Firms in Product Space: Growth, Adaptation, and Competition

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

March 14, 2023

Preliminary Draft

Abstract

Using multi-product production patterns within and across firms, we recover a continuous cost based distance between pairs of products and firms. Firms closer to potential products grow faster in sales and scope. Higher interfirm proximity slows sales growth but increases scope growth, core product focus and predicts mergers. Product distance implies a product adoption path. Each extra rank of product distance decreases the frequency of adoption by .7 percent relative to the base rate and 5 percent of the closest products explain 15 per cent of adoptions. When export demand for unproduced products induces domestic adoption, firms choose closer products.

JEL Codes: D2, L1, L23, L25.
Keywords: Multiproduct firms, firm capabilities, product classification, product space, growth paths.

Acknowledgments. We thank Nick de Roos for early suggestions and Lei Yu for excellent research assistance. We are grateful for comments from Catherine Thomas, Kirill Borusyak, Andreas Hafler, Wanyu Chung and participants at the 1st SICKCL Conference, Economics of Global Interactions, King's College London Research Day, ETSG, LSE Trade Workshop, Kent Macro Workshop, Nottingham, Zurich ETH, Paris Trade & Macro Workshop and Royal Economic Society.

[a]Aarhus University.
[b]Kings College London, CEP/LSE, and CEPR. Corresponding Author: john.morrow@kcl.ac.uk.
[c]University of Sydney.

1 Introduction

Recent work has shed light on which products or inputs firms sell or buy together and how these decisions affect growth. These studies look *within* the firm and rely on product classifications as measures of closeness, taken as given depending on the statistical source. Such classifications may be based on consumption substitution patterns, common production patterns, administrative objectives, historical accident and likely some mixture of these rationales (Jacobs and ONeill, 2003). For instance, is popcorn *grown* or *manufactured*?[1] Does popcorn compete with corn or Pringles?[2] These distinctions matter for policy. In the case of a merger between paper companies Torras and Sarrio, the European commission concluded that while coated and uncoated paper are not easily substitutable on the demand side, they are on the supply side and the relevant definition of market concentration depends on which firms can easily supply the product (Commission of the European Communities, 1992). The fact that such classifications are discrete and hierarchical also implies comparisons between arbitrary products can be difficult once assigned to distinct classification branches.

This paper instead develops a new *continuous classification* of product and firm distances based on empirical co-production patterns *within* and *across* firms. Distances are based on the intensive margin of co-production and the implied distances between chains of products that are produced in common across firms. Proximity of a firm to unproduced potential products influences adoption, sales growth and scope growth. These forces are tempered by congestion when products outside the firm are close together, or firms are close together, suggesting this continuous classification captures both technological similarity and business stealing by multi-product firms. In fact, firms close to one another are more likely to acquire one another, the ultimate form of business stealing.

While the prevailing trend of the literature has been to delve deeper into the inner workings of the firm controlling for ever more detailed fixed effects and precise policy shocks to examine product mix and growth, our approach is to construct an environment in which all products and firms have coordinates and relative distances. These distances capture the

[1] Table 14 highlights differences across even the Harmonized System, Standard Industrial Classification and North American Industry Classification System. Even the organizing principles of product 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 where producing units using identical or very similar production processes are grouped." (Girard and Trau, 2004)

[2] Whose potato content is only 42% and shape "is not found in nature" exempting Pringles from VAT for potato crisps and potato-derived snacks.

observed transitive production relationships *across* all firms. This classification based on co-production and sales takes a data driven, supply side stance on product relatedness which may reveal deeper and clearer patterns in firm behavior.

To understand our framework, consider a set of products, each characterized by a technological distance from each other, represented in a high dimensional space. Each firm's costs of production increase with the distance to each product. Productivity, product distance and fixed costs then determine the extensive margin of the firm under monopolistic competition.[3] Identifying these distances is a problem the opposite from triangulation, finding the position of cell phone users (products) by their distance from cell phone towers (firms).[4] In the context of this example, we know the distance from each product to each firm from the implied marginal cost of the product, but don't know the distances between firms or products they don't produce. We implement a procedure using the upper and lower bounds of the pairwise distances between products to infer their relative locations. Finally we locate the positions of firms to be consistent with observed revenues accounting for market conditions.

Notably, our method relies on co-production *within and across* firms. Sales of these firms do not only contain information about these firms' production capabilities but also reflect technological similarity/dissimilarity between these products. Some pairs of products are often produced together, while others are almost never produced by the same firm, as shown by a literature on co-production (for instance, Bernard et al. (2010); Goldberg et al. (2010)). It suggests that 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., 2019). Such analyses are inherently discrete, paralleling existing discrete product classifications. More could be understood with continuous classification systems that integrate relative production levels *within* firms, and transitive production patterns *across* firms, e.g. when firm A makes products 1 and 2, while firm B makes products 2 and 3, this should reveal information regarding goods 1 and 3. We construct such a continuous classification which spans every largest co-produced set of products in two digit sectors across all years, a data defined classification we call a *cluster*.[5] Then we identify the location of

[3]Note that unlike productivity as in a single product setting of Melitz (2003) or multi-product setting of Mayer et al. (2014a), the firm's position relative to the nexus is non-hierarchical.

[4]In computer science, analysis of self-positioning networks leads to a similar problem. For example, indoor positioning algorithms use the information on signal strength from multiple Wi-Fi routers with a priori unknown locations to identify the location of a cell phone user in a building with a weak GPS signal.

[5]While we define this in more detail in the Data Section, this spans over 90% of the products in our data. We are progressively moving to larger classifications as we solve computational issues with the method.

every firm relative to products and each other.

This setting allows for a rich set of testable predictions and findings, which we apply to detailed firm production data from Denmark, including:

- Products closest to a firm are the most likely to be adopted and firms closer to potential new products grow faster in sales and scope.
- In higher sales clusters, firms grow *up* with higher sales and less scope; in higher entrant clusters, firms grow *out* with lower sales and more scope.
- Firms close to each other grow more slowly but increase scope more quickly, and are more likely to merge.
- Product distance from a firm predicts a product adoption path better than the cluster sales ranking and one third of product adoptions are in the Top 10 closest products. Each extra rank of distance a product is from a firm decreases the frequency of adoption by one per cent relative to the base rate.

Literature Review

In the last decade, there has been an explosion of research on multi-product firms, especially in the context of international trade.[6] For the typical model of this literature, a firm is a collection of products, which may be linked by supply or demand linkages.[7] A product is defined as a variety produced by one firm and is characterized by some marginal cost of production and or by some demand shifter (Eckel and Neary, 2010; Bernard et al., 2011). However, which particular product a firm produces is essentially neglected: in these models, products differ in their sales and this difference is driven by differences in costs or demand. Whether a firm is producing milk and cheese or milk and silk is irrelevant. By recovering distances in a product space and positioning firms in that space, *which* products a firm produces relative to *all other firms* matters for sales and scope growth.

While we are concerned with manufacturing firms connected in a product space, a term much popularized by the groundbreaking work of authors such as Hausmann et al. (2007),

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

[7]Supply linkages include flexible manufacturing, economies and diseconomies of scope, and the presence of core and non-core products (Eckel and Neary, 2010; Nocke and Yeaple, 2014; Mayer et al., 2014b; Eckel et al., 2015; Arkolakis et al., 2021; Macedoni and Xu, 2022). Demand 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).

there is also an innovation literature more concerned with new technologies which considers how firms innovate to gain market share. Escolar et al. (2020) understand firm performance using the topology of firm networks across patent areas, instead of products.

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. Bishop et al. (2022) match UK business website data to publically reported SIC codes, using text analysis to show that four digit SIC codes mask considerable heterogeniety in firm activities. Kogan et al. (2021) categorize worker's technology exposure from patent documents and the Dictionary of Occupational Titles, finding exposure displaces both high and low skill workers.

