR. Corresponding Author: john.morrow@qmul.ac.uk.
[c]The University of Sydney.

1 Introduction

Which products are potentially produced together? When demand for a product increases, can we predict which firms will supply it? This is important for policy as knowing *which* products a firm is likely to adopt can help target interventions when key supplies are desired, such as "green" products or those important during an emergency. A key obstacle is that until a firm produces a product, costs and benefits of this decision are unknown, at least outside of the firm. We introduce a novel approach to predict a firm's absolute advantages in *unproduced* products based on observed co-production patterns within and across firms. We represent these advantages as the distance from firms to products in a multi-dimensional *product space*. This provides a parsimonious and flexible way to quantify underlying co-production capabilities across several co-occurring activities often classified in separate industries.

By constructing a product space, we can enhance standard models to identify which products are more likely to be introduced and quantify counterfactual supply responses to a wide range of policy interventions and shocks.[1] Existing models of product supply under monopolistic competition typically treat a firm's absolute advantages across products as random draws from a probability distribution. This assumption sidesteps the complex reasons why firms engage in certain combinations of activities such as technological complementarities or organizational capabilities. As a result, the interconnections across activities and products are left as a black box. Our approach opens that black box by uncovering a latent map of counterfactual advantages over unobserved activities. This map helps explain firms' potential for market expansion and growth, as well as the resulting gains in consumer surplus.

To theoretically ground our approach, we develop a model of multi-product firms under monopolistic competition with CoPaTh demand (Matsuyama and Ushchev, 2017), which allows markups to vary across firms and products and nests the constant markup, CES demand system. Product adoption depends on observable factors such as market size and firm-specific capabilities, but crucially also on a firm's absolute advantage in producing a given product. For products the firm already produces, this advantage can be inferred from observed outcomes. But for unproduced products, the very ones for which product adoption is relevant, this information is missing.

Our central idea is that the sales distribution of the current set of products a firm makes and the distribution of co-occurring absolute advantages across all firms, can inform the absolute advantage a firm has on unproduced products. The main challenge we face is the potential high dimensionality of this problem. For example, trying to predict

[1] Exciting work suggestive of such results exists at the industry level and are important for both growth and trade, see Hausmann et al. (2022).

product adoption conditional on all the possible combinations of co-production imposes unrealistic data requirements. We address this challenge by constructing a product space from observed firm-product absolute advantages. In this space, the distance between a firm and an unproduced product captures the firm's counterfactual absolute advantage for that product. This solves the dimensionality problem in a parsimonious and theory-consistent way, allowing us to estimate adoption decisions even for unproduced products and to simulate counterfactual prices and revenues from potential product entries.

We employ rich Danish firm-level data to validate our approach and test two predictions about *which* firms produce *which* products. First, a firm's counterfactual absolute advantage predicts product adoption in a highly granular way, even controlling for discrete classification systems such as the Combined Nomenclature.[2] Second, our measure has substantially higher coverage of firm-product pairs contrary to measures of relatedness from the literature. Furthermore, counterfactual absolute advantages map directly to model primitives allowing for theoretically based econometrics and counterfactuals. Crucially, we validate the prediction of which firms will supply products by instrumenting product level demand with export demand, finding that a positive demand shock induces the firms with the stronger counterfactual absolute advantage to introduce a product.

Moreover, the product space reveals opportunities and threats. Firms with high absolute advantage for unproduced products, implying they are near in the product space, may have ample opportunity for expansion, while those far from potential products may have low growth potential. The trade and growth literature has demonstrated that countries with better market access through better trade channels grow faster (Redding and Venables, 2004); and while the international trade literature has ample measurements of distance to markets, there is no comparable concept for unentered *product* markets. We introduce a Revenue Potential (RP) index of how much a firm could increase its revenue by producing all accessible products in the product space, which predicts both sales and scope growth.

We apply similar logic to potential competition in each firm's existing product portfolio and calculate a Competition Potential (CP) index. While it might seem that a firm is positioned to expand rapidly into nearby products, it is possible that it is surrounded by a large number of potential entrants, restricting growth. The CP index calculates how a firm's revenue would fall if all potential rivals choose to compete with a firm in all its existing product markets. In fact, we find that high CP restricts sales and scope growth.

Finally, counterfactual absolute advantages and markups allow measurement of potential gains from variety for consumers by predicting the impact of firm entry into

[2] These can be both haphazard and endogenous to policy objectives. See the examples of Jacobs and ONeill (2003) and Grant (2023). Table A.1 highlights differences across the HS, SIC and NAICS.

unproduced varieties on the price index. This approach is similar to that proposed by Feenstra (1994), however here counterfactual prices are generated from our novel method which capture potential (in)efficiencies in co-production. We define Potential Gains from Variety (PGV) to capture such gains. The predicted gains vary by sector, ranging from 16-30% under constant markups and rising to 46-86% under variable markups.

This section continues with a literature review. Section 2 presents the theoretical model. Section 3 introduces a method to recover latent absolute advantages. Section 4 describes the data and estimation strategy. Section 5 empirically tests the model's predictive power for product adoption. Section 6 defines three indices to quantify the implications of product space location for firm growth and consumer welfare. Section 7 concludes.

Literature Review

There has been a rapid expansion of research on multi-product firms, especially in the context of international trade.[3] For the typical model of this literature, a firm is a collection of product varieties, which may be linked by supply or demand linkages. Supply linkages arise from flexible manufacturing, economies and dis-economies of scope, and the presence of core competency and non-core competency products (Eckel and Neary, 2010; Nocke and Yeaple, 2014; Mayer et al., 2014; Eckel et al., 2015; Arkolakis et al., 2021; Macedoni and Xu, 2022). While the literature has established there are "significant departures from the theoretical benchmark of core competencies" (Fontagne et al., 2018) and that such an approach "can potentially lead to a rich set of insights concerning the buyer–seller network of firms" (Herkenhoff et al., 2024), progress in formalizing and quantifying these departures has remained limited. On the demand side, linkages mainly include cannibalization effects and demand complementarities (Feenstra and Ma, 2007; Eckel and Neary, 2010; Dhingra, 2013; Bernard et al., 2018; Flach and Irlacher, 2018; Macedoni, 2022). However, how particular product bundles a firm produces matters for the supply side and growth is understudied. A contribution of our approach is to show that considering *which* product bundles a firm produces relative to *all other firms* matters for product adoption and growth.

The literature on co-production (Bernard et al., 2010; Goldberg et al., 2010) shows that some pairs of products are often produced together, while others are almost never produced by the same firm. In this approach, when a firm expands its product range, it will likely choose products that are often co-produced within other firms, conditional on their current product mix, for a wide variety of possible explanations for linkages across inputs and outputs of firms (Boehm et al., 2022; Dhyne et al., 2022; Ding, 2023; Jakel

[3]For details, see the recent review by Irlacher (2022).

et al., 2023). We build on such co-production concepts and in our approach, a firm is more likely to expand into products that are closer to its existing capabilities. Here, closeness is determined not only by the frequency of co-production across products, but also by the efficiency with which firms co-produce them.

The concept of a product space was popularized by Hidalgo et al. (2007), Hausmann et al. (2007), and Hausmann and Hidalgo (2011), who define product distance based on the co-occurrence of Revealed Comparative Advantage (RCA) across countries.[4] A key insight from this literature is that proximity to products helps explain cross-country differences in income growth.[5] Instead of relying on RCA, we construct product distances from firm-level absolute advantages, yielding a product space micro-founded with monopolistic competition models and econometric estimation of costs and markups. This allows us to predict product adoption even when co-production is sparse. This also addresses the firm-product sparsity limitation of RCA-based approaches which results in the majority of combinations showing zero co-occurrence of RCA across firms.

Our approach to the product space deliberately departs from methods based on product characteristics, which underlie models of hedonic demand (Lancaster, 1966; Rosen, 1974; Berry et al., 1995; Feenstra and Levinsohn, 1995). For example, in Pellegrino (2025), firms are positioned in a characteristics space, where proximity reflects the degree of overlap in product attributes. In contrast, our framework can be interpreted as the *dual* of this characteristics-based approach: here, proximity between products reflects similarities on the supply side rather than the demand side. Specifically, two products are close if firms frequently exhibit co-occurring absolute advantages in producing them, i.e. if they are co-produced efficiently. This method enables us to analyze a broad set of otherwise incomparable products and requires minimal data: in the simplest case, firm-level sales data are sufficient.

Our focus on firms in a product space links our paper to an expanding body of research that uses various measures of similarity *between firms* for different applications. For example, in the R&D literature, spillovers across firms depend on the proximity between technologies and products across firms. A common method in this field involves calculating the overlap between firms' technology classes, as indicated by patents, and the overlap between their product sales (Jaffe, 1986; Bloom et al., 2013).[6]

[4]The approach has since evolved (Hidalgo, 2021), with applications ranging from research fields (Guevara et al., 2016) to pioneer firm survival (Jara-Figueroa et al., 2018) and industry transitions (Neffke et al., 2011), and more recently to green products (Mealy and Teytelboym, 2022).

[5]Hidalgo et al. (2007) show that countries tend to develop RCA in goods close to those they already specialize in, facilitating faster export upgrading. This mirrors our finding that firms tend to adopt products with higher counterfactual absolute advantage and that higher Revenue Potential accelerates growth.

[6]Escolar et al. (2023) provides a summary of the approaches used in the literature for locating firms using patents data.

A small but growing set of research has been creating new categorizations of firm activities and outputs, often using advances in text analysis to uncover new relationships. In an exciting strain of work, Hoberg and Phillips (2016) create firm locations from word vectors of SEC filings, in which firms have new relative locations each year based on cosine similarity measures which generally better explain profitability and growth than SIC or NAICS classifications. They show that, following a negative demand shock, firms either relocate toward areas of high common demand or differentiate by reducing product similarity, underscoring that the relative similarity of product portfolios is an important determinant of firm dynamics. These findings are consistent with our results based on the Revenue and Competition Potential measures. In contrast to their methodology, we generate a new production-based classification with standard product classification codes common to administrative data. Existing crosswalks, thereby, allow us to link to many potential data sets and policy changes as used in the trade literature.

We next lay out the model framework which we will use to motivate the empirical measures, estimating equations and counterfactual exercises in the rest of the paper.

2 Multi-Product Firms and Product Adoption

This section presents a model of multi-product firms engaged in monopolistic competition using the Constant Pass Through (CoPaTh) setting of Matsuyama and Ushchev (2020) which nests the well known CES demand system. The novel feature we leverage here is the presence of markups that vary within products across firms and within firms across varieties. However, markups are constant at the firm-variety level as they do not depend on other economic conditions such as the toughness of competition. Firms absolute advantages vary by product and drive product adoption, which additionally depends on variables that can be estimated with fixed effects. We derive product adoption estimating equations that motivate our construction of the product space to predict counterfactual absolute advantages on *unproduced* products.

2.1 Consumers and Firms

There is a discrete set of differentiated products indexed by ν. Each product ν is produced by a continuum of firms indexed by ω, each offering a horizontally differentiated variety. The pair (ω, ν) denotes firm ω's variety of product ν.

Consumers have nested preferences and aggregate income I. At the top level, consumption is aggregated across products ν with Cobb-Douglas preferences. Within each nest, preferences are of the CoPaTh class introduced by Matsuyama and Ushchev (2020).[7] Let $q_{\omega,\nu,t}$ and $p_{\omega,\nu,t}$ denote the quantity and price of variety (ω,ν) at time t. and $\boldsymbol{q}_{\nu,t}$ and $\boldsymbol{p}_{\nu,t}$ denote their respective vectors. The CoPaTh aggregator over consumption bundles is denoted $Q(\boldsymbol{q}_{\nu,t})$, wherein variety (ω,ν) has an elasticity of substitution $\sigma_{\omega,\nu}$.

The first layer of the utility function is

$$U\left(\{\boldsymbol{q}_{\nu,t}\}\right)=\sum_{\nu}\psi_{\nu}\ln Q\left(\boldsymbol{q}_{\nu,t}\right)\quad\text{with}\quad\sum\psi_{\nu}=1,$$

where ψ_{ν} are the constant, Cobb Douglas expenditure shares. The total revenues generated by product ν satisfy $R_{\nu}=\psi_{\nu}I$. Consumers choose $\{q_{\omega,\nu,t}\}$ to maximize utility subject to the budget constraint

$$\max_{q_{\omega,\nu,t}}U\left(\{\boldsymbol{q}_{\nu,t}\}\right)\quad\text{subject to}\quad I=\sum_{\nu}\int_{\omega\in\Omega_{\nu,t}}p_{\omega,\nu,t}q_{\omega,\nu,t}d\omega.$$

A key distinction between CoPaTh and CES preferences is the presence of two different price indices. For the price vector $\boldsymbol{p}_{\nu,t}$, the ideal price index $P(\boldsymbol{p}_{\nu,t})$ captures the cost of one unit of utility. The aggregator $A(\boldsymbol{p}_{\nu,t})$, in contrast, enters the expression for individual firm demand and determines expenditure shares. These two indices generally differ, and coincide only under CES preferences when $\sigma_{\omega,\nu}=\sigma_{\nu}$ for all ω.

Let $s_{\omega,\nu,t}$ denote the expenditure share on variety (ω,ν) at time t. The CoPaTh structure implies

$$s_{\omega,\nu,t}=\left(p_{\omega,\nu,t}/A\left(\boldsymbol{p_{\nu,t}}\right)\right)^{1-\sigma_{\omega,\nu}}.\tag{1}$$

The aggregator $A(\boldsymbol{p}_{\nu,t})$ is implicitly defined by the following equation

$$\int_{\Omega_{\nu}}\left(\frac{p_{\omega,\nu,t}}{A(\boldsymbol{p}_{\nu,t})}\right)^{1-\sigma_{\omega,\nu}}d\omega=1,\tag{2}$$

which is obtained by integrating over the shares in equation (1). Since the left hand side of Equation (2) is increasing in A, this equation admits a unique solution for $A\left(\boldsymbol{p_{\nu,t}}\right)$.

The ideal price index $P\left(\boldsymbol{p_{\nu,t}}\right)$ is given by

$$P\left(\boldsymbol{p_{\nu,t}}\right)=A\left(\boldsymbol{p_{\nu,t}}\right)\exp\left(\int_{\Omega}\frac{1}{1-\sigma_{\omega,\nu}}\left(\frac{p_{\omega,\nu,t}}{A\left(\boldsymbol{p_{\nu,t}}\right)}\right)^{1-\sigma_{\omega,\nu}}d\omega\right).\tag{3}$$

[7] Matsuyama and Ushchev (2020) provide a general framework for homothetic demand systems. A notable subclass is the Homothetic Single Aggregator (HSA) family, which exhibits constant price elasticities and therefore constant markups at the variety level which we use here. For details, see the original paper.

Under CES preferences with $\sigma_{\omega,\nu} = \sigma$, this expression reduces to

$$P\left(\boldsymbol{p}_{\nu,t}\right) = \left(\int_{\Omega_\nu} p_{\omega,\nu,t}^{1-\sigma} d\omega\right)^{1/(1-\sigma)} = A\left(\boldsymbol{p_{\nu,t}}\right). \tag{4}$$

Firm ω can produce product ν at a fixed cost $f_{\omega,\nu}$ and constant marginal cost $m_{\omega,\nu,t}$, which may vary over time.[8] Conditional on producing ν, firm ω sets the price

$$p_{\omega,\nu,t} = \frac{\sigma_{\omega,\nu}}{\sigma_{\omega,\nu} - 1} m_{\omega,\nu,t} = \mu_{\omega,\nu} m_{\omega,\nu,t}, \tag{5}$$

where $\mu_{\omega,\nu} \equiv \sigma_{\omega,\nu}/(\sigma_{\omega,\nu} - 1)$ are defined as markups and $\boldsymbol{\mu_\nu}$ the vector of markups for each product ν. Markups can vary across varieties within a firm and across firms, depending on $\sigma_{\omega,\nu}$.

Let $\overline{m}_{\nu,t}$ denote the average cost of actually produced varieties ν across firms at time t, and define the firm's Absolute Advantage Index as

$$\text{Absolute Advantage Index:} \qquad a_{\omega,\nu,t} \equiv \frac{m_{\omega,\nu,t}}{\overline{m}_{\nu,t}}. \tag{6}$$

This measure is invariant to the choice of units for each product ν. The index represents the cost for a firm to produce a variety as the percentage of the product average. The lower $a_{\omega,\nu,t}$, the higher the absolute advantage of firm ω in product ν. We refer to $a_{\omega,\nu,t}$ as absolute advantage for brevity, with $\boldsymbol{a_{\nu,t}}$ as the vector of advantages for product ν at time t.

Note from Equation (5) that $\boldsymbol{p}_{\nu,t}/\overline{m}_{\nu,t} = \boldsymbol{\mu}_\nu \odot \boldsymbol{a}_{\nu,t}$ where $\odot$ is component-wise multiplication. Then using the fact that A is homogeneous of degree one, the revenue and profit of firm ω from producing ν at time t are

$$r_{\omega,\nu,t} = \left(\frac{\mu_{\omega,\nu} a_{\omega,\nu,t}}{A(\boldsymbol{\mu}_\nu \odot \boldsymbol{a}_{\nu,t})}\right)^{1-\sigma_{\omega,\nu}} R_\nu, \qquad \pi_{\omega,\nu,t} = \frac{r_{\omega,\nu,t}}{\sigma_{\omega,\nu}}. \tag{7}$$

2.2 Product Adoption Equations

Letting $A_{\nu,t} \equiv A(\boldsymbol{\mu}_\nu \odot \boldsymbol{a}_{\nu,t})$ and using Equation (7), variable profits at time t are

$$\pi_{\omega,\nu,t} = \frac{R_\nu}{\sigma_{\omega,\nu}} \left(\frac{\mu_{\omega,\nu} a_{\omega,\nu} \varphi_{\omega,t}}{A_{\nu,t}}\right)^{1-\sigma_{\omega,\nu}},$$

[8] Matsuyama and Ushchev (2020) identify each variety with a firm, but we generalize to the multi-product case with minimal changes. We assume a fixed number of firms without modeling entry. In the empirical section, we allow the set of firms and costs to evolve as in the data.

where we have added an ex-post TFP shock $\varphi_{\omega,t}$ to estimate effects within the firm. The event that a firm adopts a product, $\text{Intro}_{\omega,\nu,t} = 1$, occurs when $\pi_{\omega,\nu,t} > f_{\omega,\nu}$. We take logs of both sides, which preservers the sign of the inequality. Rearranging, product adoption occurs when

$$-(\sigma_{\omega,\nu} - 1)\ln a_{\omega,\nu,t} - (\sigma_{\omega,\nu} - 1)\ln \varphi_{\omega,t}\mu_{\omega,\nu} + (\sigma_{\omega,\nu} - 1)\ln A_{\nu,t} > \ln \frac{\sigma_{\omega,\nu} f_{\omega,\nu}}{R_\nu}. \tag{8}$$

Equation (8) forms the basis for estimating product adoption probabilities using firm-level production data. Introducing an error term $\epsilon_{\omega,\nu,t}$ and rearranging, the adoption condition becomes:

$$-\ln a_{\omega,\nu,t} - \underbrace{\ln \varphi_{\omega,t}}_{\text{Firm-Time}} + \underbrace{\ln A_{\nu,t}}_{\text{Product-Time}} - \underbrace{\left(\frac{\ln(\sigma_{\omega,\nu} f_{\omega,\nu}/R_\nu)}{\sigma_{\omega,\nu} - 1} + \ln \mu_{\omega,\nu} \right)}_{\text{Firm-Product}} + \epsilon_{\omega,\nu,t} > 0. \tag{9}$$

Equation (9) can be estimated with a linear probability model and firm-time, product-time and firm-product fixed effects. Firm-time fixed effects capture time-varying shocks to firm-level productivity. Product-time fixed effects absorb changes in product-specific market conditions, here summarized by the aggregator A. Firm-product fixed effects control for components that are fixed within a variety over time, including the fixed cost $f_{\omega,\nu}$ and demand elasticity $\sigma_{\omega,\nu}$. The inclusion of firm-time fixed effects implies that identification relies on within-firm variation across products, and therefore requires a sample of multi-product firms.