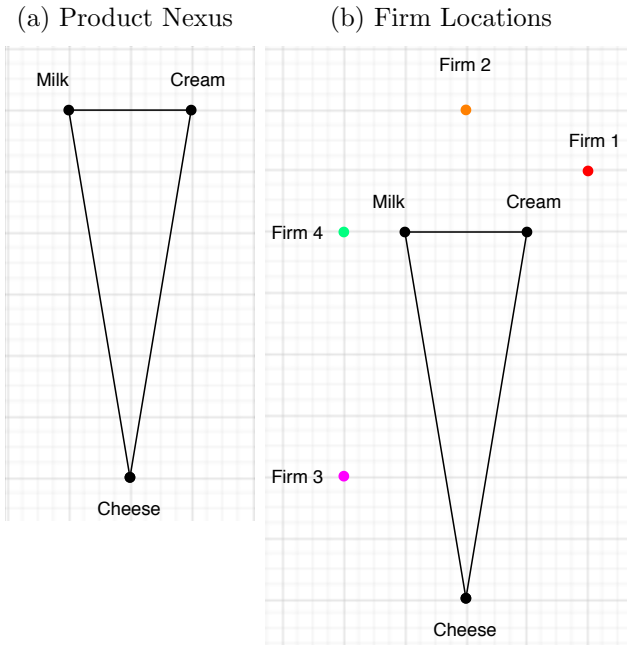
This paper proceeds as follows. A motivational section preceeds a theoretical development of how to construct the *product space*. Section 4 then details the data and construction of the space with summary statistics of its properties. Section 5 estimates various relationships of growth in sales and scope to the relative locations of products and firms. Section 6 examines how the distance rank of a product from a firm can explain adoption. Section 7 concludes.

2 Motivational Example

Consider an example with three products: milk, cream and cheese[8] so the location of each firm then can be represented in a plane. As shown, milk and cream have more similar production technologies, while cheese production is technologically distinct from either. Figure 1a represents this in our framework with a small technological distance between milk and cream and a large distance of these away from cheese.

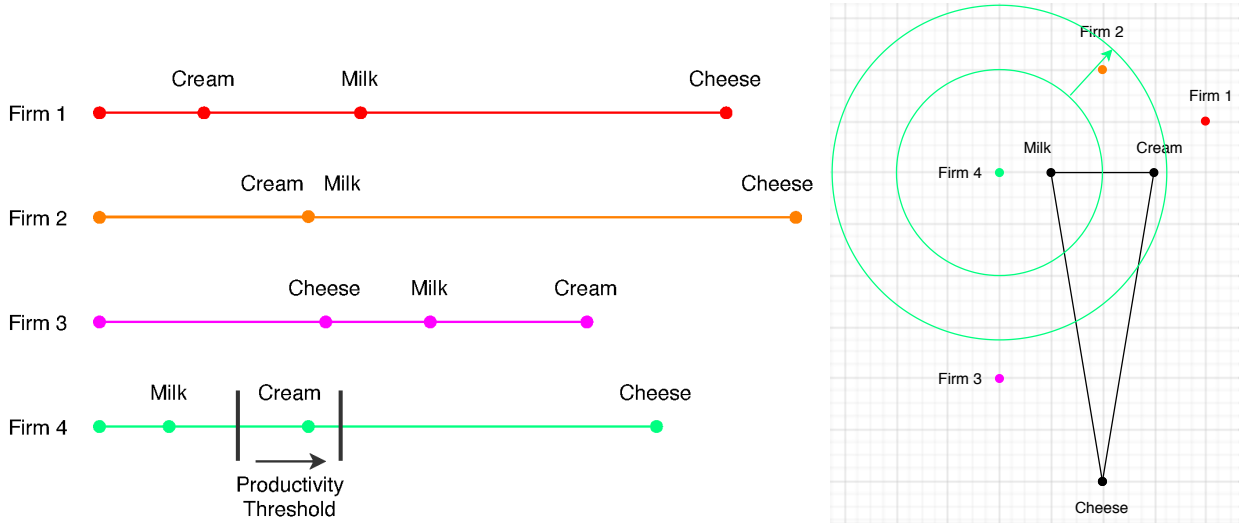
[8]Vegan options are available upon request.

Figure 1: Technological Nexus



Every firm is characterized by its position relative to products as in Figure 1b. Here Firm 1 is relatively good at cream production, less good at milk production and the least good at cheese production. Firm 2 is equally good in the production of milk and cream, but not as good at cheese production. Firm 3 is good in the production of cheese, almost as good in the production of milk, and the worst in the production of cream. These relationships are represented in a more conventional way in Figure 3a.

Figure 2: Firm and Product Locations

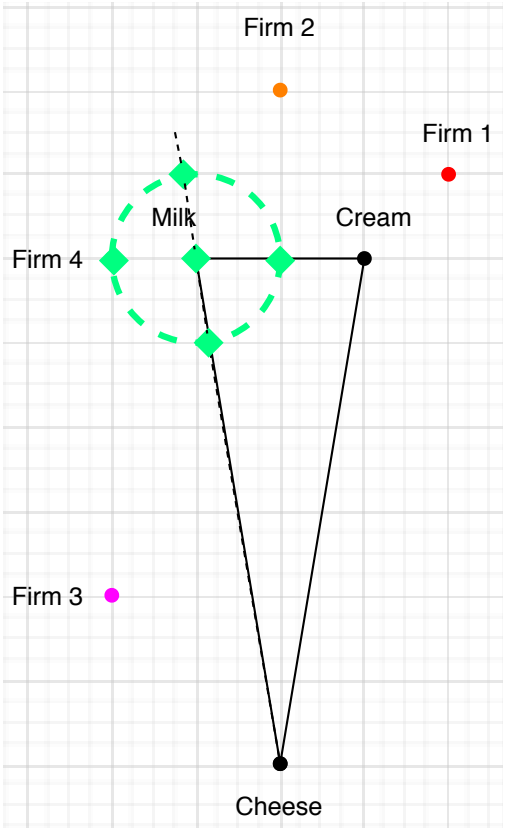


(a) Core Competencies (b) Product Adoption

Productivity of Firm 4 is represented in Figure 3b, where a circle represents a distance beyond which a firm would not make a product. In this example, Firm 4 only produces milk at first, but after a positive productivity shock it might expand to produce cream as seen in the last row of Figure 3a. This also suggests that Firm 4 needs a huge improvement in productivity to start producing cheese. Thus, the products relative to each firm may help us understand what order and under what types of shocks firms adopt or drop products.

Once the locations of products are fixed, we can also locate the firms in 'product space'. In Figure 3, possible locations of Firm 4 are illustrated on Figure 3 as a dotted green circle, where the radius of this circle is the known distance of Firm 4 from the milk point of the nexus. Green rhomboids on this circle represent the highest and the lowest possible distances from Firm 4 to cream and cheese. Furthermore, the center of this circle is the middle of both of these intervals and, in the absence, of any additional information, is a reasonable assumption for the location of Firm 4. Note that here we considered the scenario of a single product firm. If Firm 4 produces both cream and milk, a reasonable assumption is that Firm 4 is located on the line between milk and cream points. We also stipulate that the distances between Firm 4 and cream/milk are in proportion to its implied marginal cost ratio as implied by revenues, thereby fixing a location. In a similar fashion, we mathematically locate each firm as close as possible to its product portfolio while respecting the implied marginal cost ratio across products.

Figure 3: Single Product Firm Location



The next Section formalizes this concept in order to quantify these relationships.

3 Firm and Product Space

This section models firms and products co-located in space to establish a concept of distance between products. Given any sufficiently rich demand and competition structure, the revenue shares of a multi-product firm can be mapped to its marginal costs of production. We assume these costs are inversely related to the technological distances between products. While each firm makes a subset of products, the collection of all firm-product observations can be used to reveal information regarding the distance of products from one another, so long as there are chains of co-production observed between any two products. We define each set of products connected this way over all years as a *cluster*. In principle, any framework which allows us to derive this relationship is suitable; here we follow the standard setting of the literature, CES utility and monopolistic competition. The result is a continuous product classification that

explains heterogeneous firm productivities across products in terms of each firm's location in relation to products in a high dimensional product space.

3.1 Consumers and Firms

Consumer preferences for quantities of goods $q(\nu,\omega)$ across a continuum of goods varieties ν, supplied by a continuum firms ω, are given by

$$Q_\nu(q) = \left[\int_{\Omega_\nu} q(\nu,\omega)^{\frac{\sigma-1}{\sigma}} d\Omega(\omega)\right]^{\frac{\sigma}{\sigma-1}} \quad \text{with} \quad \sigma > 1,$$

where Ω_ν is the measure of varieties supplied ν. Preferences over varieties are given by

$$U(q) = \int \alpha_{\nu,t} \ln Q_\nu d\nu \quad \text{with} \quad \int_\nu \alpha_{\nu,t} d\nu = 1$$

There is a unit mass of consumers with combined income I. Variety ν is supplied by firm ω at price $p(\nu,\omega)$, and variety-firm pairs are distributed with measure Ψ in equilibrium. Consumers maximize utility through

$$\max_{q(\nu,\omega)} U(q) \quad \text{subject to} \quad I = \int \int_{\Omega_\nu} p(\nu,\omega) q(\nu,\omega) d\Psi(\nu,\omega).$$

Standard sufficient first order conditions imply that for the price index

$$P_\nu \equiv \left(\int_{\Omega_\nu} p(\nu,\omega)^{1-\sigma} d\Phi(\nu,\omega)\right)^{1/(1-\sigma)},$$

where α_ν is the budget share on variety ν and for total revenues of variety ν, R_ν we have

$$\alpha_\nu I = P_\nu Q_\nu = R_\nu.$$

This implies demand for variety ν of firm ω of

$$q(\nu,\omega) = (\alpha_\nu I)^\sigma Q_\nu^{1-\sigma} / p(\nu,\omega)^\sigma = \alpha_\nu I / P_\nu^{1-\sigma} p(\nu,\omega)^\sigma.$$

Each firm ω faces a fixed cost of production for product ν, f_ν, and marginal cost $c(\nu,\omega)$. Profit maximization yields constant markups of $p(\nu,\omega) = \frac{\sigma}{\sigma-1} c(\nu,\omega)$, so profits and costs

attributable to variety ν for firm ω selling to consumers are

$$\pi(\nu,\omega) = \sigma^{-\sigma}(\sigma-1)^{\sigma-1} R_\nu / P_\nu^{1-\sigma} c(\nu,\omega)^{\sigma-1} - f_\nu,$$
$$c(\nu,\omega) = \frac{\sigma-1}{\sigma} P_\nu \left(\frac{p(\nu,\omega)\, q(\nu,\omega)}{R_\nu} \right)^{1/(1-\sigma)}.$$

Finally, the measure of firms providing variety ν, is positive for firms with sufficently low costs

$$c(\nu,\omega) \leq \sigma^{\frac{\sigma}{1-\sigma}} (\sigma-1) (\alpha_\nu I)^{\frac{1}{\sigma-1}} P_\nu / f_\nu^{\frac{1}{\sigma-1}}. \tag{1}$$

From Equation (1), it's clear that more firms will produce variety ν incomes increase or there is less competition as measured by the price index P_ν. Now we turn to how the technological distances of products from firms determine the distribution of marginal costs and therefore the extensive margin of firms.

3.2 Technological Distance

Each of N products is represented by a location in N-dimensional technological space, as is each firm ω. The relative location of a firm to products determines its cost structure. The location of a variety ν is $\ell(\nu) = (\ell_1(\nu), ..., \ell_N(\nu))$ and the location of a firm is $\ell(\omega) = (\ell_1(\omega), ..., \ell_N(\omega))$. The technological distance between any pair of products or firms is given by the Euclidean distance, $\|\ell(\nu) - \ell(\omega)\| = \left(\sum_{i=1}^N (\ell_i(\nu) - \ell_i(\omega))^2 \right)^{1/2}$.[9] We use co-produced product distances to a firm to derive the distances between products. Since we won't be able to assign each firm's location until the product locations are fixed *using production information across all firms*, we assume firm locations are observed with additive location error $\varepsilon(\omega)$, so their location is $\ell(\omega) + \varepsilon(\omega)$. Once we have constructed the locations of products, we will fix firm locations. A firm ω's marginal cost in production of variety ν, $c(\nu,\omega)$, is equal to the distance between the firm's technological location and variety's location so that[10]

$$c_\varepsilon(\nu,\omega) \equiv \|\ell(\nu) - \ell(\omega) - \varepsilon(\omega)\|. \tag{2}$$

[9] This could be any norm, for instance the class of L^p norms could be chosen for goodness of fit.

[10] Similarly, we define the actual marginal cost of production of variety ν by firm ω without location error as $c(\nu,\omega) = \|\ell(\nu) - \ell(\omega)\|$.

We define p_ε as the constant markup over c_ε and $P_{\varepsilon\nu}$ as the corresponding price index and similarly all other variables above with firm location error using a ε subscript. Given information on prices, variety budget shares and an assumed value for σ then we observe the cost structures $c_\varepsilon(\nu, \omega)$.

3.3 Bounding Product Distances

If firm-product data were complete, i.e. if all firms produced all products in no matter what quantity, it would be possible to write down a system of equations where the distances from each firm to each nexus nodes are known. No population of firms is likely to exhibit this property and any subsample would suffer from substantial selection issues. The production patterns of multi-product firms therefore exhibits a latent variable problem. However, appealing to an idea inspired by the principle of revealed preference, we propose an alternative procedure based on the triangle inequality which allows us to evaluate upper and lower bounds of distances if there is at least one firm producing both of two goods. This idea can be expanded transitively: if Firm 1 makes products A and B and Firm 2 makes products B and C, we can also infer bounds on the distance between products A and C.[11]

3.3.1 An Example

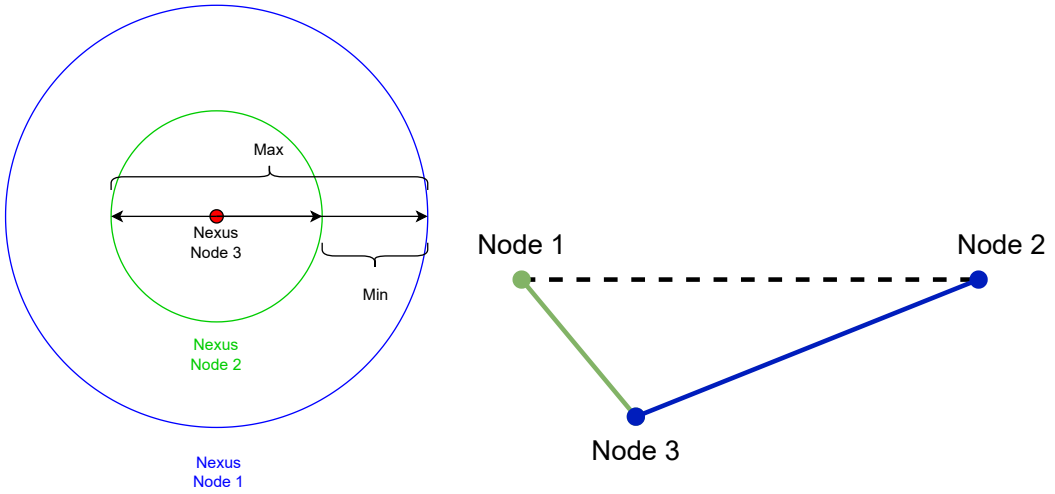
While production patterns in a multi-product firm does not tell us exactly how far apart any two products it produces are, it does provide an upper and lower bound of this distance as shown in Figure 5a. This is easy to see from the triangle inequality. Given two products ν and ν' produced by a firm ω and locations $\ell(\cdot)$, we have

$$\underbrace{\left|\|\ell(\nu) - \ell(\omega)\| - \|\ell(\nu') - \ell(\omega)\|\right|}_{\text{Minimum Distance}} \leq \|\ell(\nu) - \ell(\nu')\| \leq \underbrace{\|\ell(\nu) - \ell(\omega)\| + \|\ell(\nu') - \ell(\omega)\|}_{\text{Maximum Distance}} \quad (3)$$

and one can think about a two product firm as a dot with two concentric circles, where circles represent the potential locations of both products. It is clear then that the shortest possible distance between these products is equal to the difference of two radii and the longest to their sum as illustrated in Figure 5b.