Although fixed effects account for much of the unobserved heterogeneity in the data, a key limitation remains: the variable $a_{\omega,\nu,t}$, capturing absolute advantage, is only observable for products the firm actually produces. For potential products not currently produced, this information is missing. This creates a selection problem and introduces bias in estimating Equation (9), since the set of observed $a_{\omega,\nu,t}$ is endogenous to the firm's prior adoption decisions.

The next section introduces a novel method to recover counterfactual absolute advantages based on firm production patterns. The idea is to measure co-occurring absolute advantage between pairs of produced products as a measure of underlying co-production capabilities and extrapolate them to unproduced products. There are many possible metrics to achieve this, each corresponding to a different mapping from observed absolute advantages in co-production to counterfactual advantages. In all cases, we recover the underlying counterfactual absolute advantage to estimate Equation (9).

3 Co-occurring and Counterfactual Advantages

This section establishes a link between observed firms' behaviors and their absolute advantages for produced and unproduced products. Conditional on observing positive revenues, prices, and demand elasticities, we can use the model above to back out the set of absolute advantages for produced products. Our goal is to map these advantages to counterfactual advantages for unproduced products. This mapping could contain arbitrarily rich economies of scope from co-production. For instance, suppose that a firm's ability to produce a new product depends on the set of products it currently produces. That is, whether a firm produces product 1 and product 2, only 1, or only 2, or neither, affects its likelihood of adopting product 3. Since for M products there are 2^M possible combinations of products, estimating the effect of extensive margin combinations would impose unrealistic data requirements. We reduce the dimensionality of this problem by representing each of N products as locational vectors wherein distance between products represents the joint efficiency of co-producing them.

Before describing the main steps of the algorithm that constructs a product space, we define its key elements. The construction of the product space is done for every year t and we omit the time subscript to ease notation. The construction is applied to a given product *cluster*, indexed by c, which refers to a set of products linked by chains of co-production. These chains can be direct, as when the same firms co-produce two products, or indirect, as when two firms share one common product and their remaining products become indirectly connected through this shared production. We construct and analyze product spaces separately for each cluster.

For each cluster c, a Product Space, denoted PS_c, is a vector space with many dimensions as the number of products. Each product ν is represented by a vector ℓ_ν in this space. The distance between two products ν and τ, denoted $d_{\nu,\tau}$, is the Euclidean distance between ℓ_ν and ℓ_τ. Product vectors are located so that these distances represent the average absolute advantage to produce τ conditional on producing ν perfectly (i.e., with an absolute advantage at the zero-th percentile of costs). Smaller values of $d_{\nu,\tau}$ indicate co-occurrence of stronger absolute advantages when co-producing both goods.

Each firm ω is characterized by a vector ℓ_ω which locates the firm closer to products in which it has stronger absolute advantage. The distance from a firm to a product it does not make represents its counterfactual absolute advantage in that product, with products further away implying worse absolute advantage. Having located both the product and firm vectors in the product space, we then recover counterfactual absolute advantages for every firm-product pair in each cluster.

The rest of this section details the general problem we solve, how to construct distances and the product space, and how to locate firms which will imply counterfactual absolute advantages. In the Section 4.2, we estimate absolute advantages using two standard approaches from the literature corresponding to constant and variable markups as well as compare them to the dominant product classification hierarchy and product relatedness from the complexity literature.

3.1 Counterfactual Absolute Advantage

Our first step in estimating the product space is to construct distances $d_{v,\tau}$ between any two products ν and τ within the same cluster c. We interpret $d_{v,\tau}$ through the lens of efficiency in co-production: products that are closer in distance are more likely to be co-produced efficiently. In this sense, shorter distances reflect greater technological or organizational complementarities.

To illustrate this first step in our algorithm, consider a firm ω producing two products 1 and 2 with absolute advantages $a_{\omega,1}$ and $a_{\omega,2}$. In our product space, the distance of the firm to the two products equal these advantages and they can be visualized spatially by placing each product around the firm as in Figure 1a, where product 1 is on the circle of radius $a_{\omega,1}$ and product 2 is on the circle of radius $a_{\omega,2}$. One can imagine if the firm produced product 2 perfectly with $a_{\omega,2} = 0$, the distance from the firm to product 1 would be the distance between products 1 and 2, labeled $d_{1,2}$. An application of the triangle inequality in this example provides numerical bounds on this distance

$$|a_{\omega,1} - a_{\omega,2}| \leq d_{1,2} \leq a_{\omega,1} + a_{\omega,2}. \tag{10}$$

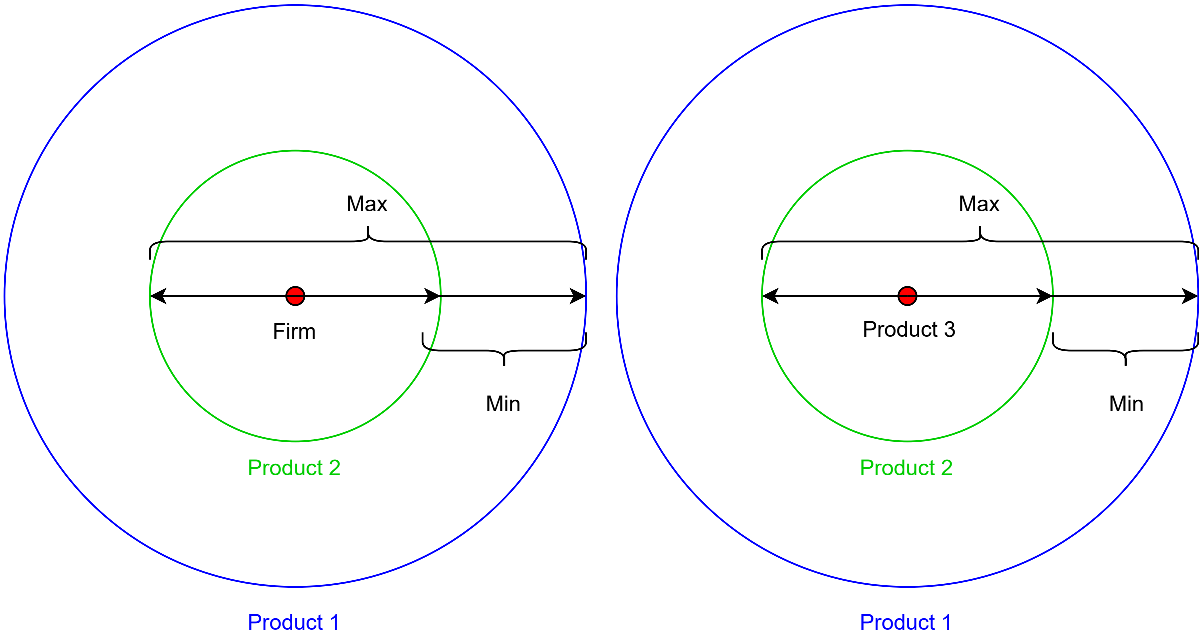
To illustrate this, consider the example in Figure 1a. Conditional on observing the firm's absolute advantages in products 1 and 2, we can apply the triangle inequality to recover bounds of the possible distance between the two products, $d_{1,2}$. The minimum distance arises when both products lie on the same line extending in one direction from the firm's location where both products located to the right. In this case, the distance between the products equals the absolute difference in their distances to the firm, i.e., $|a_{\omega,1} - a_{\omega,2}|$. Conversely, the maximum distance occurs when the two products lie on opposite directions through the firm with product 1 to the right and product 2 to the left, so that the distance between them equals the sum of the two radii, $a_{\omega,1} + a_{\omega,2}$.

For each firm ω which co-produces the product pair (ν, τ) we take the average[9] of the

[9]Given the bounds in Equation (10), one might be tempted to use order statistics to obtain a tighter fit. For example, instead of averaging, one could take the *minimum* of the upper bounds and the *maximum* of the lower bounds across observations to construct a narrower range of potential distances. In the absence of measurement error, this approach would indeed be preferable. However, in the presence of errors, the minimum of the upper bounds can be smaller than the maximum of the lower bounds,

Figure 1: Product Distance Bounds

(a) Products Co-Produced by Firms (b) Products Linked by Production Across Firms



upper and lower bounds of Equation (10) to define distance, where $N_{\nu,\tau}$ is the number of such firms

$$d_{\nu,\tau} \equiv \frac{1}{N_{\nu,\tau}} \sum_{\omega \in \Omega_{\nu,\tau}} \frac{a_{\omega,\nu} + a_{\omega,\tau} + |a_{\omega,\nu} - a_{\omega,\tau}|}{2}. \tag{11}$$

Equation (11) is an average between the of upper and lower distance bounds and therefore consistent with Equation (10) taken as an average. The distance between two products is then an index of co-occurring absolute advantages. When firms on average co-produce two products efficiently, hence exhibit co-occurring absolute advantages, the two products will be close in the space.

Although we posit a positive relationship between these distances and absolute advantages $a_{\omega,\nu}$, we do not impose a specific functional form *a priori*. Instead, we introduce a general increasing mapping between distances and absolute advantages $T(a_{\omega,\nu})$, with $T(0) = 0$, which we will employ below to better accommodate outliers in the data. This approach reflects the idea that even small differences in absolute advantage can, in some settings, translate into substantial variation in the probability or intensity of

especially as the number of observations grows. Empirically, we observe such a pattern a non-negligible fraction of our sample. For this reason, we adopt a more robust approach that averages the distance bounds across firms to mitigate the influence of errors and outliers.

co-production. The more general definition of distance is then

$$d_{\nu,\tau} \equiv \frac{1}{N_{\nu,\tau}} \sum_{\omega \in \Omega_{\nu,\tau}} \frac{T\left(a_{\omega,\nu}\right) + T\left(a_{\omega,\tau}\right) + \left|T\left(a_{\omega,\nu}\right) - T\left(a_{\omega,\tau}\right)\right|}{2}. \tag{12}$$

3.1.1 Co-production Capability for Products not Co-produced

Although the application of the triangle inequality allows us to find distances between any pair of co-produced products, our goal is to obtain distances between *all* products within a cluster, including products that are not co-produced by any firm. To get these distances, we use the known distance bounds between co-produced products and apply the triangle inequality successively. For example, if Firm A makes products 1 and 3 and Firm B makes products 2 and 3, we can infer bounds on the distance between products 1 and 2 by applying the triangle inequality to the distance bounds between pairs of products 1 and 3 and 2 and 3. We illustrate this idea in Figure 1b, mirroring Figure 1a: products 1 and 2 are co-produced with product 3 but not with each other, which allows us to construct distance bounds between products 1 and 2.

Notice that products may be linked by co-production through more than one product. For example, if Firm A produces products 1, 3, and 4, and Firm B produces products 2, 3, and 4, then products 1 and 2 are connected through products 3 and 4. Multiple linkages allow for multiple upper and lower distance bounds. Therefore, we find the least total distance across all of the distance upper bounds for any chain of indirectly co-produced products. Similarly, we find the greatest distance over all distance lower bounds. We then define the distance as the average of these least and greatest bounds.[10]

3.1.2 Assigning Product and Firm Locations

Given a complete set of pairwise distances between products in each cluster c, we represent these distances as a collection of vectors the same dimension as the number of products representing product locations, which we define as a product space PS_c. To resolve indeterminacy of locations, we impose the following normalization, which ensures a one-to-one mapping between the distance matrix and the resulting multi-dimensional product space. We start with arbitrarily numbering products, and set product 1 to zero: $\ell_1 = 0$, thus locating product 1 at the origin. For the second product, we choose the first coordinate equal to the distance between products 1 and 2 and the remaining coordinates to zero: $d_{1,2} = \|\ell_1 - \ell_2\|$. We iterate this procedure for each product i, setting its first

[10] While longer chains of production provide more bounds to take in to account, they make the estimation of sparse distance matrices noisy. We address this challenge in Section 4 by splitting products by broad categories of goods.

$i-1$ coordinates to preserve distances to each product $k<i$: $d_{i,k}=\|\ell_i-\ell_k\|$ and the rest to 0.[11]

Second, we locate multi-product firms in the product space.[12] Assuming transformed firm-product absolute advantages $T\left(a_{\omega,\nu}\right)$ are observed with errors $\epsilon_{\omega,\nu}$, the location of firm ω satisfies

$$\|\ell_\nu-\ell_\omega\|=T\left(a_{\omega,\nu}\right)\epsilon_{\omega,\nu},$$

for every product ν produced by firm ω.

Minimizing the sum of squared errors projects the firm's location onto the hyperplane determined by the products it produces $\{\nu\}$ and is given by the combination

$$\ell_\omega=\sum_\nu \frac{T\left(a_{\omega,\nu}\right)^{-2}}{\sum_k T\left(a_{\omega,k}\right)^{-2}}\ell_\nu. \tag{13}$$

The location of a firm in Equation (13) can be interpreted as a weighted average of the location of the products that it makes, with zero weight on products it does not produce. This solution implies that firms are closer to products in which they have better absolute advantage.

Given the firm's location, we can then recover the counterfactual absolute advantages for unproduced products, which we denote with $\tilde{a}_{\omega,\tau}$, by inverting $T\left(\tilde{a}_{\omega,\tau}\right)$ in the following equation

$$\|\ell_\omega-\ell_\tau\|\equiv T\left(\tilde{a}_{\omega,\tau}\right). \tag{14}$$

The proximity of a firm to unproduced products depicted in Equation (14) comes from information embedded in the product space and the location of the firm. Together they determine each firm's counterfactual absolute advantages to produce products.

We now turn to the data and construction of clusters and product spaces which will form the environment for our estimates of firm behavior.

4 Data and Estimation Procedure

In this section, we present the data, construct product spaces, and report summary statistics. For each cluster and year, we construct two different product spaces based on measuring absolute advantages when the underlying markups are constant or variable

[11]This is technically done in $M-1$ steps, requiring $M-1$ dimensions, where M is the number of products, but any representation of at least this dimension suffices, and so we use M for ease of exposition.

[12]Single-product firms will have no part in our analysis besides being used in price indexes of competition because they are not informative about co-production patterns as in Equation (9).

and compare these to product spaces based on the Combined Nomenclature and on the product relatedness measures from the product complexity literature.

4.1 Data

We rely on firm-product-level data from Danish firms, spanning from 2000 to 2018, provided by Denmark Statistics. We use two data sources: the Production Statistics (VARS) and the Trade Statistics (UHDI). The Production Statistics is a survey in which manufacturing firms with at least 10 employees are required to report their sales in quantities and values for each product they produce. Sales are recorded independently of the market in which the product is sold, thereby including both domestic and export sales. This is similar to PRODCOM data available in many European countries, and we combine it with trade statistics to isolate production for the domestic market. In the Trade Statistics, firms report their exports and imports by product and destination. Products are reported according to the eight-digit level of the Combined Nomenclature (CN) code, with the firm (CVRNR) being the reporting unit, equivalent to the Harmonized System classification at the 6-digit level.

Our data preparation closely follows Buus et al. (2022), who have provided the code for the estimation of marginal costs using the method proposed by De Loecker et al. (2016). To account for changes in product categories over time, we employ the algorithm proposed by Bernard et al. (2012), aggregating categories to the so-called CN8+ level. A product is defined at both the CN8+ and the unit of measurement level. For most CN8+ categories, firms report the same unit of measurement (kg, number, etc.). In some rare instances, the same product code is recorded with different units, and we consider these as separate products.

For the production function estimation, we also rely on the Firm Statistics Register and the Firm Accounts Statistics, which provide balance sheet, employment, and finance information on the universe of private sector Danish firms.

4.2 Absolute Advantage Estimation

Our model in Section 2.1 defines the absolute advantage of a firm ω in product ν in year t as the ratio between the firm's marginal cost $m_{\omega,\nu,t}$ and the average marginal cost across firms for the same product, $\overline{m}_{\nu,t}$. To compute these absolute advantages, we must first estimate marginal costs, which are typically unobserved by the researcher. We consider two approaches consistent with our CoPaTh framework. First, we analyze the specific case of CES preferences, where the elasticity of substitution $\sigma_{\omega,\nu} = \sigma$ is constant across all firms and products. In this setting, we invert the demand system to obtain firm-level markups. Second, we adopt the approah of by De Loecker, Goldberg, Khandelwal, and Pavcnik (2016) (DGKP) to estimate production functions and markups. These markups allow us to back out the elasticity $\sigma_{\omega,\nu}$ and recover marginal costs as the ratio between unit values and estimated markups.

While the CES method relies on strong assumptions such as constant markups and a particular nesting of demand across varieties, it does not rely on production function estimation or availability of quantities or unit values and is therefore widely applicable. By leveraging sophisticated methods and rich data, the DGKP method allows for economies of scope, scale and selection into multi-product status as well as productivity improvements that accompany production adoption. The DGKP results are therefore our preferred estimates despite more stringent data requirements.[13]

4.2.1 Demand Inversion (CES)

By combining the production statistics with the trade statistics, we can compute the total domestic sales for each firm-product-year as the difference between the total production value and total export value. We apply the same procedure to derive the domestic quantity.[14]

We calculate the marginal costs for each firm-product by assigning a value of $\sigma = 5$ for all sectors, using the domestic unit values to compute the price index of Equation (4) and combining it with domestic sales to compute the marginal costs from Equations (7) and absolute advantages from Equation (6). As we only use information on Danish firms and their domestic sales, in computing the marginal cost, we implicitly ignore foreign firms exporting to Denmark in the calculation of aggregate revenues and price indexes. This does not pose an issue in measuring domestic absolute advantage, since its derivation only requires data on domestic product expenditure shares.

4.2.2 Production Function Estimation (DGKP)

We apply the procedure outlined by De Loecker et al. (2016), which integrates the estimation of a production function and markups, to derive marginal costs for each firm-product and thereby absolute advantage. In this scenario, there is no need to assume a demand function to determine marginal costs, as these are inferred from the production decisions

[13] The industrial organization literature has proposed a number of alternatives for estimating multi-product firm behavior, see De Roux et al. (2021); Dhyne et al. (2022); Orr (2022). Some of these approaches would not be in line with our theoretical model and would lead to different theoretical formulas for our counterfactuals. These methods are of course limited to estimating absolute advantages for *produced* products. Our paper complements this line of work by developing an algorithm to recover absolute advantages for *unproduced* products, conditional on estimates.

[14] We exclude observations with negative domestic sales or quantity, which are likely due to firms engaging in carry-along trade (Bernard et al., 2018).

of firms. We closely follow Buus et al. (2022) in estimating markups and refer readers to their paper for further details.