[11] We take a bounding approach since depending on the underlying assumptions of a model, it's unclear what the joint distributions of production within a multi-product firm. What may appear as a 'stylized fact' of rational firm behavior may simply be generated by random processes often assumed in the literature, see Sheveleva (2019) in the case of multi-product firms and Bernard and Zi (2023) in the case of firm-to-firm networks.

Figure 4: Co-produced Product Distances



(a) Max and Min Product Distance

(b) Triangle Inequality Distance Bounds

For each pair of co-produced products we therefore have maximum and minimum distances. If there are enough firms producing both products, these boundaries may be narrow, but even one firm can give us narrow bounds if any firm ω is close to one product ν' (a small radius in Figure 5a) since in Equation (3) as $\|\ell(\nu') - \ell(\omega)\|$ approaches zero, the inequalities on both sides become tight.

We then refine these distances by combining bounds across firms. Formally, define the minimum of the upper bound distances for products ν and ν', $\overline{d}(\nu, \nu')$ by

$$\overline{d}(\nu, \nu') \equiv \min_{\omega} \{\|\ell(\nu) - \ell(\omega)\| + \|\ell(\nu') - \ell(\omega)\| \text{ given } r(\nu, \omega) > 0 \text{ and } r(\nu', \omega) > 0\},$$

and similarly the maximum of the lower bound distances for products ν and ν', $\underline{d}(\nu, \nu')$ by

$$\underline{d}(\nu, \nu') \equiv \max_{\omega} \{|\|\ell(\nu) - \ell(\omega)\| - \|\ell(\nu') - \ell(\omega)\|| \text{ given } r(\nu, \omega) > 0 \text{ and } r(\nu', \omega) > 0\}.$$

We further refine distances along chains of multi-product firms that produce a common product. Suppose ν'' is produced by both a firm ω who produces ν and by a firm ω' that produces ν'. From the triangle inequality,

$$\begin{aligned}\|\ell(\nu) - \ell(\nu')\| &\leq \|\ell(\nu) - \boldsymbol{\ell(\omega)}\| + \|\boldsymbol{\ell(\omega)} - \boldsymbol{\ell(\omega')}\| + \|\boldsymbol{\ell(\omega')} - \ell(\nu')\| \\ &\leq \|\ell(\nu) - \ell(\omega)\| + \|\ell(\omega') - \ell(\nu')\| + \|\ell(\omega) - \boldsymbol{\ell(\nu'')}\| + \|\boldsymbol{\ell(\nu'')} - \ell(\omega')\|\end{aligned}$$

and the distances in the second line are known. Taking all possible such combinations further refines the upper and lower bounds for distance, which label $\widetilde{d}(\nu,\nu')$ and $d(\nu,\nu')$.[12]In the instance that there are groups of products for which there is no such chain of co-production (teddy bears and nuclear weapons) then there is no basis in our method for assessing distance between these products as by definition they are in different clusters. Within each cluster, we then *define* the distance $d(\nu,\nu')$ as the average of $\widetilde{d}(\nu,\nu')$ and $d(\nu,\nu')$,

$$d(\nu,\nu') \equiv \left(\widetilde{d}(\nu,\nu') + d(\nu,\nu')\right)/2.$$

3.4 Assigning Product and Firm Locations

First, we locate products. Numbering products $\{\nu_1,\ldots,\nu_N\}$, we define the product vectors $\{\ell(\nu_1),\ldots,\ell(\nu_N)\}$ recursively to preserve the distances between products, so that $\|\ell(\nu_i)-\ell(\nu_j)\| = d(\nu_i,\nu_j)$ for every pair ν_i and ν_j. Each vector $\ell(\nu_i)$ has every coordinate k, $\ell_k(\nu_i)=0$ for $k\geq i$, so for instance $\ell(\nu_1)=0$. For products ν_i with $i>1$, $\ell(\nu_i)$ has $i-1$ unknown coordinates which must satisfy the system of $i-1$ equations

$$d(\nu_i,\nu_k) = \|\ell(\nu_i)-\ell(\nu_k)\|, \qquad \forall k<i$$

and any solution will satisfy the product distances constructed above. Finally, we use these product coordinates to fix the locations of firms as above.[13]

Second, we locat firms in the product space. The assumptions above imply each firm has observed revenues for variety ν of

$$r_\varepsilon(\nu,\omega) = \left(\frac{\sigma}{\sigma-1}\right)^{\sigma-1}\frac{\alpha_\nu I}{P_{\varepsilon\nu}^{1-\sigma}c(\nu,\omega)^{\sigma-1}} = \left(\frac{\sigma}{\sigma-1}\right)^{\sigma-1}\frac{R_{\varepsilon\nu}}{P_{\varepsilon\nu}^{1-\sigma}c(\nu,\omega)^{\sigma-1}}. \tag{4}$$

We define each firm's core variety $\overline{\nu}(\omega)$ as the variety with the highest revenues. Since the

[12]If one models noise explicitly, it's also the case that such 'chained bounds' introduce ever larger noise terms, limiting their usefulness while becoming computationally expensive. In particular for products ν and ν' produced by firm ω with location noise term $\varepsilon(\omega)$ we have

$$\begin{aligned}\|\ell(\nu)-\ell(\nu')\| &\leq \|\ell(\nu)-\ell(\omega)\| + \|\ell(\nu')-\ell(\omega)\| + 2\|\varepsilon(\omega)\| \\ &\geq \max\{\|\ell(\nu)-\ell(\omega)\|, \|\ell(\nu')-\ell(\omega)\|\} - \|\varepsilon(\omega)\|\end{aligned}$$

so as the magnitude of noise increases or comparisons are chained, the bounds become less informative.

[13]Single product firms will have no part in our analysis besides being used in price indexes of competition, both because they may differ in substantial ways from multi-product firms, but also because our fixed effects to control for omitted variables will exclude them from our analysis.

cost ratio of each variety relative to the core variety fixes distance ratios, using Equations (2) and (4) and defining this cost ratio as $\rho_\varepsilon(\nu,\omega)$ we have

$$\rho_\varepsilon(\nu,\omega) \equiv \frac{c_\varepsilon(\nu,\omega)}{c_\varepsilon(\overline{\nu}(\omega),\omega)} = \frac{\|\ell(\nu)-\ell(\omega)-\varepsilon(\omega)\|}{\|\ell(\overline{\nu}(\omega))-\ell(\omega)-\varepsilon(\omega)\|} = \frac{P_{\varepsilon\nu}}{P_{\varepsilon\overline{\nu}}}\left(\frac{r_\varepsilon(\overline{\nu}(\omega),\omega)/R_{\varepsilon\overline{\nu}}}{r_\varepsilon(\nu,\omega)/R_{\varepsilon\nu}}\right)^{\frac{1}{\sigma-1}}. \quad (5)$$

Equation (5) shows that for given observed revenue shares of each variety, less competition through the price index implies that marginal costs are higher, while revenue shares imply lower costs.

Following the intuition above, if we wished to place a firm as close as possible to its core variety while ensuring that the revenue ratios for varieties a firm produces match the observations $\{\rho_\varepsilon(\nu,\omega)\}$, then we would locate the firm by solving the following minimization problem:

$$\min_{\ell(\omega)} \|\ell(\overline{\nu}(\omega))-\ell(\omega)\| \quad \text{subject to} \quad \|\ell(\nu)-\ell(\omega)\| = \rho_\varepsilon(\nu,\omega)\|\ell(\overline{\nu}(\omega))-\ell(\omega)\|.$$

This is equivalent to the formulation:

$$\min_{\ell(\omega)} \sum_\nu \left(\frac{\|\ell(\nu)-\ell(\omega)\|}{\rho_\varepsilon(\nu,\omega)}\right)^2 \quad \text{subject to} \quad \|\ell(\nu)-\ell(\omega)\|^2 = \rho_\varepsilon(\nu,\omega)^2\|\ell(\overline{\nu}(\omega))-\ell(\omega)\|^2.$$

If we relax the constraint from this minimization problem, this amounts to running a Weighted Least Squares regression with weights $\rho_\varepsilon(\nu,\omega)$ as follows (where $\mathbb{I}$ is the identity matrix):

$$\begin{bmatrix} \ell(\nu_1)/\rho_\varepsilon(\nu_1,\omega) \\ \ell(\nu_2)/\rho_\varepsilon(\nu_2,\omega) \\ \vdots \end{bmatrix} = \begin{bmatrix} \mathbb{I}/\rho_\varepsilon(\nu_1,\omega) & 0 & \cdots \\ 0 & \mathbb{I}/\rho_\varepsilon(\nu_2,\omega) & 0 \\ \vdots & 0 & \ddots \end{bmatrix} \ell(\omega) + \eta(\omega).$$

This method of locating firms has the advantage of a clear interpretation and affords a closed form solution that is fast to compute. A firm's location who produces varieties $\{\nu_i\}$ is then given by a convex combination of the products it produces

$$\ell(\omega) = \sum_i \frac{c_\varepsilon(\nu_i,\omega)^{-2}}{\sum_j c_\varepsilon(\nu_j,\omega)^{-2}}\ell(\nu_i). \quad (6)$$

This solution shows that firms are closer to lower cost varieties, i.e. higher revenue varieties,

after accounting for preferences and competition. Notice also that a firm's proximity to unproduced varieties comes from information embedded in the product space since its location has a weight of zero on unproduced varieties. Single product firms are naturally have the same location as their product. It follows from Equation (1) that in this framework, location embodies each firms inherent capability to adopt varieties. Since adopting new products shifts the location of firms, this also implies that product scope contains information on adoption capability.

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

4 Data and Estimation Procedure

In this section we present the data and detail how the other firm- and product-level distances and coordinates are obtained from the procedure above and then present summary statistics regarding the recovered product clusters.

4.1 Cluster Construction

We use data on the value of production for 24 years, from 1995 to 2018, for Danish firms. The data is a survey which covers manufacturing firms with more than 10 employees, in which firms report the value and quantity of production (from which we can compute unit values) and is provided by Statistics Denmark.[14] This data is augmented using the VAT register provided by Statistics Denmark to net out the value of firm level exports (see Borchsenius et al., 2010). We define a product as Combined Nomenclature 8-digit (CN8) code and products for which the value of exports exceeds production are removed from the sample. In our baseline results, we separately consider 34 sectors, defined by CN 2-digit (CN2) codes.

To recover product locations, we need information on prices at the firm-product level. Since data is reported at the CN 10-digit (CN10) code, and since we define a product as a CN8 code, we aggregate quantities and values at the 8-digit level. We drop CN8 products when their corresponding CN10 products are not reported in the same unit of measure e.g., if the quantity of one CN10 good is reported as pieces and another as kilograms. We compute the price of each CN8 product as a unit value. We drop CN2 sectors that have less than 10

[14]Since the goal is to estimate the technological similarities between products and firms, we use data on the value of total production instead of data on sales in the Danish market.

CN8 products. Firms' market shares and the price indexes are computed using the sample of all firms.

When a firm produces both product A and B, we define that relationship as a *direct linkage*. When firm 1 produces both A and C and firm 2 produces both B and C, we define this relationship between product A and C as an *indirect linkage*, or if any firm produces two indirectly linked products then the products are also indirectly linked. We define a *cluster* as a group of CN8 products within a CN2 sector as the largest set of *directly or indirectly linked* products. Identifying clusters is important because we can only estimate distances for products within a cluster and the distance between products in different clusters is infinite by definition. Furthermore, we want to focus on clusters that are persistent across years and avoid clusters that only occur for a few years perhaps due to inconsistent linkages.

We proceed as follows. We recover all product distances. Following the above definition, a cluster is all products with finite distance to each other in *any* year. This is a broad definition congruent with using the average values of market share and price indexes across *all* years. This procedure returns 38 clusters from the original 34 two digit sectors.

We further refine the sample of products to avoid the possibility of having a year in which clusters contain multiple sub-clusters. We do this by selecting products in the intersection of the largest clusters in each CN2 each year. For each year and cluster, we repeat the clustering procedure above. 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 7 percent of observations at the firm-product-year level.

For each cluster and each year, we then estimate the product and firm distances, following the procedure above assuming $\sigma = 5$. The procedure returns a matrix of distances between products and firms denoted by d_{pfct} for product p, firm f, cluster c, and year t. We also compute a matrix of distances between firms denoted by d^F_{fkct} for firms f and k, cluster c, and year t which appear in the variable definitions below.[15]

4.2 Cluster Analysis

Descriptive statistics of the recovered clusters are presented in Table 1. A clusters have on average 38 products and 23 firms and the distribution of both are right skewed with high

[15]Notice that by construction a firm is defined as a firm-sector-cluster: no firm can belong to two clusters in the same sector, since that would create a link between the two clusters. Therefore we treat the same firm producing in two different sectors as two separate firms.

interquartile variation. The number of products and firms vary over time, notably dipping around the 2008 financial crisis.

Table 1: Cluster Descriptive Statistics

	Number of Products					**Number of Firms**				
Year	Avg.	Std.	Med.	25P.	75P.	Avg.	Std.	Med.	25P.	75P.
1995	42.1	40.3	26	17	60	32.6	41.5	23	5	40
1996	41.3	38.0	25	16	48	38.1	44.9	28	10	43
1997	40.4	36.8	26	15	62	31.6	39.2	19	6	39
1998	39.9	36.4	25	15	62	31.2	38.4	22	7	36
1999	40.7	35.2	28	15	47	29.2	35.9	21	5	39
2000	34.5	28.8	25	12	53	22.2	20.1	13	6	32
2001	35.4	31.3	22	14	60	22.7	20.1	15	8	32
2002	35.6	29.5	23	14	49	23.1	20.8	15	7	37
2003	35.5	28.1	27	14	55	20.8	17.0	18	7	32
2004	36.2	28.8	26	16	45	22.6	19.3	18	7	34
2005	36.3	27.7	25	14	62	22.2	18.7	22	8	33
2006	37.9	27.3	26	15	60	22.5	19.1	20	8	28
2007	31.1	20.7	25	14	48	17.2	15.3	14	6	21
2008	31.5	22.1	24	15	47	16.5	14.1	14	7	21
2009	32.7	21.8	23	14	53	16.5	12.3	14	6	25
2010	31.8	22.1	21	14	51	17.1	13.7	15	7	22
2011	32.6	22.7	24	14	48	16.8	14.5	14	6	21
2012	32.6	22.6	25	14	51	16.2	13.0	14	6	22
2013	35.6	24.4	24	17	54	16.7	13.7	12	8	22
2014	38.8	27.0	25	15	67	18.5	16.2	13	9	22
2015	39.2	26.6	25	19	70	18.7	15.6	15	10	20
2016	51.6	42.1	37	20	84	27.1	28.9	18	13	25
2017	52.9	43.5	36	22	82	27.6	27.3	19	13	26
2018	53.0	41.8	35	23	82	27.8	26.8	20	13	28
Average	**38.3**	**30.2**	**26.2**	**15.8**	**58.3**	**23.1**	**22.8**	**17.3**	**7.8**	**29.2**

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 colums) and number of firms (last four columns) across clusters in each year. In each year, there are 37 clusters (for 34 CN 2-digit sectors). The last row (Average) reports the average of the statistics across years.