The key challenge in estimating a production function for multi-product firms arises from a lack of information on how various inputs are allocated to each product. To address this challenge, De Loecker et al. (2016) propose estimating production functions for single-product firms. The production function is estimated using a control function approach that accounts for unobserved heterogeneity in both productivity and input prices. We follow Buus et al. (2022) and also incorporate product-specific export status, number of export destinations, and the square of the number of export destinations into the control functions. However, unlike Buus et al. (2022), we do not include information on export support, as it is only available for a subset of the years.

To calculate markups, this approach uses the results of cost minimization of flexible inputs.[15] Under perfect competition, with no markups, the revenue share of an input and its output elasticity are equal to each other. The markup is defined as the wedge between the revenue share of a variable production input and its output elasticity. By estimating the production function, we determine the output elasticity of materials and we interpret the ratio of this elasticity to the revenue share of materials as the price-cost markup.

Once the markups are obtained, we calculate the marginal cost as the ratio between unit value and markup. To avoid outliers, we calculate the absolute growth rate of firm-product marginal cost changes and exclude the top 2% of those. Moreover, we drop the top and bottom 3% of estimated markups. We amalgamate the CES and DGKP samples to ensure that the estimation of the product space, according to the two different marginal costs, is derived using an identical set of firm-products.

In summary, for each firm-product-year in our sample, we have estimated the absolute advantages using either the CES demand function (CES) or a production function estimation (DGKP). The subsequent step in the empirical analysis involves dividing firm-products into clusters of connected products.

4.3 Cluster Construction and Analysis

To address dimensionality challenges in our analysis and to increase precision estimating a highly sparse product space, we divide our sample into 12 sectors.[16] We further refine

[15] For details, see De Loecker and Warzynski (2012) and De Loecker (2021).

[16] We begin with the following 15 sectors: Animal products (CN 2-digit 01-05), Vegetable products (06-15), Foodstuffs (16-24), Mineral products (25-27), Chemical products (28-38), Plastics and rubber (39-40), Leather and Fur (41-43), Wood products (44-49), Textiles (50-63), Footwear and headgear (64-67), Stone and glass (68-71), Metals (72-83), Machinery and electrical (84-85), Transportation (86-89), and Miscellaneous (90-97). We then drop any sector with less than 10 products, to be constistent with our model of monopolistic competition with differentiated goods, which leaves us with 12 sectors.

our selection of products and firms for analysis, as the estimation of the product space is applicable only to a set of products that are either co-produced or linked by chains of co-production. For this reason, we exclude single-product firms from our sample, as they do not offer any insight into co-production capability. For the remainder of the analysis, we exclusively focus on firms that are multi-product.

Within each sector, we identify the clusters of all products connected by co-production to each other in *any* year within each sector. We further refine the sample of products to avoid the possibility that one cluster in one year breaks down by 2 or more separate clusters, or sub-clusters. If there are more than one sub-cluster per cluster each year, we keep the sub-cluster with the largest number of products in it, and drop the products that are in the other clusters for all years. This procedure drops approximately 15% percent of observations at the firm-product-year level and 19% at the product-year level. For each sector, we find a single cluster.

By definition, each cluster stands distinct from the others. Given this distinction, we will refer to a firm as a firm-cluster for the rest of our analysis. Notably, firms that extend across multiple clusters represent 8% of the total number of firms and account for 23% of overall sales. These are firms whose products span multiple sectors, e.g., a firm making both chemical products and plastic products. In a robustness test, we employ a more aggregated sector definition, yielding similar outcomes (refer to Appendix E).[17] When dealing with larger clusters, the proportion of firms spanning multiple clusters diminishes, making up 5% of firms and contributing 4% to sales. Furthermore, with larger clusters, the procedure that drops the smallest sub-clusters per cluster each year leads us to drop 8% percent of observations at the firm-product-year level and 12% at the product-year level.

Table 1 reports the descriptive statistics for the identified clusters. On average, a cluster comprises 88 products and is associated with 53 firms. The distributions of both products and firms exhibit a right skew and display significant interquartile variation. Notably, the counts of products and firms both witnessed a decline around the time of the 2008 financial crisis.

Table 2 presents summary statistics that describe the average size of each cluster, both in terms of products and firms. Among these, the Foodstuffs cluster has the highest count of products and firms, while the Transportation cluster registers the lowest. The Metals and Machinery and Electrical clusters both tally above-average numbers in terms of products and firms. Meanwhile, the Textiles cluster stands out with its large product count but a smaller firm count.

[17] In this context, the sectors are categorized as: Animals/Vegetables/Food (CN 2-digit 01-24), Minerals/Chemicals/Plastics (25-40), Textiles/Footwear (41-43, 50-67), Stone/Metals (68-83), Machinery/Transportation (84-89), and Miscellaneous (44-49, 90-97).

Table 1: Cluster Descriptive Statistics

	Number of Products					Number of Firms				
Year	Avg.	Std.	Med.	25P.	75P.	Avg.	Std.	Med.	25P.	75P.
2000	88	87	53	27	131	62	44	56	26	84
2001	90	83	70	18	138	62	42	54	35	87
2002	91	80	67	23	141	68	40	67	39	95
2003	90	77	62	23	137	61	36	54	39	79
2004	97	77	80	30	150	73	38	68	44	113
2005	94	76	82	18	149	65	41	53	38	107
2006	98	81	89	18	153	60	43	56	20	100
2007	68	58	47	23	102	41	32	32	16	66
2008	73	61	63	22	108	41	27	33	20	63
2009	77	66	71	14	123	44	30	42	18	67
2010	75	63	63	13	113	47	33	36	19	79
2011	83	67	78	23	127	49	36	36	19	78
2012	88	69	84	24	136	46	32	38	21	80
2013	88	71	66	21	141	46	35	32	19	78
2014	91	74	68	16	155	47	36	40	20	78
2015	93	76	66	21	156	46	38	33	17	82
2016	94	73	92	20	151	47	38	37	17	81
2017	97	76	93	18	151	46	39	39	9	79
2018	97	76	98	25	146	48	37	39	17	81
Average	**88**	**73**	**73**	**21**	**137**	**53**	**37**	**44**	**24**	**83**

In each year, we compute average (Avg.), standard deviation (Std.), median (Med.), and 25th and 75th percentiles (25P. and 75P.) of the number of products (first four columns) and number of firms (last four columns) across clusters. In each year, there are 12 clusters (for 12 sectors defined as groups of CN 2-digit codes). The last row (Average) reports the average of the statistics across years.

4.4 Quantifying Co-production Capability

Using the CES demand system inversion and the DGKP approach, we derive two distinct measures for the marginal costs for product ν of firm ω, $m^i_{\omega,\nu}$, where i =CES, DGKP. This implies observed absolute advantages $a^i_{\omega,\nu} = m^i_{\omega,\nu}/\overline{m}^i_{\nu}$.

In our framework, we assume that there is an increasing and monotone relationship between the observed absolute advantages and the distance between a firm to a product through a positive and increasing transformation $T\left(a^i_{\omega,\nu}\right)$ with $T(0) = 0$ as in Equation (12). We represent this firm-product distance as a true distance $d_{\omega,\nu}$ and a measurement error $\varepsilon_{\omega,\nu}$: $d_{\omega,\nu}\varepsilon_{\omega,\nu}$ and consider three transformations, with log as our primary specification:

1. Log (Baseline): $d_{\omega,\nu}\varepsilon_{\omega,\nu} = \ln\left(1 + a^i_{\omega,\nu}\right)$. The natural logarithm of absolute advantage plus one to maintain positive distances when $a^i_{\omega,\nu} < 1$.

2. Level: $d_{\omega,\nu}\varepsilon_{\omega,\nu,t} = a^i_{\omega,\nu}$. Here, we directly use absolute advantage.

3. Inverse Hyperbolic Sine: $d_{\omega,\nu}\varepsilon_{\omega,\nu} = \ln\left(a^i_{\omega,\nu} + \sqrt{1 + \left(a^i_{\omega,\nu}\right)^2}\right)$, another approach

Table 2: Cluster Descriptive Statistics

	Number of Products		Number of Firms	
Cluster	Avg.	Std.	Avg.	Std.
Animal Products	106	35	37	19
Foodstuffs	227	19	107	13
Mineral products	21	7	14	8
Chemical Products	148	26	46	11
Plastics and rubber	54	13	41	10
Wood products	19	7	38	25
Textiles	178	33	39	22
Stone and glass	15	5	25	15
Metals	103	23	95	21
Machinery and Electrical	127	16	92	12
Transportation	10	3	10	2
Miscellaneous	48	11	89	34

In each cluster, we compute average (Avg.) and standard deviation (Std.) of the number of products (first two columns) and number of firms (last two columns) across years.

maintaining positive distances (see Chen and Roth (2024)).

We now have all the elements to determine the location of products and firms and the respective distances. For each cluster-year, we construct the product space using the procedure outlined in Section 3, and the three transformations for absolute advantage.

4.5 Product Space Based on Combined Nomenclature (CN)

To evaluate our method against a discrete classification system, we consider an alternate product space derived from the CN classification system. In the CN classification, the distances between firms and products, as well as between firms themselves, are discrete. We assume these distances can have one of four distinct values.

Recall that a product ν is defined as a combination of a unit and CN8+ code. For every firm-cluster in a given year, the core product $\bar{\nu}_\omega$ is defined as the product ν with the largest sales for the firm ω. Then, the distance of a firm ω to a product ν is given by

$$d_{\omega,\nu} = 1 + \mathbf{1}_{\mathrm{CN6}(\bar{\nu}_\omega)\neq\mathrm{CN6}(\nu)} + \mathbf{1}_{\mathrm{CN4}(\bar{\nu}_\omega)\neq\mathrm{CN4}(\nu)} + \mathbf{1}_{\mathrm{CN2}(\bar{\nu}_\omega)\neq\mathrm{CN2}(\nu)}.$$

To elucidate, a firm has a distance of one to all products that fall under the same CN6 code of its core product. It has a distance of two to products under the same CN4 code of its core, but a different CN6 code, and so on.

4.6 Product Space Based on Product Relatedness (RCA)

To further validate the efficacy of our approach, we compare the performance of our model in predicting product adoption using a product space based solely on co-production, as it is typically done in the literature on the product space for countries (Hidalgo et al., 2007). Extending this concept to firm co-production, we compute a measure of proximity between products based on how frequently firms have Revealed Comparative Advantage (RCA) in their production at the same time, and then construct a measure of relatedness of firms to products. To compare with our baseline exercise, we compute the measure of product proximity and firm-product relatedness within our clusters.

First, we define the $RCA_{\omega,\nu}$ of a firm ω for product ν as the share of firm ω revenues on product ν, relative to the aggregate sales share of product ν in the total cluster sales

$$RCA_{\omega,\nu} = \frac{r_{\omega,\nu}}{\sum_{\nu} r_{\omega,\nu}} / \frac{\sum_{\omega} r_{\omega,\nu}}{\sum_{\omega}\sum_{\nu} r_{\omega,\nu}}. \tag{15}$$

Then, we compute a measure of proximity between product ν and ν' as

$$\phi_{\nu,\nu'} = \min\{\text{Prob}(RCA_{\omega,\nu} > 1 | RCA_{\omega,\nu'} > 1),\ \text{Prob}(RCA_{\omega,\nu'} > 1 | RCA_{\omega,\nu} > 1)\}. \tag{16}$$

$\text{Prob}(RCA_{\omega,\nu} > 1 | RCA_{\omega,\nu'} > 1)$ is the probability that a firm has an RCA on ν conditional on having an RCA on ν'. It is computed as the share of firms that have an RCA on ν out of the total number of firms that have an RCA on ν'.

Finally, we compute product relatedness of RCA between firm ω and a product ν as

$$\text{Relatedness}_{\omega,\nu} = \sum_{\nu'} \phi_{\nu,\nu'} \cdot 1\left(RCA_{\omega,\nu'} > 1\right) / \sum_{\nu'} \phi_{\nu,\nu'}, \tag{17}$$

where $1\left(RCA_{\omega,\nu'} > 1\right)$ is one when $RCA_{\omega,\nu'} > 1$ and zero otherwise. The distance of a firm to a product is then given by

$$d_{\omega,\nu} = 1 - \text{Relatedness}_{\omega,\nu}. \tag{18}$$

4.7 Product Space Analysis

Here, we present descriptive information about product spaces. The summary statistics for the estimated distances in the year 2000 are shown in Table 3. We observe that both product-to-product and firm-to-product distances are smaller in the CES than in the DGKP specification. Table 4 presents the correlation between various measures for the year 2000. All correlations are statistically significant and positive. Notably, the correlation between DGKP and CES distances is 26%,[18] while there is a 29% correlation between the CN and RCA measures, showing the affinity of these pairs of measures.

Product spaces maintain stability over time. We examine the distribution of the growth rate in the distance between a firm and a product, defined as the difference $d_{\omega,\nu,t} - d_{\omega,\nu,t-1}$, since the distances are expressed in logarithms. Both the average and the median firm-to-product growth rate of distance is close to zero on a year to year basis. In the CES framework, distances vary by 5 percentage points within the 25-75 percentile range, indicating general stability of our approach. Distances growth for DGKP is slightly more volatile, varying by 14 percentage points for the same range. We provide these results in Table D.1 of the appendix.

Table 3: Estimated Distances: Summary Statistics (2000)

	Product-to-Product		Firm-to-Product	
	CES	DGKP	CES	DGKP
Average	0.39	0.75	0.23	0.58
Std. Dev.	0.23	0.32	0.09	0.21
5th Perc.	0.11	0.27	0.10	0.20
10th Perc.	0.15	0.38	0.13	0.31
25th Perc.	0.23	0.56	0.17	0.46
50th Perc.	0.33	0.69	0.22	0.58
75th Perc.	0.50	0.91	0.28	0.70
90th Perc.	0.70	1.11	0.34	0.81
95th Perc.	0.82	1.29	0.39	0.88

The table reports the distribution of product-to-product and firm-to-product distances in the year 2000. The sample comprises all product-to-product and all firm-to-product distances across all clusters. For confidentiality reasons, we divide distances data in 100 bins, compute the average distance within bin, and report these values in the rows with percentiles.

Table 4: Correlations Between Measures (2000)

	Distance (CES)	Distance (DGKP)	Distance (CN)	Distance (RCA)
Distance (CES)	1			
Distance (DGKP)	0.258***	1		
Distance (CN)	0.097***	0.149***	1	
Distance (RCA)	0.121***	0.134***	0.288***	1

***: significant at 99%, ** at 95%, * at 90%. The table reports the correlation values across distances in our four specifications.

Figure 2 illustrates the distribution of firm-product distances for the year 2000 across four different specifications. The CES and DGKP approaches yield continuous distributions. Conversely, the CN approach categorizes distances into four groups, predominantly

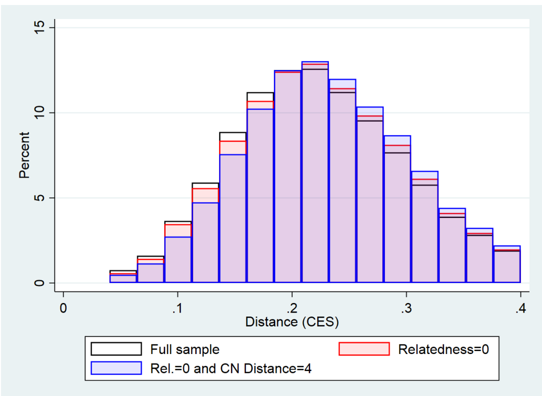
[18]There is a higher correlation of 41% between the estimated marginal costs in these two approaches.

placing them at a distance of four, in different CN 2-digit sectors. Using the RCA measure, a significant peak in the data occurs at the maximum distance of one, indicating that many products are not co-produced within the same year. This peak accounts for over 80% of observations.

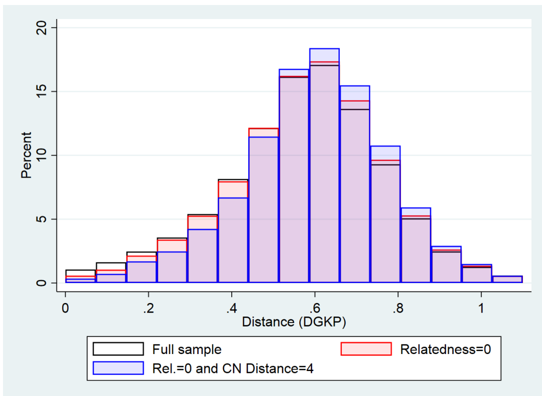
Figures 2a and 2b display the distribution of firm-product distances for the CES and DGKP methods under two conditions: when RCA-based relatedness is zero (red bars) and when both RCA-relatedness is zero and CN distance is at its maximum (blue bars). These distributions shift to the right and exhibit considerable variance. This rightward shift results from the correlation between distance measures based on absolute advantages and those derived from CN classifications and RCA co-occurrence. The substantial variance in CES and DGKP distances, even within these restrictive samples, indicates that relying solely on discrete classification systems or RCA co-occurrence metrics fails to capture the full granularity of counterfactual absolute advantages.

Figure 2: Distribution of Distances (2000)

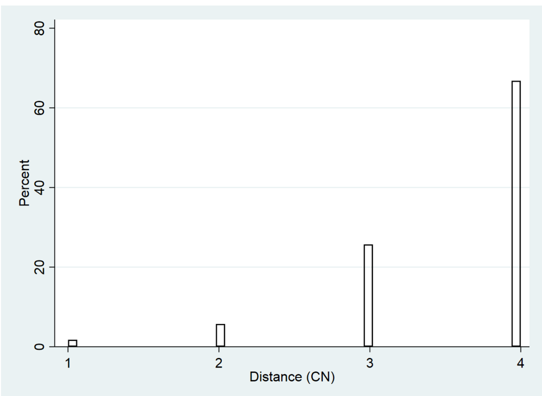
(a) CES



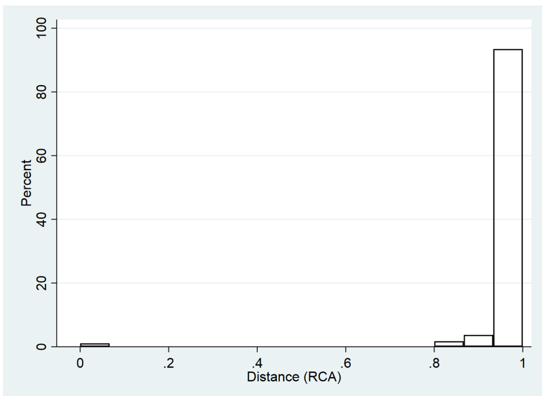
(b) DGKP



(c) CN



(d) RCA

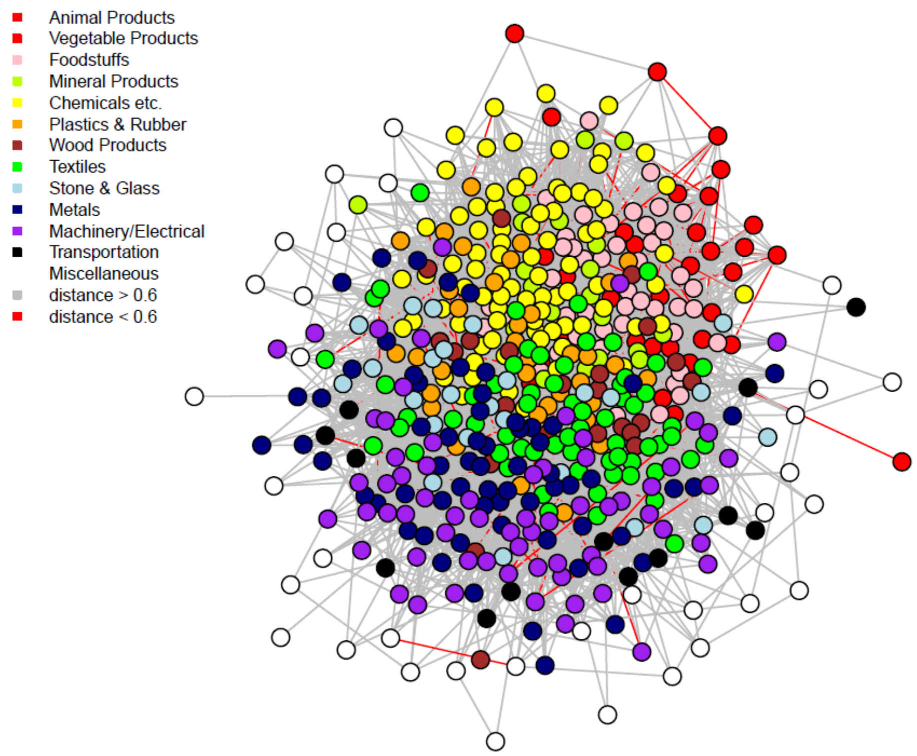


This figure illustrates the distribution of firm-product distances for the year 2000 using various approaches. For confidentiality, we have restricted the samples to ensure that each bar contains a sufficient number of observations. Consequently, the percentage values shown here are higher than those calculated from the full sample.

To offer some insight into the relationship between distances within this space, we conduct several regression analyses in Appendix D.1. We find that products with higher sales are generally more isolated than those with smaller sales. Moreover, we observe a hump-shaped relationship between product-to-product distances and co-production frequency, suggesting that products are further apart when only a few firms co-produce them. However, as the number of co-producing firms increases beyond certain thresholds, the distance between products starts to decrease. Moreover, we find that firms tend to be closer to the products they produce, which validates our approach.

Graphical Representation. Due to confidentiality, we are unable to provide a visual representation of product spaces at the CN8 level but can do so with aggregation to the CN4 level. Our clustering algorithm produces a single cluster at this level, meaning that each product is directly or indirectly linked to all others.

Figure 3: Product Space



This figure provides a graphical representation of the product space and is generated using product-to-product distances estimated using our algorithm to a cluster of all directly and indirectly connected CN 4-digit products. We use the DGKP approach with the baseline log transformation. Marginal costs from DGKP are estimated at the CN 8-digit product-units which distinguish 8-digit products by unit of measure. To obtain marginal costs at the CN 4-digit product-unit we use a quantity-weighted average across the CN 8-digit product-units within that CN 4-digit product-unit. In the Figure 3, lines are drawn to create the smallest tree including all products. Subsequently, lines below a certain cutoff distance are added to the graph. Some distances are omitted to enhance readability.

Figure 3 provides a network representation of the product space using maximum-spanning trees, constructed following methods similar to those in Hidalgo et al. (2007). The figure illustrates that products within similar sectors tend to cluster together, such as Chemical products alongside Plastic and Rubber, or Machinery in close proximity to Metals and Transportation. Additionally, the periphery is marked by an assortment of miscellaneous products.

We now move to our estimates of firm behavior with counterfactual absolute advantages from the product space.

5 Product Adoption and Counterfactual Advantages

While firm-level shocks, such as a boost in productivity, and product-level shocks, like rising demand or changes in tariffs, might prompt a firm to introduce or discontinue a product, the precise prediction of *which* firms will be involved has largely been overlooked in the literature. This section first quantifies the predictive capability of the product space in determining which products will be introduced. In Appendix D.3, we also examine the predictive capability regarding product discontinuations. The section continues with an IV strategy to test whether demand shocks at the product level leads firms with better counterfactual advantages to adopt them. The results show that this is indeed the case.

5.1 Product Adoption and Counterfactual Advantage

For each firm, we identify products not produced in its initial year of observation. For example, if a firm first appears in the data in 2000, we consider as candidate products all items in the firm's cluster not produced in that year. In subsequent years, we construct a product introduction indicator $\text{Intro}_{\omega,\nu,t}$ for firm ω, product ν, and year t, equal to 1 if the firm produces the product in that year, and 0 otherwise. We normalize this outcome by the average product introduction rate, which is approximately 0.4 percent.

We estimate the following regression:

$$\text{Intro}_{\omega,\nu,t} = \beta \cdot \ln\left(1 + \tilde{a}_{\omega,\nu,t-1}\right) + \phi_{\omega,t} + \rho_{\nu,t} + \epsilon_{\omega,\nu,t}, \tag{19}$$

where the main independent variable is the lagged transformed counterfactual absolute advantage $T(\tilde{a}_{\omega,\nu,t-1}) = \ln(1 + \tilde{a}_{\omega,\nu,t-1})$, as in our baseline product space. We lag this variable to avoid contemporaneous feedback from product adoption into the construction of the product space.[19]

[19] Similar structural equations for product introduction are derived in Boehm et al. (2022), although here markups are allowed to vary by firm-time and product-time.

For specifications using CN classification or product relatedness (RCA), we substitute $\ln(1+\tilde{a}_{\omega,\nu,t-1})$ with the corresponding lagged firm-product distance measure.

Following our theoretical framework, we include product-year fixed effects ($\rho_{\nu,t}$) to control for time-varying shocks to demand or technology at the product level. Firm-year fixed effects ($\phi_{\omega,t}$) absorb contemporaneous shocks such as changes in productivity, ensuring that identification comes from within-firm variation across products rather than across firms.[20] Throughout this section, standard errors are clustered at the product-year level. The findings are illustrated in Table 5. The columns demonstrate that product adoption rates diminish as counterfactual advantage (CN distance, product relatedness) worsens.

Table 5: Product Introduction and Advantage (All)

	Dependent Variable: $\text{Intro}_{\omega,\nu,t} = 1$			
	(CES)	(DGKP)	(CN)	(RCA)
Lagged Advantage	-3.685***	-1.543***	-1.688***	-13.474***
	(0.469)	(0.200)	(0.081)	(0.980)
Product-Time FE	Yes	Yes	Yes	Yes
Firm-Time FE	Yes	Yes	Yes	Yes
R^2	0.07	0.07	0.08	0.07
# Obs.	645290	645290	645290	645290

Results from OLS estimation of (19). Cluster standard errors in parenthesis. Cluster: product-year. ***: significant at 99%, ** at 95%, * at 90%.

Discrete versus Continuous Classifications. As shown in Table 5, distance measures based on the CN classification are predictive of product introduction. This suggests that the CN classification captures meaningful proximity between products: firms tend to introduce products located in nearby CN categories. However, a key limitation of the CN classification is its discreteness. It cannot differentiate between products within the same category, whereas our continuous measures are more granular. For example, a firm whose core product is in cotton fabrics would, under the CN classification, appear equally likely to expand into apparel and into carpets, even though these outputs differ substantially in production processes and inputs. This is because both apparel and carpets lie in different CN 2-digit codes relative to cotton but are grouped within the same cluster, and hence, are at the same distance to the firm.

To illustrate the value of continuous classifications, Table 6 reports results from estimating (19) while controlling for CN distance bins via fixed effects. The coefficients

[20] As discussed above, when a firm produces across multiple, highly distinct CN two-digit clusters, we treat each firm-cluster pair as a distinct production unit. This ensures that the fixed effects also absorb cluster-year shocks.

on absolute advantage remain negative and statistically significant, indicating that our continuous measures retain predictive power even within CN-defined hierarchies.

Table 6: Product Introduction and Advantage (CN Controls)

	Dependent Variable: $\text{Intro}_{\omega,\nu,t} = 1$		
	(CES)	(DGKP)	(RCA)
Lagged Advantage	-1.615***	-0.737***	-8.914***
	(0.466)	(0.199)	(0.943)
Product-Time FE	Yes	Yes	Yes
Firm-Time FE	Yes	Yes	Yes
CN Distance FE	Yes	Yes	Yes
R^2	0.08	0.08	0.08
# Obs.	645290	645290	645290

Results from OLS estimation of (19). Cluster standard errors in parenthesis. Cluster: product-year. ***: significant at 99%, ** at 95%, * at 90%.

Absolute Advantage versus RCA We perform a similar analysis using distance measures based on RCA. These metrics also exhibit predictive power for product introductions, with results comparable to those obtained using our continuous classification. However, a major limitation of the RCA approach is its lack of coverage: over 80% of firm-products have a relatedness of zero, which corresponds to distance taking the maximum value of one. As shown in Table 7, our advantage based measures retain predictive power even when the RCA-based relatedness measure is uninformative. This is reflected in the negative and statistically significant coefficients on absolute advantage.[21]

Table 7: Product Introduction and Advantage (Relatedness=0)

	Dependent Variable: $\text{Intro}_{\omega,\nu,t} = 1$		
	(CES)	(DGKP)	(CN)
Lagged Advantage	-1.631***	-0.472***	-1.194***
	(0.413)	(0.172)	(0.089)
Product-Time FE	Yes	Yes	Yes
Firm-Time FE	Yes	Yes	Yes
Lagged Relatedness =	0	0	0
R^2	0.08	0.08	0.08
# Obs.	542445	542445	542445

Results from OLS estimation of (19). Cluster standard errors in parenthesis. Cluster: product-year.. ***: significant at 99%, ** at 95%, * at 90%.

[21] As a robustness check, we estimate a specification based on ranking products by their distance rather than using the continuous values of distance. The results, reported in Appendix D.2, confirm that firms are more likely to introduce products that are closer in rank, even when including CN fixed effects or when RCA-based relatedness is zero.

5.2 Product Adoption and Demand Shocks

As highlighted in the preceding section, a myriad of reasons can prompt a firm to introduce a new product. These can span from firm-specific factors, like enhancements in efficiency, to product-specific elements, such as changes in demand. Here we examine how a positive demand shock influences the likelihood of a product being introduced, and how this relationship is modulated by a firm's counterfactual absolute advantages in the presence of a demand shock.

To quantify a demand shock that is credibly exogenous to supply conditions, we instrument for positive export demand shocks on product introduction. We estimate

$$\text{Intro}_{\omega,\nu,t} = \beta \cdot \ln\left(1 + \tilde{a}_{\omega,\nu,t-1}\right) + \gamma \cdot \text{Exports}_{\nu,t} \times \ln\left(1 + \tilde{a}_{\omega,\nu,t-1}\right) + \phi_{\omega,t} + \rho_{\nu,t} + \kappa_{\omega,\nu} + \epsilon_{\omega,\nu,t}, \tag{20}$$

where the dependent variable $\text{Intro}_{\omega,\nu,t}$ and counterfactual advantages mirror those used in regression (19). $\text{Exports}_{\nu,t}$ denotes the log of the total exports from Denmark of product ν in year t. Equation (20) closely follows the linear probability model for product adoption predicted by our theory in Equation (9), where the probability of introducing a product is log-linear in absolute advantage and depends on product-year, firm-year, and firm-product fixed effects. We restrict the sample to firm-products that have been introduced in the years considered, thereby excluding firm-products ω, ν such that $\max_t \text{Intro}_{\omega,\nu,t} = 0$. Therefore, the identifying variation is in the switch of a product from being not produced to being produced within each firm.

We instrument $\text{Exports}_{\nu,t}$ by adhering to the approach by Hummels et al. (2014), similar to Dhyne et al. (2021), using either global exports or those from comparable nations to instrument for Danish exports by calculating $\text{Exports IV}_{\nu,t} = \text{Log} \sum_{k \neq DNK} \text{Exports}_{k,\nu,t}$, with the total exports of all countries except Denmark.[22] Notice that export data, sourced from BACI, is reported at the 6-digit level, while our products are defined at a more granular 8-digit level.

In Table 8, we show that our key parameter of interest, the interaction between distance and exports, is negative and statistically significant under the CES and DGKP cost specifications. This outcome shows that the impact of export demand shocks to increase product introduction is amplified for products where co-production capability is stronger and attenuated for products with less counterfactual advantage. When demand

[22] As a robustness, we consider $\text{Exports IV}_{\nu,t} = \text{Log} \sum_{k \in K} \text{Exports}_{k,\nu,t}$ where K denotes two different sets of countries: the same set of countries as Autor et al. (2013): Australia, Finland, Germany, Japan, New Zealand, Spain, Switzerland, and USA, and a set of EU countries: Austria, Belgium, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, Netherlands, Portugal, Spain, and Sweden. Results are in appendix D.4, along with the results of the first stage regressions.

Table 8: Export Shocks and Product Adoption

	Dependent Variable: $\text{Intro}_{\omega,\nu,t} = 1$							
	(CES) OLS	(CES) IV	(DGKP) OLS	(DGKP) IV	(CN) OLS	(CN) IV	(RCA) OLS	(RCA) IV
Advantage	47.60	401.5***	25.54	96.81***	3.720	8.426	-60.62***	-52.54***
	(66.15)	(121.4)	(17.99)	(31.23)	(5.447)	(10.03)	(4.094)	(19.20)
Advantage	-10.77*	-41.70***	-3.429**	-9.555***	-0.561	-0.971	-3.737***	-20.43
X Exports	(5.650)	(10.52)	(1.522)	(2.659)	(0.465)	(0.869)	(1.088)	(39.32)
Firm-Time FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm-Product FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Prod-Time FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
# Obs.	13782	13782	13782	13782	13782	13782	13782	13782
F-Stat		404.49		634.29		484.58		7.66

Results from OLS estimation of (20). Cluster standard errors in parenthesis. Cluster: product-year. ***: significant at 99%, ** at 95%, * at 90%. We instrument Log $\text{Exports}_{\nu t} \times \ln\left(1+\tilde{a}_{\omega,\nu,t-1}\right)$ with Log Exports $\text{IV}_{\nu t} \times \ln\left(1+\tilde{a}_{\omega,\nu,t-1}\right)$ where Log Exports $\text{IV}_{\nu t} = \text{Log} \sum_{k \neq DNK} \text{Exports}_{k\nu t}$ is the total exports of all countries except Denmark.

increases for a product, closer firms in the product space supply that demand. Notice that both the CN-based and the RCA-based distance measures do not attain statistical significance in the IV specification, indicating that our advantage based method better predicts adoption following a demand shock.

6 Firm Growth Potential and Gains from Variety

In this section, we use the counterfactual absolute advantages recovered above to investigate how potential product expansion by firms can predict growth. We introduce two new measures akin to Market Potential in the trade literature (Redding and Venables, 2004). The first is *Revenue Potential*, which quantifies latent revenues for a firm in unproduced products representing pro-growth opportunities. The second is *Competition Potential*, which quantifies potential losses from competitors introducing the current products of each firm. These measures rely upon on the model developed in Section 2.

Using the counterfactual absolute advantages, we can compute the counterfactual revenues a firm would obtain upon introducing a new variety. To do so, we extend the model of Section 2 to allow aggregate revenues per product and demand elasticities of substitution to vary over time. Given aggregate market size, counterfactual absolute advantages recovered above and counterfactual markups recovered here, we can compute counterfactual revenues for each firm-product-time combination in each cluster according to the structure of our model.

Mirroring our construction of product spaces above, we construct these potential measures for both constant and variable markups. For constant markups (CES), we again set the elasticity of substitution to $\sigma = 5$ for a markup of 1.25. In the variable case, estimate the elasticities of substitution using the firm-product specific estimated markups with the DGKP methodology. For unproduced products, we predict markups with firm and product fixed effects, which we show are consistent so long as all products and firms are in the same cluster. We detail this step of estimating counterfactual variable markups in Section 6.1. We then define our Revenue and Competition Potential measures and estimate their impact on firm growth. Finally, we explore the implications of counterfactual absolute advantage and markups for gains from variety.

6.1 Counterfactual Variable Markups

Here, we detail how to estimate counterfactual firm-product markups with a model of firm and product fixed effects. Given the markup $\mu_{\omega,\nu,t}$ for product ν produced by firm ω, we propose a simple structure consistent with CoPaTh preferences. We estimate markups within each cluster according to an OLS estimate of $\beta_t = \begin{bmatrix} \beta_{\nu,t} & \beta_{\omega,t} \end{bmatrix}$ in

$$\log \mu_{\omega,\nu,t} = \begin{bmatrix} \chi_{\omega,t} & \chi_{\nu,t} \end{bmatrix} \beta_t^T + \eta_{\omega,\nu,t}, \tag{21}$$

where $\chi_{\nu,t}$ and $\chi_{\omega,t}$ are row vectors corresponding to the fixed effects of variety ν and firm ω and log is the natural logarithm. This specification allows us to estimate firm- and product specific components of markups for every produced product. We then use estimated $\hat{\beta}_t$ by cluster to predict counterfactual markups $\hat{\mu}_{\omega,\nu,t}$ as

$$\hat{\mu}_{\omega,\nu,t} \equiv \exp\left(\begin{bmatrix} \tilde{\chi}_{\omega,t} & \tilde{\chi}_{\nu,t} \end{bmatrix} \hat{\beta}_t^T\right), \tag{22}$$

where $\tilde{\chi}_{\nu,t}$ and $\tilde{\chi}_{\omega,t}$ are row vectors ranging over all unproduced firm-variety pairs in each cluster. While estimates of β_t in Equation (21) are not uniquely identified due to different possible normalizations, all estimates obtained from this regression imply identical predicted counterfactual markups in Equation (22) within cluster, as stated here with the proof provided in Appendix C.[23]

Proposition. *Out of sample firm-product markup predictions using two way firm and product fixed effects are consistent within a cluster.*

With counterfactual advantages and markups in hand, we now can define our measures of Revenue and Competition Potential.

[23] The predicted markups are trimmed so markups range between 1.1 and 2.5, by changing values outside those ranges to the closest in the range. Our positive results are robust to a much wider potential range of markups, although the normative results are sensitive to the presence of high markups. Adapting the markups estimated from this method for normative conclusions is an area for future work.

6.2 Revenue and Competition Potential

Some firms, while having counterfactual advantages for products with large hypothetical revenues, do not (yet) produce the corresponding goods. This implies that such firms possess substantial potential to augment their revenue, relative to a firm with less underlying co-production capability. We now define a measure to capture this firm potential for growth. First, we compute revenues from product ν for a firm ω for our potential measures, applying Equation (7) from above, using predicted markups and counterfactual advantages for produced (P) and also unproduced (U) products for model consistency.[24] Accordingly, revenues for produced products are denoted with $r^P_{\omega,\nu,t}$ and with $r^U_{\omega,\nu,t}$ for unproduced products. These revenues equal

$$r^P_{\omega,\nu,t} = \left(\frac{\hat{\mu}_{\omega,\nu,t}\tilde{a}_{\omega,\nu,t}}{A\left(\hat{\boldsymbol{\mu}}_{\boldsymbol{\omega,\nu,t}} \odot \tilde{\boldsymbol{a}}_{\boldsymbol{\omega,\nu,t}}\right)}\right)^{\frac{1}{\hat{\mu}_{\omega,\nu,t}-1}} R_{\nu,t}, \quad \text{and} \quad r^U_{\omega,\nu,t} = \left(\frac{\dot{\mu}_{\omega,\nu,t}\dot{a}_{\omega,\nu,t}}{A\left(\dot{\boldsymbol{\mu}}_{\boldsymbol{\omega,\nu,t}} \odot \dot{\boldsymbol{a}}_{\boldsymbol{\omega,\nu,t}}\right)}\right)^{\frac{1}{\dot{\mu}_{\omega,\nu,t}-1}} R_{\nu,t}, \tag{23}$$

where tilde superscripts indicate counterfactuals for produced varieties and dot superscripts for counterfactuals of produced varieties plus the additional hypothetical variety of the firm.