As Table 1 emphasizes, the cluster calculation is done every year, which we take as the 'state space' to look forward in the empirics. Throughout we will include cluster-time fixed effects for comparability across years. While our method allows distances between products to change over time, they are fairly stable: cluster-time and product-pair fixed effects explain 93 per cent of the variation in product-to-product distance (see Appendix Table 11).[16]

[16]This corresponds to a sample without merged customs data at the moment, but the results are very

5 Estimation

In this section, we use recovered firm and product distances and Danish production data to study their relationships to firm performance and to product mix changes. We test how good are various measures of distance from products and from firms are at predicting sales and scope growth.[17] The main results are:

- Products closest to a firm are the most likely to be adopted and firms closer to potential new products grow faster in sales and scope.
- In clusters with higher total sales, firms grow *up* with higher sales and less scope; in higher entrant clusters, firms grow *out* with lower sales and more scope.
- Firms close to each other grow more slowly but increase scope more quickly, and are more likely to merge.

5.1 Product Distances and Growth Channels

We first consider the relationship between the proximity/dispersion of potential products and the sales growth and product scope of firms. We estimate the following equation:

$$g_{fct} = \beta \text{Distance}_{fct-1} + \text{c}_{ct} + \text{b}_f + \epsilon_{fct} \tag{7}$$

where g_{fct} is the growth rate of firm sales or scope (number of products), Distance_{fct-1} is a lagged measure of distance, c_{ct} is a cluster-year fixed effect, and b_f is a firm fixed effect. From above, log distance is basically proportional to log revenues, although this measure is based on the average of the largest lower bound of distances and the smallest upper bound of distance. The Distance measures we employ in Equation (7), which are all relative to the cluster containing the linked firms and products, are:

- Log distance of a firm from the closest product outside of its scope:

$$\text{Closest}_{fct} = \min_p \ln d_{pfct}$$

  where p is an index over all products not produced by the firm in the cluster.

similar.

[17] Note that in the present formulation, sales are proportional to profits under constant markups and since measurements are in logs, the estimates are equivalent.

- Average log distance of a firm from potential products:

$$\text{Average}_{fct} = \frac{1}{N_{fct}} \sum_{p} \ln d_{pfct}$$

  where N_{fct} is the number of products not produced by the firm in the cluster.

- Standard deviation of the log distance of potential products:

$$\text{Std. Dev.}_{fct} = \left(\frac{1}{N_{fct}} \sum_{p} \left(\ln d_{pfct} - \text{Average}_{fct} \right)^2 \right)^{1/2}.$$

Table 2 shows results when the dependent variable is the growth rate of sales. Firms with closer potential products, as measured by the closest product distance and by average product distance, grow faster. Table 3 examines the relationship of potential product distance and the growth rate of scope, finding the same qualitative results. Firms far away from potential products have lower scope growth rates.

Potential product dispersion, measured by the standard deviation of distance, is associated with higher future sales growth. The result suggests that a more concentrated product nexus is less conducive to growth, perhaps because of tougher competition that emerges when products or firms are relatively close. Examining the impact of dispersion on firm scope, we see a similar relationship: scope grows more quickly when potential products are less congested. Both facts show that if the set of potential products are less congested *due to a firm's relative position in the cluster*, the firm will grow faster along both intensive (up) and extensive (out) margins. *Room to grow in a cluster is correlated with growing up and out.*

Table 2: Sales Growth and Potential Product Distance

	Dependent Variable: Growth Rate of Sales		
	(1)	(2)	(3)
Closest Potential Product	-0.063**		
	(0.029)		
Average Potential Distance		-0.241***	
		(0.074)	
Std. Dev. of Potential Distance			0.355***
			(0.092)
Cluster-Time FE	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes
R^2	0.06	0.06	0.06
# Obs.	15588	15588	15588

Results from OLS estimation of (7). Cluster robust standard errors in parenthesis. Cluster: firm-cluster. ***: significant at 99%, ** at 95%, * at 90%. Dist. Closest is the lagged log distance of the firm from its closest non-produced product. Avg. Dist. is the lagged average log distance of the firm from the products that are not produced. Std. Dev. of Dist. is the lagged standard deviation of the log distance of products outside the scope of the firm.

Table 3: Scope Growth and Potential Product Distance

	Dependent Variable: Growth Rate of Scope		
	(1)	(2)	(3)
Closest Potential Product	-0.081***		
	(0.009)		
Average Potential Distance		-0.029	
		(0.021)	
Std. Dev. of Potential Distance			0.180***
			(0.030)
Cluster-Time FE	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes
R^2	0.05	0.04	0.04
# Obs.	15588	15588	15588

Results from OLS estimation of (7). Cluster robust standard errors in parenthesis. Cluster: firm-cluster. ***: significant at 99%, ** at 95%, * at 90%. Dist. Closest is the lagged log distance of the firm from its closest non-produced product. Avg. Dist. is the lagged average log distance of the firm from the products that are not produced. Std. Dev. of Dist. is the lagged standard deviation of the log distance of products outside the scope of the firm.

In Table 4, we examine heterogeneous effects across cluster characteristics on sales growth. We interact the measures of distance with the lagged log of total cluster sales and with the log number of firms in the cluster. The positive coefficient on the total sales interaction shows that *larger cluster sales magnify the role of distance* from outside products. Similarly, *less competitive clusters magnify the effect of distance*. In the language of monopolistic competition models, lower demand and higher firm entry mean being close to outside products is more important for sales growth.

Table 5 examines the same relationships for the growth of firm scope, and the results

are striking. While the baseline effects of the Closest, Average and Dispersion of potential product distance remain the same sign for Sales and Scope, cluster characteristics have exactly the opposite effects. *Smaller cluster sales magnify the role of distance* on scope growth and *more competitive clusters magnify the effect of distance*. This suggests that *demand and firm entry within clusters have opposite effects on Sales and Scope growth*. Examining the third column of each Table, this conclusion extends to the dispersion of potential products a firm faces due to its location.

Table 4: Sales Growth and Cluster Characteristics

	Dependent Variable: Growth Rate of Sales		
	(Dist. Closest)	(Avg. Dist.)	(Std. Dev. of Dist.)
Measure	-0.074***	-0.254***	0.344***
	(0.028)	(0.072)	(0.092)
(Measure)x(Total Sales)	0.065***	0.060***	0.288***
	(0.004)	(0.003)	(0.023)
(Measure)x(N.Firms)	-0.185***	-0.157***	-0.973***
	(0.014)	(0.023)	(0.075)
Cluster-Time FE	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes
R^2	0.08	0.09	0.08
# Obs.	15588	15588	15588

Results from OLS estimation of (7). Cluster robust standard errors in parenthesis. Cluster: firm-cluster. ***: significant at 99%, ** at 95%, * at 90%. The corresponding Nexus Measure in each regression is described in the column headers. Dist. Closest is the lagged log distance of the firm from its closest non-produced product. Avg. Dist. is the lagged average log distance of the firm from the products that are not produced. Std. Dev. of Dist. is the lagged standard deviation of the log distance of products outside the scope of the firm. Total Sales = lagged log of total output value in the cluster-time. N.Firms = lagged log of the number of firms in the cluster-time.

Table 5: Scope Growth and Cluster Characteristics

	Dependent Variable: Growth Rate of Scope		
	(Dist. Closest)	(Avg. Dist.)	(Std. Dev. of Dist.)
Measure	-0.079***	-0.027	0.177***
	(0.009)	(0.021)	(0.029)
(Measure)x(Total Sales)	-0.004***	-0.005***	-0.019***
	(0.001)	(0.001)	(0.006)
(Measure)x(N.Firms)	0.035***	0.025***	0.009
	(0.004)	(0.007)	(0.023)
Cluster-Time FE	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes
R^2	0.05	0.04	0.05
# Obs.	15588	15588	15588

Results from OLS estimation of (7). Cluster robust standard errors in parenthesis. Cluster: firm-cluster. ***: significant at 99%, ** at 95%, * at 90%. The corresponding Nexus Measure in each regression is described in the column headers. Dist. Closest is the lagged log distance of the firm from its closest non-produced product. Avg. Dist. is the lagged average log distance of the firm from the products that are not produced. Std. Dev. of Dist. is the lagged standard deviation of the log distance of products outside the scope of the firm. Total Sales = lagged log of total output value in the cluster-time. N.Firms = lagged log of the number of firms in the cluster-time.