We define Revenue Potential (RP) in the following way:

$$RP_{\omega,t} \equiv \frac{\sum_\nu r^U_{\omega,\nu,t}}{\sum_\nu r^P_{\omega,\nu,t} + \sum_\nu r^U_{\omega,\nu,t}}, \tag{24}$$

which would approach one as a firm produced one or a few of low-revenue products and zero if a firm were somehow to produce all products in a cluster. Consequently, a higher RP is associated with elevated potential revenues if the firm were to broaden its scope.

We also define a Competition Potential (CP) index, using the equations defined above. Let $\mathring{r}^P_{\omega,\nu,t}$ be firm $\omega's$ revenues in producing product ν, under a hypothetical worst-case scenario in which all firms in the cluster opt to compete by producing varieties of ν. CP is defined as one minus the ratio of a firm ω's worst case total revenue to the baseline revenue of firm ω

$$CP_{\omega,t} = 1 - \frac{\sum_\nu \mathring{r}^P_{\omega,\nu,t}}{\sum_\nu \tilde{r}^P_{\omega,\nu,t}}. \tag{25}$$

[24]The key difference between the two revenue formulas lies in the argument of the aggregator A. For currently produced products, we compute the aggregator based on the set of goods already in all firms' portfolios and use it to determine revenues. Any new unproduced product would change the price vector for the calculation of the aggregator. Therefore, for each unproduced product, we recompute the aggregator under the hypothetical scenario in which the product is added and then use the new value of A to compute the corresponding counterfactual revenue. This procedure is applied separately to each candidate product.

As with Equation (23), it is straightforward to calculate each $\mathring{r}_{\omega,\nu,t}$.[25] If all firms produced all of a firm's products, then there would be zero Competition Potential, whereas in a competitive limit of a large number of firms that could produce the firm's product, CP would be 1. Hence, higher Competition Potential corresponds to threats to future revenues.

With this theoretical grounding here, our Potential measures can be computed under constant and variable markups to consistently compare them. We now estimate the influence of these measures on firm growth.

6.3 Sales and Scope Growth

Here we estimate how RP and CP can predict firm growth in sales and scope. We estimate the following regression

$$y_{\omega,c,t} = \beta_1 RP_{\omega,c,t-1} + \beta_2 CP_{\omega,c,t-1} + \phi_\omega + \xi_{c,t} + \epsilon_{\omega,c,t}, \tag{26}$$

where $y_{\omega,c,t}$ represents two performance variables for firm ω in cluster c at year t, comprising: 1) the growth rate of domestic sales, $\ln \text{Sales}_{\omega,c,t} - \ln \text{Sales}_{\omega,c,t-1}$, with domestic sales constructed as described above and 2) the growth rate of the number of products, or scope, $\ln \text{Scope}_{\omega,c,t} - \ln \text{Scope}_{\omega,c,t-1}$. $RP_{\omega,c,t-1}$ and $CP_{\omega,c,t-1}$ are the lagged values of equations (24) and (25), respectively. We include firm ϕ_ω and cluster-time $\xi_{c,t}$ fixed effects, clustering errors at the firm level.

Table 9: Firm Sales Growth

	Dependent Variable: Growth Rate of Sales	
	Constant Markups	Variable Markups
Lagged RP	0.525***	0.358***
	(0.072)	(0.084)
Lagged CP	-0.077*	-0.164***
	(0.044)	(0.045)
Cluster-Time FE	Yes	Yes
Firm FE	Yes	Yes
R^2	0.16	0.15
# Obs.	8435	8435

Results from OLS estimation of (26). Clustered standard errors in parenthesis. Cluster: firm. ***: significant at 99%, ** at 95%, * at 90%.

[25] Computing these revenues requires recalculating the aggregator A under the counterfactual scenario in which all firms that were not previously producing product ν begin to do so.

Table 10: Firm Scope Growth

	Dependent Variable: Growth Rate of Scope	
	Constant Markups	Variable Markups
Lagged RP	0.005	0.108***
	(0.020)	(0.026)
Lagged CP	-0.124***	-0.049***
	(0.016)	(0.016)
Cluster-Time FE	Yes	Yes
Firm FE	Yes	Yes
R^2	0.14	0.13
# Obs.	8435	8435

Results from OLS estimation of (26). Clustered standard errors in parenthesis. Cluster: firm. ***: significant at 99%, ** at 95%, * at 90%.

Tables 9 and 10 show that a higher Revenue Potential is associated with higher growth rates of sales and of scope. In fact, the coefficient on RP is positive and statistically significant in almost all specifications. As we control for firm fixed effects, the interpretation is that the growth rate of sales and scope is higher, relative to the average firm growth rate, in the presence of initial higher revenue potential. In contrast, the negative and significant coefficients for CP in both tables indicate that higher potential competition is associated with lower sales and scope growth. In Appendix D.6, we present an instrumental variable approach to estimate (26). In this approach, we instrument for the RP and CP of a firm using the average RP and CP of the firms' competitors. The results demonstrate robustness of this specification.

Finally, we use our rich set of counterfactuals and flexible model setting to quantify potential gains from variety.

6.4 Potential Gains from Variety

Here, we use our counterfactual absolute advantages and markups to quantify potential gains from variety by quantifying how price levels would change when all observed firms in a cluster manufacture all products. Since the CoPaTh price index is ideal, this measures the change in real income from Gains from Variety. This scenario assumes that all firms in a cluster that are not currently producing a given product experience a positive shock that leads them to introduce such products, as reflected in the counterfactuals computed above.[26] Such a welfare gain counterfactual would usually be determined by the assumption of an unseen cost distribution and therefore quantitatively subjective in

[26]While this ignores potential fixed costs of entering a variety, it provides an empirically informed estimate of gains from variety, holding income constant. Given the difficulty of finding convincing measures of fixed costs when goods are *actually* produced, much less unproduced, we leave this endeavor for future work, especially as changes in income may be correlated with product adoption and exit decisions.

model choice, as is done in well known frameworks such as Eaton and Kortum based models (Frechet draws) or Melitz based models (Pareto draws).

To address this question, we define the Potential Gains from Variety (PGV) for product ν as the ratio of the price index when all firms in the cluster choose to produce product ν at time t, $P\left(\acute{\boldsymbol{p}}_{\boldsymbol{\nu},\boldsymbol{t}}\right)$, to the actual price index $P\left(\boldsymbol{p}_{\boldsymbol{\nu},\boldsymbol{t}}\right)$ computed using counterfactuals from above as

$$PGV_{\nu,t} \equiv \frac{P\left(\acute{\boldsymbol{p}}_{\boldsymbol{\nu},\boldsymbol{t}}\right)}{P\left(\boldsymbol{p}_{\boldsymbol{\nu},\boldsymbol{t}}\right)} = \frac{P\left(\acute{\boldsymbol{p}}_{\boldsymbol{\nu},\boldsymbol{t}}/\overline{m}_{\nu,t}\right)}{P\left(\boldsymbol{p}_{\boldsymbol{\nu},\boldsymbol{t}}/\overline{m}_{\nu,t}\right)} = \frac{P\left(\acute{\boldsymbol{\mu}}_{\boldsymbol{\nu},\boldsymbol{t}} \odot \acute{\boldsymbol{a}}_{\boldsymbol{\nu},\boldsymbol{t}}\right)}{P\left(\boldsymbol{\mu}_{\boldsymbol{\nu},\boldsymbol{t}} \odot \boldsymbol{a}_{\boldsymbol{\nu},\boldsymbol{t}}\right)}, \tag{27}$$

where the second equation follows from the homogeneity of the price index. The difference between the two price indexes is driven solely by gains from new varieties implied by the counterfactual pattern of advantages and markups, similar to those proposed by Feenstra (1994).

PGV is computed for every product ν and year t. Subsequently, we compute a weighted average and weighted standard deviation of $PGV_{\nu,t}$ for each cluster and each year, where the weights are the expenditure shares on each product ν. In Table 11, we report the average and standard deviation of PGV across years for each cluster.

Table 11: Potential Gains from Variety by Sector ($\times$100)

	Constant Markups		Variable Markups	
Cluster	Avg.	Std.	Avg.	Std.
Animals	72.3	21.7	19.4	18.2
Foodstuffs	69.6	19.0	14.8	21.0
Mineral products	73.0	24.8	40.3	16.8
Chemical Products	77.0	21.4	19.5	23.0
Plastics and rubber	75.4	22.8	22.4	21.9
Wood	79.3	18.7	35.5	29.0
Textiles	70.2	15.5	23.0	24.4
Stone and glass	80.5	26.2	47.5	27.7
Metals	78.6	23.9	24.6	20.0
Machinery and electrical	70.8	22.5	14.0	21.8
Transportation	82.2	23.0	63.5	30.4
Miscellaneous	83.2	15.4	25.9	23.4
Average	**76.0**	**21.2**	**29.2**	**23.1**
Green Products	80.3	22.1	27.7	25.1

Average and standard deviation of the EP for each cluster. We set $\sigma = 5$ in the calculation of the price index for the CES case. For the DGKP case, details on the calculations of the price index are in the main text.

Gains from variety under constant markups come from firms entering each product market, with potentially better advantages than the existing pool of firms. Under variable markups, gains can additionally arise from a wide locus of advantages and markups with a higher potential for consumer gains. Under constant markups, the price index falls on average to 76% of its initial value, with Foodstuffs recording the largest drop (69.6%) and Miscellaneous the smallest (83.2%). Under variable markups, we obtain a larger average drop in the price index to 29.2%. Quantitatively, the difference in results is largely driven by high-markup varieties. When all firms begin to make more differentiated high-markups products, the gains are larger relative to when all products are equally differentiated. Similarly, if firms with high markups make all products, there are extra gains due to the high differentiation provided to consumers.

As a sample application, we also compute the Potential Gains from Variety for a list of 6-digit "green products" provided by Mealy and Teytelboym (2023), based on the work by Mealy and Teytelboym (2022) and Andres et al. (2023) in the last row of the table.[27] The results are in line with our baseline results: if all firms begin to make all of the green products, the price index for these products would be 28-80% that of the current one. This suggests high gains in varieties where removable barriers to entry exist.

7 Conclusion

This paper develops a framework to predict which products firms will adopt by recovering counterfactual absolute advantages over unproduced goods. We construct a data-driven product space from observed co-production patterns to map firms' latent capabilities, embed it in a multi-product model with heterogeneous markups, and show that firm–product proximity predicts adoption and firm growth, outperforming Combined Nomenclature and Revealed Comparative Advantage based relatedness at highly granular levels. Empirically, using Danish data and an identification strategy based on export-demand shocks, we find that firms closer in the product space are more likely to enter following positive demand shifts. The framework also delivers firm-specific indices, Revenue Potential and Competition Potential that predict firm growth. Finally, an index of Potential Gains from Variety links economy wide latent capabilities to aggregate supply responses.

A policy relevant implication of the model is that targeted interventions, such as those aiming to foster adoption of strategic or green technologies, could benefit from mapping firms' counterfactual absolute advantages by product cluster. Counterfactual simulations indicate sizable potential welfare gains if firms expanded production to currently unproduced products: between 16–30% under constant markups and 46–86% under variable markups.

This framework opens several avenues for future research. First, extending the model

[27]Out of the 295 products provided in the original list, only 123 are included in our clusters. The remaining products that are not matched are either not produced in Denmark or excluded from our clusters given the algorithm we outlined above.

to explicitly incorporate dynamic firm decision-making under uncertainty or economies of scope could enrich predictions about product adoption. Second, applying the product space mapping to other contexts, such as services, global value chains, or emerging economies, would further test the external validity of the method. Third, incorporating demand-side features, such as brand spillovers or product characteristics, could complement the current supply-driven approach. Finally, integrating entry/exit responses among firms more directly into the product space could better capture competitive responses and market structure dynamics.

References

Andres, P., P. Mealy, N. Handler, and S. Fankhauser (2023): "Stranded nations? Transition risks and opportunities towards a clean economy," *Environmental Research Letters*, 18, 045004.

Arkolakis, C., S. Ganapati, and M.-A. Muendler (2021): "The Extensive Margin of Exporting Products: A Firm-level Analysis," *American Economic Journal: Macroeconomics*, 13.

Autor, D., D. Dorn, and G. H. Hanson (2013): "The China Syndrome: Local Labor Market Effects of Import Competition in the United States," *American Economic Review*, 103, 2121–2168.

Bernard, A. B., E. J. Blanchard, I. Van Beveren, and H. Vandenbussche (2018): "Carry-Along Trade," *Review of Economic Studies*, 86, 526–563.

Bernard, A. B., S. J. Redding, and P. K. Schott (2010): "Multiple-Product Firms and Product Switching," *American Economic Review*, 100, 70–97.

Bernard, A. B., I. Van Beveren, and H. Vandenbussche (2012): "Concording eu trade and production data over time," Publisher: CEPR Discussion Paper No. DP9254.

Berry, S., J. Levinsohn, and A. Pakes (1995): "Automobile Prices in Market Equilibrium," *Econometrica*, 841–890.

Bloom, N., M. Schankerman, and J. Van Reenen (2013): "Identifying Technology Spillovers and Product Market Rivalry," *Econometrica*, 81, 1347–1393.

Boehm, J., S. Dhingra, and J. Morrow (2022): "The comparative advantage of firms," *Journal of Political Economy*, 130, 3025–3100.

Buus, M. T., J. R. Munch, J. Rodrigue, and G. Schaur (2022): "Do export support programs affect prices, quality, markups and marginal costs? Evidence from a natural policy experiment," *Review of Economics and Statistics*, 1–45.

Chen, J. and J. Roth (2024): "Logs with zeros? Some problems and solutions," *Quarterly Journal of Economics*, 139, 891–936, publisher: Oxford University Press.

De Loecker, J. (2021): "Comment on *(Un)pleasant...* by Bond et al (2020)," *Journal of Monetary Economics*.

De Loecker, J., P. K. Goldberg, A. K. Khandelwal, and N. Pavcnik (2016): "Prices, Markups, and Trade Reform," *Econometrica*, 84, 445–510.

De Loecker, J. and F. Warzynski (2012): "Markups and firm-level export status," *American economic review*, 102, 2437–2471, publisher: American Economic Association.

De Roux, N., M. Eslava, S. Franco, and E. Verhoogen (2021): "Estimating production functions in differentiated-product industries with quantity information and external instruments," Tech. rep., National Bureau of Economic Research.

Dhingra, S. (2013): "Trading Away Wide Brands for Cheap Brands," *American Economic Review*, 103, 2554–84.

Dhyne, E., A. K. Kikkawa, M. Mogstad, and F. Tintelnot (2021): "Trade and Domestic Production Networks," *Review of Economic Studies*, 88, 643–668.

Dhyne, E., A. Petrin, V. Smeets, and F. Warzynski (2022): "Theory for Extending Single-Product Production Function Estimation to Multi-Product Settings," *National Bureau of Economic Research Working Paper*.

Ding, X. (2023): "Industry Linkages from Joint Production," *Working Paper*.

Eckel, C., L. Iacovone, B. Javorcik, and J. P. Neary (2015): "Multi-product firms at home and away: Cost- versus quality-based competence," *Journal of International Economics*, 95, 216–232.

Eckel, C. and P. J. Neary (2010): "Multi-Product Firms and Flexible Manufacturing in the Global Economy," *Review of Economic Studies*, 77, 188–217.

Escolar, E. G., Y. Hiraoka, M. Igami, and Y. Ozcan (2023): "Mapping firmsâ locations in technological space: A topological analysis of patent statistics," *Research Policy*, 52, 104821.

Feenstra, R. and H. Ma (2007): "Optimal Choice of Product Scope for Multiproduct Firms under Monopolistic Competition," *in E. Helpman, D. Marin and T. Verdier, eds., The Organization of Firms in a Global Economy, Harvard University Press.*

Feenstra, R. C. (1994): "New Product Varieties and the Measurement of International Prices (Appendix)," *American Economic Review*, 84, 157–177.

Feenstra, R. C. and J. A. Levinsohn (1995): "Estimating Markups and Market Conduct with Multidimensional Product Attributes," *The Review of Economic Studies*, 62, 19.

Flach, L. and M. Irlacher (2018): "Product versus Process: Innovation Strategies of Multiproduct Firms," *American Economic Journal: Microeconomics*, 10, 236–77.

Fontagne, L., A. Secchi, and C. Tomasi (2018): "Exporters product vectors across markets," *European Economic Review*, 110, 150–180.

Girard, M. and A. Trau (2004): "Implementing the North American Industry Classification System: The Canadian Experience," Tech. rep., Statistics Canada.

Goldberg, P. K., A. K. Khandelwal, N. Pavcnik, and P. Topalova (2010): "Multiproduct firms and product turnover in the developing world: Evidence from india," *Review of Economics and Statistics*, 92, 1042–1049.

Grant, M. (2023): "Where Do the Data Come From? Endogenous Classification in Administrative Data," *Working Paper.*

Guevara, M. R., D. Hartmann, M. Aristaran, M. Mendoza, and C. A. Hidalgo (2016): "The Research Space: Using Career Paths to Predict the Evolution of the Research Output of Individuals, Institutions, and Mations," *Scientometrics*, 109, 1695–1709.

Hausmann, R. and C. A. Hidalgo (2011): "The network structure of economic output," *Journal of economic growth*, 16, 309–342.

Hausmann, R., J. Hwang, and D. Rodrik (2007): "What you export matters," *Journal of Economic Growth*, 12, 1–25.

Hausmann, R., D. P. Stock, and M. A. Yildirim (2022): "Implied Comparative Advantage," *Research Policy*, 51, 104143.

Herkenhoff, P., S. Krautheim, and P. Saure (2024): "A simple model of buyerâseller networks in international trade," *European Economic Review*, 170, 104868.

Hidalgo, C. A. (2021): "Economic complexity theory and applications," *Nature Reviews Physics*, 3, 92–113.

Hidalgo, C. A., B. Klinger, A.-L. Barabàsi, and R. Hausmann (2007): "The Product Space Conditions the Development of Nations," *Science*, 317, 482–487.

Hoberg, G. and G. Phillips (2016): "Text-based network industries and endogenous product differentiation," *Journal of Political Economy*, 124, 1423–1465.

Hummels, D., R. Jorgensen, J. Munch, and C. Xiang (2014): "The wage effects of offshoring: Evidence from Danish matched worker-firm data," *American Economic Review*, 104, 1597–1629.

Irlacher, M. (2022): "Multi-Product Firms in International Economics," *CESifo Working Paper*.

Jacobs, G. and C. ONeill (2003): "On the reliability (or otherwise) of SIC codes," *European Business Review*, 15, 164–169.

Jaffe, A. (1986): "Technological Opportunity and Spillovers of R&D: Evidence from Firms, Patents, Profits and Market Value," *American Economic Review*, 76, 984–1001.

Jakel, I. C., C. Eckel, L. Macedoni, and R. Riezman (2023): "Explorations in Product Space: Multi-Product Firms and their Product Portfolios," *Working Paper*.

Jara-Figueroa, C., B. Jun, E. L. Glaeser, and C. A. Hidalgo (2018): "The Role of Industry-Specific, Occupation-Specific, and Location-specific Knowledge in the Growth and Survival of New Firms," *Proceedings of the National Academy of Sciences*, 115, 12646–12653.

Lancaster, K. J. (1966): "A new approach to consumer theory," *Journal of political economy*, 74, 132–157.

Macedoni, L. (2022): "Large multiproduct exporters across rich and poor countries: Theory and evidence," *Journal of Development Economics*, 156, 102835.

Macedoni, L. and M. J. Xu (2022): "Flexibility and Productivity: Towards the Understanding of Firm Heterogeneity," *International Economic Review*, 63, 1055–1108.

Matsuyama, K. and P. Ushchev (2017): "Beyond CES: Three Alternative Classes of Flexible Homothetic Demand Systems," *SSRN Electronic Journal*.

——— (2020): "Constant Pass-Through," *Working Paper*, publisher: CEPR Discussion Paper No. DP15475.

Mayer, T., M. J. Melitz, and G. I. Ottaviano (2014): "Market Size, Competition, and the Product Mix of Exporters," *American Economic Review*, 104, 495–536.

Mealy, P. and A. Teytelboym (2022): "Economic complexity and the green economy," *Research Policy*, 51, 103948.

——— (2023): "Green Transition Navigator," .

Neffke, F., M. Henning, and R. Boschma (2011): "How Do Regions Diversify over Time? Industry Relatedness and the Development of New Growth Paths in Regions: ECONOMIC GEOGRAPHY," *Economic Geography*, 87, 237–265.

Nocke, V. and S. Yeaple (2014): "Globalization and multiproduct firms," *International Economic Review*, 55, 993–1018.

Orr, S. (2022): "Within-Firm Productivity Dispersion: Estimates and Implications," *Journal of Political Economy*, 130, 2771–2828.

Pellegrino, B. (2025): "Product Differentiation and Oligopoly: A Network Approach," *American Economic Review*, 115, 1170–1225.

Redding, S. and A. J. Venables (2004): "Economic geography and international inequality," *Journal of International Economics*, 62, 53–82.

Rosen, S. (1974): "Hedonic prices and implicit markets: product differentiation in pure competition," *Journal of political economy*, 82, 34–55.

A Product Classifications Across HS, SIC and NAICS

The organizing principles of classification systems vary: "NAICS differs significantly from the SICs because it is based on a single organizing principle, contrary to the SICs where entities are sometimes grouped according to production-oriented principles and sometimes grouped according to demand-based principles. NAICS is based on a production-oriented or supply based conceptual framework [...] very similar production processes are grouped." (Girard and Trau, 2004). Table A.1 provides example products from the Harmonized System which vary by main classification across two North American classification systems, the SIC and NAICS and even change main classification over time.

Table A.1: Product Classification Differences

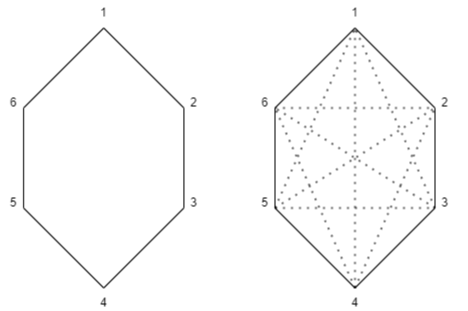
	HS		SIC		NAICS	
	Code	Description	Code	Description	Code	Description
1	1005904040	Popcorn, Unpopped, Except Seed	119	Cash Grains, Not Elsewhere Classified-Con. (Major Group 01 Agricultural production-crops DIVISION-A Agriculture, Forestry, and Fishing)	Before 2014: 111150 After 2014: 311999	111150 Corn Farming (111 Crop Production Sector) 311999 All Other Miscellaneous Food Manufacturing (311-Food Manufacturing)
2	714101000	Cassava (manioc) frozen	2037	Frozen Fruits, Fruit Juices, and Vegetables (Major Group 20 Food and kindred products DIVISION-D Manufacturing)	111130	Dry Pea and Bean Farming (111 Crop Production Sector)
3	1504102000	Fish-liver oils and their fractions	2077	Animal and Marine Fats and Oils (Major Group 20 Food and kindred products DIVISION-D Manufacturing)	114111	Finfish Fishing (114 Fishing, Hunting and Trapping Sector)
4	2301200010	Flours, meals and pellets, of meat or meat offal; greaves	2077	Animal and Marine Fats and Oils (Major Group 20 Food and kindred products DIVISION-D Manufacturing)	114111	Finfish Fishing (114 Fishing, Hunting and Trapping Sector)
5	1702202210 1702202290 1702202410 1702202490 1702202810 1702202890 1702204010 1702204090	Maple sugar and maple syrup	2099	Food Preparations, Not Elsewhere Classified (Major Group 20 Food and kindred products DIVISION-D Manufacturing)	111998	All Other Miscellaneous Crop Farming (111 Crop Production Sector)
6	5808101000	Braids, in the piece	2241	Narrow Fabric and Other Smallwares Mills: Cotton, Wool, Silk, and Manmade Fiber (Major Group 22 Textile mill products DIVISION-D Manufacturing)	315990	Apparel Accessories and Other Apparel (315-Apparel Manufacturing, 314-Textile Product Mills)

Figure B.1: Distances Across Co-produced Products

(a) Co-produced Distances

$$\begin{bmatrix} 0 & v_{12} & NA & NA & NA & v_{16} \\ v_{21} & 0 & v_{23} & NA & NA & NA \\ NA & v_{32} & 0 & v_{34} & NA & NA \\ NA & NA & v_{43} & 0 & v_{45} & NA \\ NA & NA & NA & v_{54} & 0 & v_{56} \\ v_{61} & NA & NA & NA & v_{65} & 0 \end{bmatrix}$$

(b) Length 1 (c) New Length 2



B "Filling in" Cost Bounds Algorithm

This section describes how to determine the upper and lower bounds for the distances between products that are not co-produced by linked by co-production. As shown in Figure B.1, we start with known upper and lower distance averages v_{ij} for pairs v_{12}, v_{23}, v_{34}, v_{45}, v_{56}, and v_{61}. These distances are based on direct co-production, with at one intermediary firm involved, assigning them a chain length of one.

Our next step involves recovering the distance bounds for pairs v_{ij} where products i and j are not co-produced but but bounds can be inferred through a two-step co-production chain via an intermediary product k, thus having a chain length of two (since it involves two intermediary firms). To recover these "chain length two" distances, we apply the triangle inequality. Specifically, we calculate the upper bound by finding the smallest maximum difference $|v_{ik} - v_{kj}|$ across all intermediaries k, and the lower bound by identifying the largest minimum sum $v_{ik} + v_{kj}$ across all k. These calculated bounds are then used to populate distances for indirect connections, as illustrated in Figure B.2.

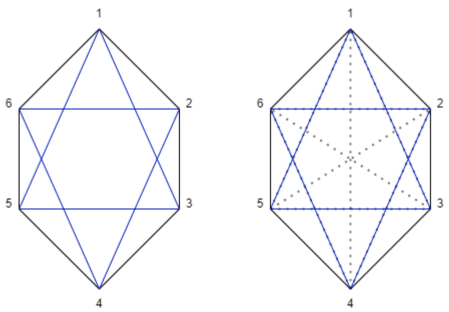
We repeat this process with the newly populated distances, although now the upper and lower bounds of length 2 correspond to length 3 separations of the original distances. This algorithm continues until $N - 1$ steps for N products are completed, which will result in a maxi min of lower bound distances and mini max of upper bound distances. This process is depicted in Figure B.3.

Once the matrices of upper and lower bound distances are populated as described, we compute the final distances d_{ij} as the average of the corresponding upper and lower v_{ij}.

Figure B.2: Distances Across Products Not Co-produced (Length 2)

(a) Indirectly Linked Distances

$$\begin{tikzcd}
0 & v_{12} \arrow[r] & v_{13} & NA & v_{15} & v_{16} \arrow[l] \arrow[d] \\
v_{21} & 0 & v_{23} \arrow[u] \arrow[r] & v_{24} & NA & v_{26} \\
v_{31} & v_{32} \arrow[l] \arrow[d] & 0 & v_{34} \arrow[u] \arrow[r] & v_{35} & NA \\
NA & v_{42} & v_{43} \arrow[l] \arrow[d] & 0 & v_{45} \arrow[u] \arrow[r] & v_{46} \\
v_{51} & NA & v_{53} & v_{54} \arrow[l] \arrow[d] & 0 & v_{56} \\
v_{61} \arrow[u] \arrow[r] & v_{62} & NA & v_{64} & v_{65} & 0
\end{tikzcd}$$

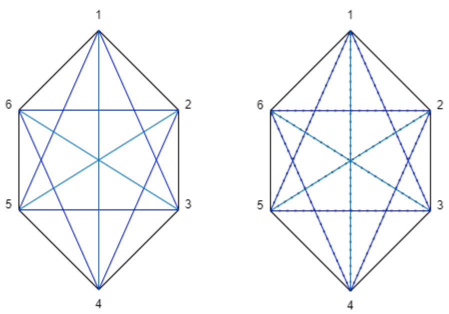


(b) Length 2 (c) New Length 3

Figure B.3: Distances Across Products Not Co-produced (Length 3)

(a) Indirectly Linked Distances

$$\begin{tikzcd}
0 & v_{12} & v_{13} \arrow[r] & v_{14} & v_{15} \arrow[d] & v_{16} \\
v_{21} & 0 & v_{23} & v_{24} \arrow[u] \arrow[r] & v_{25} & v_{26} \arrow[l] \arrow[d] \\
v_{31} & v_{32} & 0 & v_{34} & v_{35} \arrow[u] \arrow[r] & v_{36} \\
v_{41} & v_{42} \arrow[l] \arrow[d] & v_{43} & 0 & v_{45} & v_{46} \\
v_{51} \arrow[u] \arrow[r] & v_{52} & v_{53} \arrow[l] \arrow[d] & v_{54} & 0 & v_{56} \\
v_{61} & v_{62} \arrow[u] \arrow[r] & v_{63} & v_{64} & v_{65} & 0
\end{tikzcd}$$



(b) Length 3 (c) New Length 4

C Uniqueness of Counterfactual Markups

Proposition. *Out of sample firm-product markup predictions using two way firm and product fixed effects are consistent within a cluster.*

Proof. Fix a product ν' not produced by firm ω' (and therefore out of sample) and we wish to show for any two estimators of Equation (21), say $\widehat{\beta}$ and $\widehat{\gamma}$, that they produce the same predicted values, i.e. $\begin{bmatrix} m_{\nu'} & m_{\omega'} \end{bmatrix} \left(\widehat{\beta} - \widehat{\gamma}\right) = 0$. Considering Equation (21) in matrix block form $\mu = M\beta + \eta$, since $\widehat{\beta}$ and $\widehat{\gamma}$ are OLS estimators, $M\left(\widehat{\beta} - \widehat{\gamma}\right) = 0$. Now define ν_0 as some product produced by firm $\omega_0 \equiv \omega$ and by assumption, ν' and ν_0 are connected by co-production, so there exists a sequence of firms $\{\omega_i\}_{i=0}^{N}$ such that each firm ω_i produces ν_i and ν_{i+1} and $\nu_{N+1} = \nu'$. Let $\underline{m}_i$ be the row vector of M corresponding to firm ω_i producing ν_i and $\overline{m}_i$ the row vector corresponding to firm ω_i producing ν_{i+1}. Now if firm ω_0 did produce ν_N, it would have a row vector of M equal to

$$\widetilde{m} \equiv \overline{m}_0 - \sum_{i=1}^{N-1} \left(\underline{m}_i - \overline{m}_i\right),$$

where each addition of $\underline{m}_i - \overline{m}_i$ preserves the fixed effect of ω_0 but swaps the fixed effect of ν_i for ν_{i+1}, ending with the fixed effect of firm $\omega' = \omega_0$ and variety $\nu' = \nu_N$. Since $\widetilde{m}$ is a sum of actually produced firm-product combinations and therefore rows of M, $M\left(\widehat{\beta} - \widehat{\gamma}\right) = 0$ implies that $\widetilde{m}\left(\widehat{\beta} - \widehat{\gamma}\right) = 0$ as desired.

D Estimation

D.1 Distance Analysis

Here we provide descriptive statistics regarding estimated distances in each cluster which inform our model based absolute advantages.

Stability of product space. Table D.1 provides the summary statistics of the growth rate in the distance between a firm and a product as $d_{\omega,\nu,t} - d_{\omega,\nu,t-1}$.

Table D.1: Changes in Firm-to-Product Distances: Summary Statistics

	CES	DGKP
Average	-0.01	-0.05
Std. Dev.	10.26	23.67
5th P.	-17.07	-40.23
10th P.	-12.08	-29.47
25th P.	-5.73	-14.44
50th P.	-0.06	-0.18
75th P.	5.59	13.73
90th P.	11.44	27.64
95th P.	15.34	36.94

The table reports the distribution of the yearly percentage change in firm-to-product distances (all values are multiplied by 100). The sample comprises all firm-to-product distances across all clusters. For confidentiality reasons, we divide the percentage change in distances in 100 bins, compute the average change within bin, and report these values in the rows with percentiles.

Shape and Characteristics of the Product Space. First, we focus on product-to-product distances, and estimate the following model

$$d_{\nu,\nu',t} = \alpha \text{Log Sales}_{\nu,t} + \beta \text{Log Sales}_{\nu',t} + \gamma \text{Co-production}_{\nu,\nu',t} + FE_{c,t} + \epsilon_{\nu,\nu',t}, \quad \text{(D.1)}$$

where Log $\text{Sales}_{\nu,t}$ represents the total sales for product ν summed over all multi-product firms. The term $\text{Co-production}_{\nu,\nu',t}$ quantifies the extent to which the two products are co-produced. This is measured either by the number of firms co-producing both products ν and ν', or by the proportion of firms that co-produce these products relative to the greater of the number of firms producing ν or ν'.

Table D.2: Product-to-Product Distances

	Dependent Variable: Product-to-Product Distance					
	(CES)	(CES)	(CES)	(DGKP)	(DGKP)	(DGKP)
Log Sales ν	0.008***	0.003***	0.009***	0.009***	0.007***	0.009***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Log Sales ν'	0.007***	0.002***	0.008***	0.009***	0.007***	0.009***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
# Firms Co-producing ν, ν'		0.233***			0.077***	
		(0.000)			(0.000)	
# Share Firms Co-producing ν, ν'			0.715***			0.243***
			(0.001)			(0.001)
Cluster-Time FE	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.06	0.41	0.41	0.07	0.09	0.09
# Obs.	1447074	1447074	1447074	1447074	1447074	1447074

Results from OLS estimation of (D.1). Standard errors in parenthesis. ***: significant at 99%, ** at 95%, * at 90%.

Table D.2 reveals that products with higher sales are generally more isolated than those with smaller sales. This observation is supported by the positive and statistically significant coefficient for Log Sales across all specifications. Contrary to expectations, the coefficient on co-production is also positive and statistically significant. This suggests that products often produced together tend to be more distantly located. Our analysis below indicates that firms are typically closer to the products within their production scope. Therefore, the observed positive correlation between co-production and product distance implies that our algorithm tends to position firms centrally among the products they produce, with those outside their scope being more distant from the firm but relatively closer to the products within the scope.

Further investigation in Table D.3 explores the possibility of a non-linear relationship between co-production variables and product distances. We observe a hump-shaped relationship, suggesting that products are further apart when only a few firms co-produce them. However, as the number of co-producing firms increases beyond certain thresholds (14 for CES and 10 for DGKP), the distance between products starts to decrease. Products co-produced by more than 28 firms in the CES specification and 21 in the DGKP specification tend to be closer than those never co-produced together. These findings from Tables D.2 and D.3 illustrate that our approach yields insights markedly different from those derived solely from co-production patterns.

Table D.3: Product-to-Product Distances

	Dependent Variable: Product-to-Product Distance			
	(CES)	(CES)	(DGKP)	(DGKP)
Log Sales ν	0.002***	0.005***	0.007***	0.007***
	(0.000)	(0.000)	(0.000)	(0.000)
Log Sales ν'	0.002***	0.004***	0.007***	0.007***
	(0.000)	(0.000)	(0.000)	(0.000)
Co-production (Number)	0.301***		0.104***	
	(0.000)		(0.001)	
Squared Co-production (Number)	-0.011***		-0.005***	
	(0.000)		(0.000)	
Co-production (Share)		2.018***		0.881***
		(0.002)		(0.004)
Squared Co-production (Share)		-1.643***		-0.805***
		(0.002)		(0.004)
Cluster-Time FE	Yes	Yes	Yes	Yes
R^2	0.48	0.56	0.10	0.11
# Obs.	1447074	1447074	1447074	1447074

Results from OLS estimation of (D.1). Standard errors in parenthesis. ***: significant at 99%, ** at 95%, * at 90%.

We extend our analysis to firm-to-product distances using the following regression model

$$d_{\omega,\nu,t} = \alpha \text{Log Sales}_{\nu,t} + \beta \text{Log Sales}_{\omega,t} + \text{Dummy for Produced}_{\omega,\nu,t} + FE_{c,t} + \epsilon_{\omega,\nu',t}, \quad \text{(D.2)}$$

where Log $\text{Sales}_{\omega t}$ represents the total sales of firm ω and Dummy for $\text{Produced}_{\omega,\nu' t}$ is a binary variable that takes the value of one if firm ω manufactures product ν in year t.

The findings, presented in Table D.4, reveal some intriguing patterns. In the CES specification, firms are generally further from products with larger sales, whereas in the DGKP specification, the opposite trend is observed - firms are closer to products with larger sales. Moreover, there is a noticeable decrease in the distance of firms to products as the total sales of the firms increase. Additionally, firms are consistently closer to the products they produce, underlining a strong link between firm production profiles and their proximity to specific products in the product space.

Table D.4: Product-to-Product Distances

	Dependent Variable: Firm-to-Product Distance					
	(CES)	(CES)	(CES)	(DGKP)	(DGKP)	(DGKP)
Log Sales Product	0.002***	0.002***	0.002***	-0.001***	-0.001***	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Log Sales Firm		-0.002***	-0.001***		-0.001***	-0.000***
		(0.000)	(0.000)		(0.000)	(0.000)
Production Dummy			-0.063***			-0.176***
			(0.000)			(0.001)
Cluster-Time FE	Yes	Yes	Yes	Yes	Yes	Yes
Product FE	0.11	0.11	0.13	0.10	0.10	0.13
Firm FE	1394297	1394297	1394297	1394297	1394297	1394297

Results from OLS estimation of (D.2). Standard errors in parenthesis. ***: significant at 99%, ** at 95%, * at 90%.

D.2 Product Adoption by Proximity

As a robustness check, we estimate a version of the product adoption equation that uses rank-based measures of distance. Specifically, we estimate

$$\text{Intro}_{\omega,\nu,t} = \beta \cdot \text{Rank}_{\omega,\nu,t-1} + \phi_{\omega,t} + \rho_{\nu,t} + \epsilon_{\omega,\nu,t}, \quad \text{(D.3)}$$

where $\text{Rank}_{\omega,\nu,t-1}$ denotes the rank of products not produced by firm ω based on counterfactual absolute advantage (CN distance or product relatedness) in year $t-1$. The closest product receives rank 1.