Until now, we have only examined the product landscape, and the not the competitors who inhabit it and for that we need to look at the *proximity of firm's to each other*. This will reveal what the role of firm to firm competition has on Sales versus Scope.

5.2 Firm Proximity and Growth Channels

We now examine the role of the distance between firms. The take home message is that in the abscence of nearby competitors, firms are more likely to *grow up* on the intensive margin rather than *grow out* on the extensive product margin. The distance between two firms depends on their product mix: for instance, if two firms produce the same products with a similar distribution of market shares, they will also be close to one another. However, interfirm distance does not rely only on the similarity of product mix, as it contains information about the technological distance of all products. Hence, two firms may be close even if they produce disjoint sets of products, since the products might be close as observed in the co-production of third party firms.

Paralleling our definitions between firms and products, we consider here the following measures of distance for a firm f in cluster c and year t. Let d^F_{fkct} denote the distance between firm f and firm k.

- Log distance of a firm from the closest competitor:

$$\text{Closest}^F_{fct} = \min_k \ln d^F_{fkct}.$$

- Average log distance of a firm from all other firms in the cluster:

$$\text{Average}^F_{fct} = \frac{1}{N^F_{ct} - 1} \sum_k \ln d^F_{fkct},$$

  where N^F_{ct} is the number of firms in the cluster-year.

- Standard deviation of the log distance between a firm and all other firms in the cluster:

$$\text{Std. Dev.}^F_{fct} = \left(\frac{1}{N^F_{ct} - 1} \sum_k \left(\ln d^F_{fkct} - \text{Average}^F_{fct} \right)^2 \right)^{1/2}.$$

First, we estimate equations similar to Equation (7), using measures of distance between

firms, with the same fixed effect definitions as above:

$$g_{fct} = \beta \text{Distance}^{F}_{fct-1} + \text{c}_{ct} + \text{b}_{f} + \epsilon_{fct}. \tag{8}$$

Results are presented for Sales in Table 6 and for Scope in Table 7. Now the impact of having a *technologically close competitor* is clear. The farther the closest competitor is away, the faster a firm grows in sales and less in scope. Notably the effect of a close competitor on Sales is opposite that of Table 2: having a close potential product is conducive to sales growth, but not a close competitior. In fact, *the effect of a close competitor hurts sales grown more than an equally close product helps.*

Having a close competitor decreases sales growth but instead a firm grows by increasing its scope. In short, *close competition makes firms grow out rather than up.* This fits closely with theories of the firm where firms adopt new products to grow in the face of exhausted or contested markets. The level of competition in a cluster as measured by the average distance away from other firms shows that competition eats into both Sales and Scope growth as one might expect.[18] In the language of monopolistic competition models, this aggregate measure of competition has the usual effect, but nearby individual firms cause adaption in growth strategies.[19]

Table 6: Sales Growth and Interfirm Distance

	Dependent Variable: Growth Rate of Sales		
	(1)	(2)	(3)
Dist. Closest	0.198***		
	(0.012)		
Avg. Dist.		0.841***	
		(0.055)	
Std. Dev. of Dist.			-0.767***
			(0.059)
Cluster-Time FE	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes
R^2	0.13	0.14	0.13
# Obs.	15588	15588	15588

Results from OLS estimation of (7) using measures of distance that are firm-to-firm. Cluster robust standard errors in parenthesis. Cluster: firm-cluster. ***: significant at 99%, ** at 95%, * at 90%. Dist. Closest is the lagged log distance of the firm from its closest competitor. Avg. Dist. is the lagged average log distance of the firm from its competitors, namely the other firms in the cluster-time. Std. Dev. of Dist. is the standard deviation of the log distance of the competitors.

[18]The role of competitive dispersion relative to a firm is less clear: controlling for cluster-time averages, the more dispersed competition is relative to a firm, the lower sales growth but this does not significantly effect scope.

[19]This is in line with theories contrasting product versus process innovation, see Dhingra (2013).

Table 7: Scope Growth and Interfirm Distance

	Dependent Variable: Growth Rate of Scope		
	(1)	(2)	(3)
Dist. Closest	-0.022***		
	(0.004)		
Avg. Dist.		0.094***	
		(0.017)	
Std. Dev. of Dist.			0.019
			(0.019)
Cluster-Time FE	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes
R^2	0.04	0.04	0.04
# Obs.	15588	15588	15588

Results from OLS estimation of (7) using measures of distance that are firm-to-firm. Cluster robust standard errors in parenthesis. Cluster: firm-cluster. ***: significant at 99%, ** at 95%, * at 90%. Dist. Closest is the lagged log distance of the firm from its closest competitor. Avg. Dist. is the lagged average log distance of the firm from its competitors, namely the other firms in the cluster-time. Std. Dev. of Dist. is the standard deviation of the log distance of the competitors.

We've seen that the presence of a closeby competitor decreases Sales growth and increases Scope, in line with strategies to exploit new markets or diversify a firm's activities. Another strategy in the presence of a technologically close competitor could be to acquire them. After all, "if you can't beat them, join them." We now turn to the role of interfirm distance and mergers and acquisitions.

5.3 Mergers and Acquisitions with Nearby Firms

Here we use the distances between firms to predict the probability of mergers and acquisitions. We use data on mergers and acquisitions (M&As) across Danish firms from Chan et al. (2022) who identify mergers using the method proposed by Smeets et al. (2016). The economic motivations for M&As are twofold: on the one hand, M&As increase the market power of the acquirer, while on the other they can lead to reduction in costs, due to sharing of fixed input requirements or to other synergies. Our distance measure speaks to both channels, since the distance between two firms reflects differences in both technology and production portfolios. Similar to the above specifications, we estimate the following equation:

$$\text{Merger}_{fkct} = \ln d^F_{fkct} + \text{c}_{ct} + \text{b}_f + \text{e}_k + \epsilon_{fkct} \tag{9}$$

where $\text{Merger}_{fkct} = 1$ if firms f acquires k, or vice versa, in year $t+1$ and cluster c, and 0 otherwise. We normalize this variable using the average merger rate.[20] The results in Table 8 show that nearby firms are more likely to merge. In fact, the coefficients on firm-to-firm distance both one and two years before the merger are negative and significant. *So another way of 'growing out' besides expanding scope is to acquire a nearby competitor.*

Table 8: Mergers and Interfirm Distance

	Dependent Variable: Merger Next Period			
	(1)	(2)	(3)	(4)
Log Firm-to-Firm Distance	-1.439***	-1.653***		
	(0.460)	(0.502)		
Lagged Log Firm-to-Firm Distance			-1.100**	-1.426***
			(0.458)	(0.496)
Cluster-Time FE	Yes	Yes	Yes	Yes
Firms FE	No	Yes	No	Yes
R^2	0.01	0.04	0.01	0.05
# Obs.	454163	454163	267137	267137

Results from OLS estimation of (9). Cluster robust standard errors in parenthesis. Cluster: firm-to-firm pair. ***: significant at 99%, ** at 95%, * at 90%.

We've now seen that firm-product and firm-firm distances impact the growth of sales and product scope. We now turn examine the extent to which these distances predict *which* products firms are likely to introduce.

6 Product Expansion Paths

While many changes in the economic environment may incentivize firms to introduce or discontinue products, predicting product adoption relative to the entire set of potential products is much less understood. For instance, in the standard models in which the product mix decision of a firm is based on a core competence, models predict that a positive productivity shock increases the scope of a firm and that the new product is far from the core. However, such models cannot ex-ante predict whether the new product introduced is product A or product B. This is because these studies are based on looking *within the firm*, whereas we combine *information across firms* to arrive at richer predictions of product expansion paths. In this section, we quantify the ability of our joint firm and product location classification to predict which products are introduced by a firm.