Table D.5 presents the baseline estimates. Across all specifications, we find that product introduction rates decline with distance rank. Moving from the M-th to the $M+1$-th closest product reduces the likelihood of introduction by approximately 0.5 percentage points under both CES and DGKP specifications.[28]

Table D.5: Product Introduction and Product Rankings

	Dependent Variable: Dummy=1 for Product Introduction			
	(CES)	(DGKP)	(CN)	(RCA)
Lagged Rank (=1 closest)	-0.005***	-0.005***	-0.008***	-0.009***
	(0.001)	(0.001)	(0.000)	(0.000)
Product-Time FE	Yes	Yes	Yes	Yes
Firm-Time FE	Yes	Yes	Yes	Yes
R^2	0.07	0.07	0.07	0.07
# Obs.	645290	645290	645290	645290

Results from OLS estimation of (D.3). Cluster standard errors in parenthesis. Cluster: product-year. ***: significant at 99%, ** at 95%, * at 90%.

The results are robust to controlling for CN distance fixed effects (Table D.6) and to restricting the sample to firm-product pairs with zero RCA-based relatedness (Table D.7).

Table D.6: Product Introduction and Product Rankings

	Dependent Variable: Dummy=1 for Product Introduction		
	(CES)	(DGKP)	(RCA)
Lagged Rank (=1 closest)	-0.002***	-0.003***	-0.006***
	(0.001)	(0.001)	(0.000)
Product-Time FE	Yes	Yes	Yes
Firm-Time FE	Yes	Yes	Yes
CN Distance FE	Yes	Yes	Yes
R^2	0.08	0.08	0.08
# Obs.	645290	645290	645290

Results from OLS estimation of (D.3). Cluster standard errors in parenthesis. Cluster: product-year. ***: significant at 99%, ** at 95%, * at 90%.

D.3 Product Drops

In this section, we quantify the ability of the product space to predict product drops. As the measure of distance within a firm is based on absolute advantages that we observe, the estimation of the full product space is not necessary. As a result, we consider this exercise as a sanity check for our measure of absolute advantages. For each firm, we select the products that the firm produces in its first year in the data. For instance, if a firm enters the data set in 2000, the sample of products we consider are the products in the cluster of the firm that the firm produces in 2000. For each product in the years

[28] For the CN classification, product pairs within the same distance bin are randomly ranked due to the discrete nature of the taxonomy. Despite this limitation, the rank-based approach still captures meaningful variation in product proximity.

Table D.7: Product Introduction and Product Rankings

	Dependent Variable: Dummy=1 for Product Introduction		
	(CES)	(DGKP)	(CN)
Lagged Rank (=1 closest)	-0.002***	-0.002***	-0.005***
	(0.000)	(0.000)	(0.000)
Product-Time FE	Yes	Yes	Yes
Firm-Time FE	Yes	Yes	Yes
Lagged Relatedness =	0	0	0
R^2	0.08	0.08	0.08
# Obs.	542445	542445	542445

Results from OLS estimation of (D.3). Cluster standard errors in parenthesis. Cluster: product-year. ***: significant at 99%, ** at 95%, * at 90%.

after entry, we compute a production indicator $\text{Drop}_{\omega,\nu,t}$ for product ν, firm ω, and year t, which equals 1 if the product is not produced (i.e., dropped) and zero otherwise.

We estimate the following regression

$$\text{Drop}_{\omega,\nu,t} = \beta \cdot \ln\left(1 + \tilde{a}_{\omega,\nu,t-1}\right) + a_{\omega,t} + b_{\omega,t}, +\varepsilon_{\omega,\nu,t} \tag{D.4}$$

where $Rank_{\omega,\nu,t-1}$ is the rank of produced products based on the lagged rank of absolute advantage of the product for each firm $\tilde{a}_{\omega,\nu,t-1}$, so that for the farthest product the product rank in the previous year equals one.

Results are shown in Table D.8 and D.9. Products with less absolute advantage are more likely to be dropped, as the coefficient on the lagged absolute advantage is positive and statistically significant in each specification. As shown for the case of product introduction, using a discrete classification cannot shed light on which products, within a certain aggregation of codes, are more likely to get dropped, while our measure based on absolute advantage is able to distinguish between these products.

Table D.8: Product Drop and Product Rankings

	Dependent Variable: Dummy=1 for Product Drop			
	(CES)	(DGKP)	(CN)	(RCA)
Lagged Distance	6.419***	1.550***	0.155***	1.918***
	(0.571)	(0.139)	(0.032)	(0.249)
Product-Time FE	Yes	Yes	Yes	Yes
Firm-Time FE	Yes	Yes	Yes	Yes
R^2	0.60	0.60	0.60	0.60
# Obs.	22642	22642	22642	22642

Results from OLS estimation of (D.3). Cluster standard errors in parenthesis. Cluster: product-year. ***: significant at 99%, ** at 95%, * at 90%.

Table D.9: Product Drop and Product Rankings

	Dependent Variable: Dummy=1 for Product Drop		
	(CES)	(DGKP)	(RCA)
Lagged Distance	4.737***	1.195***	1.589***
	(0.619)	(0.149)	(0.248)
Product-Time FE	Yes	Yes	Yes
Firm-Time FE	Yes	Yes	Yes
CN Distance FE	Yes	Yes	
R^2	0.60	0.60	0.60
# Obs.	22642	22642	22642

Results from OLS estimation of (D.3). Cluster standard errors in parenthesis. Cluster: product-year. ***: significant at 99%, ** at 95%, * at 90%.

D.4 Export Shocks and Product Adoption

Table D.10: Export Shocks - IV = Autor et al. (2013) Countries

	Dependent Variable: Dummy=1 for Product Introduction							
	(CES)	(CES)	(DGKP)	(DGKP)	(CN)	(CN)	(RCA)	(RCA)
Lagged Advantage	47.600	546.423***	25.540	101.694***	3.720	18.018	-60.616***	-50.157***
	(66.152)	(158.317)	(17.992)	(38.854)	(5.447)	(12.369)	(4.094)	(11.136)
Lagged Advantage	-10.772*	-54.372***	-3.429**	-9.974***	-0.561	-1.806*	-3.737***	-25.362
X Exports	(5.650)	(13.698)	(1.522)	(3.309)	(0.465)	(1.067)	(1.088)	(21.181)
Firm-Time FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm-Product FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Prod-Time FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Spec.	OLS	IV	OLS	IV	OLS	IV	OLS	IV
# Obs.	13782	13782	13782	13782	13782	13782	13782	13782
F-Stat		191.63		288.95		228.46		23.54

Results from OLS estimation of (20). Cluster standard errors in parenthesis. Cluster: product-year. ***: significant at 99%, ** at 95%, * at 90%. We instrument Log Exports$_{\nu t} \times \ln\left(1+\tilde{a}_{\omega,\nu,t-1}\right)$ with Log Exports IV$_{\nu t} \times \ln\left(1+\tilde{a}_{\omega,\nu,t-1}\right)$ where Log Exports IV$_{\nu t}$ = Log $\sum_{k\in K}$ Exports$_{k\nu t}$ where K denotes the same set of countries as Autor et al. (2013): Australia, Finland, Germany, Japan, New Zealand, Spain, Switzerland, and USA.

Table D.11: Export Shocks - IV = EU Countries

	Dependent Variable: Dummy=1 for Product Introduction							
	(CES)	(CES)	(DGKP)	(DGKP)	(CN)	(CN)	(RCA)	(RCA)
Lagged Advantage	47.600	509.799***	25.540	116.219***	3.720	14.576	-60.616***	-43.784***
	(66.152)	(140.652)	(17.992)	(35.409)	(5.447)	(11.370)	(4.094)	(15.438)
Lagged Advantage	-10.772*	-51.171***	-3.429**	-11.222***	-0.561	-1.506	-3.737***	-38.540
X Exports	(5.650)	(12.167)	(1.522)	(3.016)	(0.465)	(0.983)	(1.088)	(30.896)
Firm-Time FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm-Product FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Prod-Time FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Spec.	OLS	IV	OLS	IV	OLS	IV	OLS	IV
# Obs.	13782	13782	13782	13782	13782	13782	13782	13782
F-Stat		257.25		419.77		304.73		11.75

Results from OLS estimation of (20). Cluster standard errors in parenthesis. Cluster: product-year. ***: significant at 99%, ** at 95%, * at 90%. We instrument Log Exports$_{\nu t}$ $\times \ln\left(1+\tilde{a}_{\omega,\nu,t-1}\right)$ with Log Exports IV$_{\nu t}$ $\times \ln\left(1+\tilde{a}_{\omega,\nu,t-1}\right)$ where Log Exports IV$_{\nu t}$ = Log $\sum_{k\in K}$ Exports$_{k\nu t}$ where K denotes the set of EU countries: Austria, Belgium, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, Netherlands, Portugal, Spain, and Sweden.

D.5 First Stage Regressions

Table D.12: Export Shocks - IV. All Countries (First Stage)

	Dependent Variable: Lagged Advantage X Exports			
	(CES)	(DGKP)	(CN)	(RCA)
Lagged Advantage X Exports IV	0.781***	0.792***	0.769***	0.034***
	(0.003)	(0.003)	(0.004)	(0.002)
Firm-Time FE	Yes	Yes	Yes	Yes
Firm-Product FE	Yes	Yes	Yes	Yes
Prod-Time FE	Yes	Yes	Yes	Yes
# Obs.	13782	13782	13782	13782
R^2	1.00	1.00	1.00	0.93

Results from OLS estimation of the first stage of (20). Cluster standard errors in parenthesis. Cluster: product-year. ***: significant at 99%, ** at 95%, * at 90%. Log Exports IV$_{\nu t}$ = Log $\sum_{k\neq DNK}$ Exports$_{k\nu t}$.

Table D.13: Export Shocks - IV. Autor et al. (2013) Countries (First Stage)

	Dependent Variable: Lagged Advantage X Exports			
	(CES)	(DGKP)	(CN)	(RCA)
Laggend Advantage X Exports IV	0.889***	0.908***	0.872***	0.040***
	(0.004)	(0.004)	(0.006)	(0.003)
Firm-Time FE	Yes	Yes	Yes	Yes
Firm-Product FE	Yes	Yes	Yes	Yes
Prod-Time FE	Yes	Yes	Yes	Yes
# Obs.	13782	13782	13782	13782
R^2	1.00	0.99	1.00	0.93

Results from OLS estimation of the first stage of (20). Cluster standard errors in parenthesis. Cluster: product-year. ***: significant at 99%, ** at 95%, * at 90%. Log Exports IV$_{\nu t}$ = Log $\sum_{k\in K}$ Exports$_{k\nu t}$ where K denotes the same set of countries as Autor et al. (2013): Australia, Finland, Germany, Japan, New Zealand, Spain, Switzerland, and USA.

Table D.14: Export Shocks - IV. EU Countries (First Stage)

	Dependent Variable: Lagged Advantage X Exports			
	(CES)	(DGKP)	(CN)	(RCA)
Lagged Advantage X Exports IV	0.857***	0.871***	0.838***	0.038***
	(0.004)	(0.004)	(0.005)	(0.003)
Firm-Time FE	Yes	Yes	Yes	Yes
Firm-Product FE	Yes	Yes	Yes	Yes
Prod-Time FE	Yes	Yes	Yes	Yes
# Obs.	13782	13782	13782	13782
R^2	1.00	0.99	1.00	0.93

Results from OLS estimation of the first stage of (20). Cluster standard errors in parenthesis. Cluster: product-year. ***: significant at 99%, ** at 95%, * at 90%. Log Exports $\text{IV}_{\nu t}$ = Log $\sum_{k \in K}$ $\text{Exports}_{k\nu t}$ where K denotes denotes the set of EU countries: Austria, Belgium, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, Netherlands, Portugal, Spain, and Sweden.

D.6 Instruments for Revenue and Competition Potential

In this section, we introduce instrumental variables to examine the relationship between Revenue Potential (RP), Competition Potential (CP), and a firm's growth in sales and scope. Following the approach of Berry et al. (1995), we instrument the $RP_{\omega,c,t}$ of a firm ω in cluster c during year t using the average $RP_{\omega,c,t}$ of its competitors. We employ a similar strategy for $CP_{\omega,c,t}$. To ensure sufficient variation in these measures, we exclude the closest three firms from the average calculation for each firm ω.

Formally, the two instruments are computed as follows

$$RP_{\omega,c,t} = \sum_{w \in \tilde{\Omega}_{\omega,c,t}} RP_{w,c,t} / \left(N_{c,t} - 4\right), \tag{D.5}$$

$$CP_{\omega,c,t} = \sum_{w \in \tilde{\Omega}_{\omega,c,t}} CP_{w,c,t} / \left(N_{c,t} - 4\right), \tag{D.6}$$

where $\tilde{\Omega}_{\omega,c,t}$ is the set of firms in cluster-year (c, t) which excludes the 3 closest firms to firm ω and $N_{c,t}$ is the number of firms in the cluster-year.

The results are presented in Tables D.15 and D.16. The relevance condition is satisfied, as indicated by an F-statistic greater than the standard threshold of 10. Moreover, the results are robust to the IV strategy: higher lagged RP leads to a faster growth rate of sales, while higher lagged CP results in a lower growth rate of scope.

Table D.15: Firm Sales Growth

	Dependent Variable: Growth Rate of Sales	
	Constant Markups	Variable Markups
Lagged RP	0.378***	0.334**
	(0.127)	(0.154)
Lagged CP	-0.059	-0.167**
	(0.078)	(0.076)
Cluster-Time FE	Yes	Yes
Firm FE	Yes	Yes
F-Stat	36.59	45.22
# Obs.	8435	8435

Results from IV estimation of (26). Clustered standard errors in parenthesis. Cluster: firm. ***: significant at 99%, ** at 95%, * at 90%.

Table D.16: Firm Scope Growth

	Dependent Variable: Growth Rate of Scope	
	Constant Markups	Variable Markups
Lagged RP	0.047	0.057
	(0.044)	(0.054)
Lagged CP	-0.113***	-0.061**
	(0.030)	(0.029)
Cluster-Time FE	Yes	Yes
Firm FE	Yes	Yes
F-Stat	36.59	45.22
# Obs.	8435	8435

Results from IV estimation of (26). Clustered standard errors in parenthesis. Cluster: firm. ***: significant at 99%, ** at 95%, * at 90%.

D.7 Absolute Advantage transformed in Levels and IHS

In this section, we replicate the baseline results presented in the main text but using transformations corresponding to Levels or Inverse Hyperbolic Sine (IHS) for robustness. Summary statistics for the distances measured from product-to-product and firm-to-product are provided in Table D.17 for Levels and Table D.18 for IHS. The results for product introduction with lagged distance modulating various fixed effects and sub-samples are all qualitatively similar to those above, so we omit them for brevity.

Table D.17: Product Space Summary Statistics (Levels Transformation, Year 2000)

	Product-to-Product		**Firm-to-Product**	
	CES	DGKP	CES	DGKP
Average	0.87	2.74	0.53	1.39
Std. Dev.	0.49	2.16	0.22	0.71
5th Perc.	0.21	0.52	0.22	0.38
10th Perc.	0.30	0.79	0.27	0.59
25th Perc.	0.48	1.00	0.37	0.91
50th Perc.	0.80	1.92	0.50	1.31
75th Perc.	1.11	3.99	0.65	1.76
90th Perc.	1.53	5.93	0.81	2.19
95th Perc.	1.75	6.91	0.91	2.51

The table reports the distribution of product-to-product and firm-to-product distances in the year 2000. The sample comprises all product-to-product and all firm-to-product distances across all clusters. For confidentiality reasons, we divide distances data in 100 bins, compute the average distance within bin, and report these values in the rows with percentiles.

Table D.18: Product Space Summary Statistics (IHS Transformation, Year 2000)

	Product-to-Product		**Firm-to-Product**	
	CES	DGKP	CES	DGKP
Average	0.53	0.97	0.32	0.75
Std. Dev.	0.29	0.40	0.12	0.27
5th Perc.	0.15	0.36	0.15	0.26
10th Perc.	0.21	0.50	0.18	0.40
25th Perc.	0.31	0.72	0.24	0.59
50th Perc.	0.45	0.88	0.31	0.75
75th Perc.	0.68	1.16	0.39	0.90
90th Perc.	0.91	1.43	0.47	1.04
95th Perc.	1.06	1.66	0.52	1.13

The table reports the distribution of product-to-product and firm-to-product distances in the year 2000. The sample comprises all product-to-product and all firm-to-product distances across all clusters. For confidentiality reasons, we divide distances data in 100 bins, compute the average distance within bin, and report these values in the rows with percentiles.

Tables D.19 and D.20 examine the impact of export demand shocks on product introduction, with the respective first stage regressions detailed in Tables D.21 and D.22. The estimates show that the qualitative conclusions are the same as in the baseline specification.

Table D.19: Export Shocks and Product Adoption (Levels Transformation)

	Dependent Variable: Dummy=1 for Product Introduction			
	(CES)	(CES)	(DGKP)	(DGKP)
Lagged Advantage	10.117	59.895***	1.462	11.970*
	(10.153)	(19.114)	(4.096)	(6.356)
Lagged Advantage X Exports	-1.889**	-6.320***	-0.383	-1.271**
	(0.878)	(1.689)	(0.335)	(0.530)
Firm-Time FE	Yes	Yes	Yes	Yes
Firm-Product FE	Yes	Yes	Yes	Yes
Prod-Time FE	Yes	Yes	Yes	Yes
Spec.	OLS	IV	OLS	IV
# Obs.	13782	13782	13782	13782
F-Stat		526.48		595.68

Results from OLS estimation of (20). Clustered standard errors in parenthesis. Cluster: product-year. ***: significant at 99%, ** at 95%, * at 90%. We instrument Log Exports$_{\nu t} \times \tilde{a}_{\omega,\nu,t-1}$ with Exports IV$_{\nu t} \times \tilde{a}_{\omega,\nu,t-1}$ where Exports IV$_{\nu t}$ = Log $\sum_{k \neq DNK}$ Exports$_{k\nu t}$ is the total exports of all countries except Denmark.

Table D.20: Export Shocks and Product Adoption (IHS Transformation)

	Dependent Variable: Dummy=1 for Product Introduction			
	(CES)	(CES)	(DGKP)	(DGKP)
Lagged Advantage	69.421	302.434***	15.288	70.766***
	(47.155)	(84.589)	(13.665)	(23.694)
Lagged Advantage X Exports	-10.349**	-30.700***	-2.203*	-6.968***
	(4.023)	(7.319)	(1.155)	(2.017)
Firm-Time FE	Yes	Yes	Yes	Yes
Firm-Product FE	Yes	Yes	Yes	Yes
Prod-Time FE	Yes	Yes	Yes	Yes
Spec.	OLS	IV	OLS	IV
# Obs.	13782	13782	13782	13782
F-Stat		456.18		577.29

Results from OLS estimation of (20). Clustered standard errors in parenthesis. Cluster: product-year. ***: significant at 99%, ** at 95%, * at 90%. We instrument Log Exports$_{\nu t}$ $\times \ln\left(\tilde{a}_{\omega,\nu,t-1} + \sqrt{1+\tilde{a}^2_{\omega,\nu,t-1}}\right)$ with Exports IV$_{\nu t}$ $\times \ln\left(\tilde{a}_{\omega,\nu,t-1} + \sqrt{1+\tilde{a}^2_{\omega,\nu,t-1}}\right)$ where Exports IV$_{\nu t}$ = Log $\sum_{k \neq DNK}$ Exports$_{k\nu t}$ is the total exports of all countries except Denmark.

Table D.21: Export Shocks - IV = All Countries (Levels, First Stage)

	Dependent Variable: Lagged Advantage X Exports	
	(CES)	(DGKP)
Lagged Advantage X Exports IV	0.768***	0.799***
	(0.004)	(0.003)
Firm-Time FE	Yes	Yes
Firm-Product FE	Yes	Yes
Prod-Time FE	Yes	Yes
# Obs.	13782	13782
R^2	1.00	1.00

Results from OLS estimation of the first stage of (20). Clustered standard errors in parenthesis. Cluster: product-year. ***: significant at 99%, ** at 95%, * at 90%. Exports IV$_{\nu t}$ = Log $\sum_{k \neq DNK}$ Exports$_{k\nu t}$.

Table D.22: Export Shocks - IV = All Countries (IHS, First Stage)

	Dependent Variable: Lagged Advantage X Exports	
	(CES)	(DGKP)
Lagged Advantage X Exports IV	0.782***	0.792***
	(0.003)	(0.003)
Firm-Time FE	Yes	Yes
Firm-Product FE	Yes	Yes
Prod-Time FE	Yes	Yes
# Obs.	13782	13782
R^2	1.00	1.00

Results from OLS estimation of the first stage of (20). Clustered standard errors in parenthesis. Cluster: product-year. ***: significant at 99%, ** at 95%, * at 90%. Exports $\text{IV}_{\nu t} = \text{Log} \sum_{k \neq DNK} \text{Exports}_{k\nu t}$.

Furthermore, Tables D.23 and D.24 investigate how Revenue Potential and Competition Potential influence the growth rates of sales and scope using the Levels and IHS transformations. The tables show that the results are robust across specifications.

Table D.23: Firm Sales and Scope Growth (Levels Transformation)

	Dependent Variable: Growth Rate of Sales	
	Constant Markups	Variable Markups
Lagged RP	0.542***	0.376***
	(0.075)	(0.076)
Lagged CP	-0.079*	-0.180***
	(0.043)	(0.043)
Cluster-Time FE	Yes	Yes
Firm FE	Yes	Yes
R^2	0.16	0.16
# Obs.	8435	8435

	Dependent Variable: Growth Rate of Scope	
	Constant Markups	Variable Markups
Lagged RP	0.028	0.092***
	(0.020)	(0.024)
Lagged CP	-0.135***	-0.050***
	(0.016)	(0.015)
Cluster-Time FE	Yes	Yes
Firm FE	Yes	Yes
R^2	0.14	0.13
# Obs.	8435	8435

Results from OLS estimation of (26). Clustered standard errors in parenthesis. Cluster: firm. ***: significant at 99%, ** at 95%, * at 90%.

Table D.24: Firm Sales and Scope Growth (IHS Transformation)

	Dependent Variable: Growth Rate of Sales	
	Constant Markups	Variable Markups
Lagged RP	0.564***	0.370***
	(0.078)	(0.084)
Lagged CP	-0.075*	-0.166***
	(0.046)	(0.045)
Cluster-Time FE	Yes	Yes
Firm FE	Yes	Yes
R^2	0.16	0.15
# Obs.	8435	8435

	Dependent Variable: Growth Rate of Scope	
	Constant Markups	Variable Markups
Lagged RP	0.007	0.110***
	(0.021)	(0.026)
Lagged CP	-0.125***	-0.049***
	(0.017)	(0.016)
Cluster-Time FE	Yes	Yes
Firm FE	Yes	Yes
R^2	0.14	0.13
# Obs.	8435	8435

Results from OLS estimation of (26). Clustered standard errors in parenthesis. Cluster: firm. ***: significant at 99%, ** at 95%, * at 90%.

Finally, Tables D.25 and D.26 present the outcomes of our counterfactual gains from variety scenario, in which all firms introduce every variety. The results are similar to the baseline specification, as they are for Green Products (not reported).

Table D.25: Potential Gains from Variety by Sector ($\times 100$)

	Constant Markups		Variable Markups	
Cluster	Avg.	Std.	Avg.	Std.
Animals	72.5	21.5	21.0	19.4
Foodstuffs	70.2	20.1	17.4	23.0
Mineral products	77.7	21.7	43.5	17.8
Chemical Products	77.8	22.3	22.8	24.8
Plastics and rubber	77.5	22.8	26.4	23.7
Wood products	81.3	19.4	39.1	29.1
Textiles	70.3	16.1	25.0	25.6
Stone and glass	82.1	24.6	50.9	27.1
Metals	79.3	24.3	29.5	22.5
Machinery and electrical	71.7	23.7	16.8	23.9
Transportation	85.3	19.2	66.0	30.2
Miscellaneous	83.6	16.3	31.2	24.9
Average	**77.4**	**21.0**	**32.5**	**24.3**

Average and standard deviation of the EP for each cluster. We set $\sigma = 5$ in the calculation of the price index for the CES case. For the DGKP case, details on the calculations of the price index are in the main text.

Table D.26: Potential Gains from Variety by Sector (×100)

	Constant Markups		Variable Markups	
Cluster	Avg.	Std.	Avg.	Std.
Animals	71.9	21.5	19.5	18.3
Foodstuff	69.1	19.2	14.9	21.2
Mineral products	73.7	24.7	40.7	16.6
Chemical Products	77.0	21.8	19.8	23.1
Plastics and rubber	75.2	22.9	22.8	22.1
Wood	79.4	18.8	35.9	29.2
Textiles	70.2	15.6	23.1	24.5
Stone and glass	80.7	25.5	47.6	27.9
Metals	78.8	23.9	25.0	20.3
Machinery and electrical	70.7	23.0	14.4	22.1
Transportation	83.0	21.7	63.8	30.3
Miscellaneous	83.0	15.5	26.2	23.6
Average	**76.1**	**21.2**	**29.5**	**23.3**

Average and standard deviation of the EP for each cluster. We set $\sigma = 5$ in the calculation of the price index for the CES case. For the DGKP case, details on the calculations of the price index are in the main text.

E Larger Cluster Analysis

In this section, we replicate the baseline results presented in the main text, using a more aggregate definition of sectors. Specifically, we consider six sectors: Animals/Vegetables/Food (CN 2-digit 01-24), Minerals/Chemicals/Plastics (25-40), Textiles/Footwear (41-43, 50-67), Stone/Metals (68-83), Machinery/Transportation (84-89), and Miscellaneous (44-49, 90-97). This replication uses the same distance transformation as the baseline. Tables E.1 and E.2 present the descriptive statistics for the cluster characteristics. Table E.3 provides summary statistics for the distances measured from product-to-product and firm-to-product. Notice that relative to our baseline results, the average distances increase by little, with the larger increases concentrated in the higher percentiles.

Tables E.4, E.5, and E.6 detail the correlation between the likelihood of product introduction and the rankings based on the firm's product distance. Tables E.7, E.8, and E.9 explore the relationship between the probability of introducing a product and its lagged distance to the firm. Table E.10 examines the impact of export demand shocks on product introduction, with the first stage regression detailed in Table E.11. Furthermore, Tables E.12 and E.13 investigate how Revenue Potential and Competition Potential influence the growth rates of sales and scope. Finally, Tables E.14 and E.15 present the outcomes of the counterfactual in which all firms introduce every variety.

Table E.1: Cluster Descriptive Statistics

	Number of Products					Number of Firms				
Year	Avg.	Std.	Med.	25P.	75P.	Avg.	Std.	Med.	25P.	75P.
2000	207	141	191	84	260	134	64	111	90	196
2001	211	126	189	113	258	150	66	129	115	190
2002	208	141	174	91	260	143	72	129	83	190
2003	200	137	171	85	247	135	68	121	77	183
2004	205	123	175	118	250	144	56	133	128	182
2005	205	124	166	117	249	141	63	127	120	194
2006	211	128	174	124	258	136	60	130	105	192
2007	160	102	128	95	203	87	33	88	78	106
2008	163	107	130	111	208	84	34	88	79	93
2009	163	110	136	108	209	86	37	88	78	100
2010	173	97	143	123	203	100	41	111	87	131
2011	190	106	156	133	210	108	48	129	91	140
2012	195	103	162	135	225	100	40	112	82	133
2013	205	104	169	154	239	104	43	118	85	129
2014	212	108	167	166	267	105	44	117	94	129
2015	214	114	171	156	286	106	43	115	95	139
2016	208	120	160	154	288	100	46	109	80	140
2017	217	119	171	150	290	106	45	115	107	137
2018	219	123	158	151	292	107	45	118	100	140
Average	**198**	**118**	**162**	**125**	**247**	**114**	**50**	**115**	**93**	**150**

In each year, we compute average (Avg.), standard deviation (Std.), median (Med.), and 25th and 75th percentiles (25P. and 75P.) of the number of products (first four columns) and number of firms (last four columns) across clusters. In each year, there are 6 clusters (for 6 sectors defined as groups of CN 2-digit codes). The last row (Average) reports the average of the statistics across years.

Table E.2: Cluster Descriptive Statistics

	Number of Products		Number of Firms	
Cluster	Avg.	Std.	Avg.	Std.
Animals/Vegetables/Food	407	36	158	25
Machinery/Transportation	145	17	109	12
Minerals/Chemicals/Plastics	247	31	107	19
Miscellaneous	84	20	158	66
Stone/Metals	127	29	116	24
Textiles/Footwear	178	33	39	22

In each cluster, we compute average (Avg.) and standard deviation (Std.) of the number of products (first two columns) and number of firms (last two columns) across years.

Table E.3: Product Space Summary Statistics (Baseline Transformation, Year 2000)

	Product-to-Product		**Firm-to-Product**	
	CES	DGKP	CES	DGKP
Average	0.39	0.82	0.25	0.65
Std. Dev.	0.20	0.38	0.09	0.23
5th Perc.	0.12	0.32	0.11	0.29
10th Perc.	0.16	0.41	0.14	0.37
25th Perc.	0.24	0.59	0.18	0.50
50th Perc.	0.35	0.75	0.24	0.62
75th Perc.	0.48	0.99	0.29	0.77
90th Perc.	0.69	1.23	0.35	0.94
95th Perc.	0.74	1.42	0.40	1.05

The table reports the distribution of product-to-product and firm-to-product distances in the year 2000. The sample comprises all product-to-product and all firm-to-product distances across all clusters. The product space is estimated using marginal costs in logs. For confidentiality reasons, we divide distances data in 100 bins, compute the average distance within bin, and report these values in the rows with percentiles.

Table E.4: Product Introduction and Product Rankings

	Dependent Variable: Dummy=1 for Product Introduction			
	(CES)	(DGKP)	(CN)	(RCA)
Lagged Rank (=1 closest)	-0.004***	-0.004***	-0.004***	-0.006***
	(0.000)	(0.000)	(0.000)	(0.000)
Product-Time FE	Yes	Yes	Yes	Yes
Firm-Time FE	Yes	Yes	Yes	Yes
R^2	0.05	0.05	0.05	0.05
# Obs.	1332970	1332970	1332970	1332970

Results from OLS estimation of (D.3). Clustered standard errors in parenthesis. Cluster: product-year. ***: significant at 99%, ** at 95%, * at 90%.

Table E.5: Product Introduction and Product Rankings

	Dependent Variable: Dummy=1 for Product Introduction		
	(CES)	(DGKP)	(RCA)
Lagged Rank (=1 closest)	-0.002***	-0.002***	-0.004***
	(0.000)	(0.000)	(0.000)
Product-Time FE	Yes	Yes	Yes
Firm-Time FE	Yes	Yes	Yes
CN Distance FE	Yes	Yes	Yes
R^2	0.06	0.06	0.06
# Obs.	1332970	1332970	1332970

Results from OLS estimation of (D.3). Clustered standard errors in parenthesis. Cluster: product-year. ***: significant at 99%, ** at 95%, * at 90%.

Table E.6: Product Introduction and Product Rankings

	Dependent Variable: Dummy=1 for Product Introduction		
	(CES)	(DGKP)	(CN)
Lagged Rank (=1 closest)	-0.001***	-0.001***	-0.002***
	(0.000)	(0.000)	(0.000)
Product-Time FE	Yes	Yes	Yes
Firm-Time FE	Yes	Yes	Yes
Lagged Relatedness =	0	0	0
R^2	0.06	0.06	0.06
# Obs.	1173802	1173802	1173802

Results from OLS estimation of (D.3). Clustered standard errors in parenthesis. Cluster: product-year. ***: significant at 99%, ** at 95%, * at 90%.

Table E.7: Product Introduction and Advantage

	Dependent Variable: Dummy=1 for Product Introduction			
	(CES)	(DGKP)	(CN)	(RCA)
Lagged Advantage	-6.166***	-2.510***	-2.555***	-23.188***
	(0.468)	(0.195)	(0.103)	(1.457)
Product-Time FE	Yes	Yes	Yes	Yes
Firm-Time FE	Yes	Yes	Yes	Yes
R^2	0.05	0.05	0.05	0.05
# Obs.	1332970	1332970	1332970	1332970

Results from OLS estimation of (D.3). Clustered standard errors in parenthesis. Cluster: product-year. ***: significant at 99%, ** at 95%, * at 90%.

Table E.8: Product Introduction and Advantage

	Dependent Variable: Dummy=1 for Product Introduction		
	(CES)	(DGKP)	(RCA)
Lagged Advantage	-2.886***	-1.196***	-15.649***
	(0.453)	(0.192)	(1.358)
Product-Time FE	Yes	Yes	Yes
Firm-Time FE	Yes	Yes	Yes
CN Distance FE	Yes	Yes	Yes
R^2	0.06	0.06	0.06
# Obs.	1332970	1332970	1332970

Results from OLS estimation of (D.3). Clustered standard errors in parenthesis. Cluster: product-year. ***: significant at 99%, ** at 95%, * at 90%.

Table E.9: Product Introduction and Advantage

	Dependent Variable: Dummy=1 for Product Introduction		
	(CES)	(DGKP)	(CN)
Lagged Advantage	-1.700***	-0.650***	-1.712***
	(0.343)	(0.150)	(0.107)
Product-Time FE	Yes	Yes	Yes
Firm-Time FE	Yes	Yes	Yes
Relatedness =	0	0	0
R^2	0.06	0.06	0.06
# Obs.	1173802	1173802	1173802

Results from OLS estimation of (D.3). Clustered standard errors in parenthesis. Cluster: product-year. ***: significant at 99%, ** at 95%, * at 90%.

Table E.10: Export Shocks and Product Adoption

	Dependent Variable: Dummy=1 for Product Introduction							
	(CES)	(CES)	(DGKP)	(DGKP)	(CN)	(CN)	(RCA)	(RCA)
Lagged Advantage	-53.495	430.422	48.417*	125.458**	6.846	17.724	-112.107***	-26.161
	(114.169)	(266.203)	(29.045)	(55.870)	(8.115)	(16.099)	(6.531)	(145.616)
Lagged Advantage	-6.754	-48.850**	-5.755**	-12.347***	-1.022	-1.963	-5.474***	-173.642
X Exports	(9.801)	(23.082)	(2.455)	(4.759)	(0.687)	(1.387)	(1.741)	(284.484)
Firm-Time FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm-Product FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Prod-Time FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
# Obs.	16266	16266	16266	16266	16266	16266	16266	16266
F-Stat		1882.60		3624.34		2964.54		0.68

Results from OLS estimation of (20). Clustered standard errors in parenthesis. Cluster: product-year. ***: significant at 99%, ** at 95%, * at 90%. We instrument $\text{Exports}_{\nu t} \times \ln\left(1+\tilde{a}_{\omega,\nu,t-1}\right)$ with $\text{Exports IV}_{\nu t} \times \ln\left(1+\tilde{a}_{\omega,\nu,t-1}\right)$ where $\text{Exports IV}_{\nu t} = \text{Log} \sum_{k \neq DNK} \text{Exports}_{k\nu t}$ is the total exports of all countries except Denmark.

Table E.11: Export Shocks - IV = All Countries (First Stage)

	Dependent Variable: Lagged Advantage X Exports			
	(CES)	(DGKP)	(CN)	(RCA)
Lagged Advantage X Exports IV	0.788***	0.797***	0.777***	0.035***
	(0.003)	(0.003)	(0.004)	(0.002)
Firm-Time FE	Yes	Yes	Yes	Yes
Firm-Product FE	Yes	Yes	Yes	Yes
Prod-Time FE	Yes	Yes	Yes	Yes
# Obs.	16266	16266	16266	16266
R^2	1.00	1.00	1.00	0.92

Results from OLS estimation of the first stage of (20). Clustered standard errors in parenthesis. Cluster: product-year. ***: significant at 99%, ** at 95%, * at 90%. $\text{Exports IV}_{\nu t} = \text{Log} \sum_{k \neq DNK} \text{Exports}_{k\nu t}$.

Table E.12: Firm Sales Growth

	Dependent Variable: Growth Rate of Sales	
	Constant Markups	Variable Markups
Lagged RP	0.452***	0.286***
	(0.071)	(0.090)
Lagged CP	-0.058	-0.134***
	(0.043)	(0.044)
Cluster-Time FE	Yes	Yes
Firm FE	Yes	Yes
R^2	0.14	0.14
# Obs.	9552	9552

Results from OLS estimation of (26). Clustered standard errors in parenthesis. Cluster: firm. ***: significant at 99%, ** at 95%, * at 90%.

Table E.13: Firm Scope Growth

	Dependent Variable: Growth Rate of Scope	
	Constant Markups	Variable Markups
Lagged RP	-0.006	0.050*
	(0.023)	(0.028)
Lagged CP	-0.128***	-0.044***
	(0.015)	(0.016)
Cluster-Time FE	Yes	Yes
Firm FE	Yes	Yes
R^2	0.12	0.11
# Obs.	9552	9552

Results from OLS estimation of (26). Clustered standard errors in parenthesis. Cluster: firm. ***: significant at 99%, ** at 95%, * at 90%.

Table E.14: Potential Gains from Variety by Sector ($\times 100$)

	Constant Markups		Variable Markups	
Cluster	Avg.	Std.	Avg.	Std.
Animals/Vegetables/Food	63.0	19.9	8.0	16.5
Machinery/Transportation	69.3	22.4	13.6	21.9
Minerals/Chemicals/Plastics	63.0	31.1	14.6	14.9
Miscellaneous	81.5	18.2	17.4	22.4
Stone/Metal	76.9	25.0	16.3	16.9
Textiles/Footwear	70.2	15.5	23.0	24.4
Average	**70.7**	**22.0**	**15.5**	**19.5**

Average and standard deviation of the EP for each cluster. We set $\sigma = 5$ in the calculation of the price index for the CES case. For the DGKP case, details on the calculations of the price index are in the main text.

Table E.15: Potential Gains from Variety for Green Products ($\times 100$)

	Constant Markups		Variable Markups	
	Avg.	Std.	Avg.	Std.
Green Products	76.47	23.19	15.17	17.59

Average and standard deviation of the EP for green products. We set $\sigma = 5$ in the calculation of the price index for the CES case. For the DGKP case, details on the calculations of the price index are in the main text.