[20] In our dataset there are 150 mergers between manufacturing firms, which implies an average merger rate of 0.01 percent relative to all firm-pairs in all years.

Throughout the section we will consider the distance rank of potential products in ascending order of distance. Consider a product p that can be introduced by a firm f in a cluster c and year t. We assign to each product a ranking depending on how close it is to the firm as follows:

- $Rank_{pfct}$ is the rank of the closest potential products based on the distance d_{pfct}, so that for the closest product $Rank_{pfct} = 1$ and for the second closest $Rank_{pfct} = 2$, etc.

The main results regarding the adoption path of new products are:

- Proximity to potential products predicts adoption better than the cluster sales ranking and about as well as the cluster frequency of produced products. One third of product adoptions are in the Top 10 closest products.

- The distance rank of potential products predicts adoption, and remains significant when estimated with the cluster sales and frequency ranks. Each extra rank of distance a product is from a firm decreases the frequency of adoption by one per cent relative to the base rate.

First we will examine how this ranking performs in explaining which products a firm introduces and then the section proceeds by examining product expansion paths.

6.1 Predicting Product Adoption

We compare the explantory power of this distance based ranking to two other potential predictors of the products a firm may introduce, which are specific to each cluster and which might be positively or negatively related to adoption:

- Intensive Margin Rank: the sales rank of the product in the cluster. One rationale is that products with larger sales can sustain larger demand and so encourage adoption. Conversely, such products may be oversupplied and crowd out entry.

- Extensive Margin Rank: the number of firms producing the product in the cluster. Products with more firms might sustain more entry or exhibit tougher competition discourage entry.

Note that these Rankings are cluster specific to the 37 clusters (roughly equivalent to a two digit industrial classification), so they are already highly selected. We compare our distance ranking of products away from each firm with the Intensive and Extensive Margin

Ranks in Table 9 for the subset of clusters with at least 30 products.[21] Rows 1 and 3 show that the extensive margin of production frequency is more predictive of which products are introduced than the intensive margin of sales, predicting over three times the correct introduction frequency. The product distance rank is similarly strong, predicting the correct product introductions about three times as often as the intensive margin. Looking across the Table shows that longer surviving products are even better predicted by product distance, approaching the same explanatory percentanges as the extensive margin.

Table 9: Predicting Product Introduction

	Produced $\geq$ 1 Year			Produced $\geq$ 5 Years		
	Top 1	Top 5	Top 10	Top 1	Top 5	Top 10
Intensive Margin (Sales)						
Cluster Rank	0%	4.4%	9.4%	0%	4.9%	10.2%
Product Distance ($Rank_{pfct}$)						
Cluster Rank	4.0%	17.9%	30.9%	3.6%	20.6%	34.9%
Extensive Margin (Number)						
Cluster Rank	6.6%	20.4%	33.4%	6.1%	18.9%	32.7%

Share of products that have been introduced and that are in the top 1, 5, and 10 of products by three rankings; products are ranked by their distance from the firm (Nexus), by the total number of firms producing them (Number of Firms), and by the value of total production (Total Output). The ranking is such that, for instance, the top 1 product is the closest product to the firm.

6.2 Potential Product Rankings and Adoption Path

We evaluate the goodness of the firm-productdistance ranking as follows. First, for each firm, we select the products that the firm does not produce in its first year in the data. For instance, if a firm enters the dataset in 1997, the sample of products we consider are the products in the cluster of the firm that the firm does not produce in 1997. For each product in the years after entry, we compute a production indicator Intro_{pfct} for product p, firm f, cluster c, and year t, which equals 1 if the product is produced by the firm and zero otherwise. We then divide this by the average rate of product introduction, which aproximately equals 0.05 percent. We estimate the following equation:

$$\text{Intro}_{pfct} = \beta SomeRank_{pfct-1} + a_{ft} + b_{pt} + \epsilon_{pfct} \tag{10}$$

[21] Using all clusters does not drastically change the results, however looking at the Top 5 or 10 introduced products isn't terribly informative in small clusters. As the median number of products in a cluster is on average 27, this is a bit under half of the clusters.

where $SomeRank_{pfct-1}$ is a product rank in the previous year. We consider three rankings based on the cluster specific measures of:

1. The distance of the product from the firm ($Rank_{pfct-1}$).
2. The total output value for the product (Intensive Margin Rank).
3. The total number of firms producing the product (Extensive Margin Rank).

We include firm-year fixed effects (a_{ft}) and product-year fixed effect (b_{pt}). By definition, each firm and product belongs to only one cluster, so the set of fixed effects already control for cluster-year shocks.

The results are presented in Table 10. The first two columns shows that product adoption rates decrease with the distance ranking from a firm. This result persists even controlling for product-time fixed effects and, thus, control for shocks to demand or technology that might affect a product in a year. The third and fourth columns show that firms are more likely to introduce products that are lower ranked in terms of cluster sales. The fifth and sixth columns show that less frequently produced products are less likely to be introduced, but this effect is insignificant controling for product-time fixed effects. The final two columns of the Table includes all three rankings and show that only the product distance ranking is robust when product-time fixed effects are included. Looking across the table, we see that for each extra rank of distance a product is from a firm, it decreases the frequency of adoption by about one per cent relative to the base adoption rate.

Table 10: Product Introduction and Product Rankings

	Dependent Variable: Intro_{pfct}							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Lagged Rank	-0.010***	-0.011***					-0.008***	-0.011***
($Rank_{pfct-1}$)	(0.001)	(0.001)					(0.001)	(0.001)
Lagged Rank			0.008***	0.005			0.003***	0.005
Intensive			(0.001)	(0.013)			(0.001)	(0.013)
Lagged Rank					-0.011***	0.004	-0.008***	0.004*
Extensive					(0.001)	(0.002)	(0.001)	(0.002)
Firm-Time FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Prod-Time FE	No	Yes	No	Yes	No	Yes	No	Yes
R^2	0.10	0.17	0.10	0.17	0.10	0.17	0.10	0.17
# Obs.	402915	402915	402915	402915	402915	402915	402915	402915

Results from OLS estimation of (10). Standard errors in parenthesis. ***: significant at 99%, ** at 95%, * at 90%.

7 Conclusion and Future Directions

Using observed production patterns within and across firms allows us to construct a continuous, high dimensional product classification or *product space*. Firms can also be located in this space to reveal how proximity to potential products or competitors shape sales and scope growth, competition and mergers. Distance to potential products explains which products a firm will adopt, tempered by local competition characteristics. Letting the data reveal such a classification system thereby reveals new dimensions of firm behavior with respect to both products and other firms. The distance rank of potential products away from firms helps explain the path of product adoption. Continued work might use Hummels style export demand instrument (Dhyne et al., 2021) to test theoretical mechanisms and explore different competitive strategies such as strategic patents.

Future work might take the classification here to decompose what it captures and how well it can explain the same facts in different contexts. Left unexplored also is the role for sales and scope growth and product adoption in export markets. By being a continuous and data driven classification method, it also naturally opens up an analysis of firms engaged in non-manufacturing activities like services or hybrid activities since firms often provide both goods and services. Using household scanner data with a similar approach could also produce a complementary product and household classification based on consumption, rather than production and sales patterns. While only a first pass at understanding growth strategies and competition in terms of firm-product and firm-firm distances constructed from industry wide patterns, it does so without relying on handed down classifications of economic activities. It has the potential to free subsequent analysis from some of the vagaries of categorization systems across countries and over time.

References

Arkolakis, C., S. Ganapati, and M.-A. Muendler (2021): "The Extensive Margin of Exporting Products: A Firm-level Analysis," *American Economic Journal: Macroeconomics*, 13. 7

Bernard, A. B., E. J. Blanchard, I. Van Beveren, and H. Vandenbussche (2018): "Carry-Along Trade," *The Review of Economic Studies*, 86, 526–563. 7

Bernard, A. B., S. J. Redding, and P. K. Schott (2010): "Multiple-Product Firms and Product Switching," *American Economic Review*, 100, 70–97. 1

——— (2011): "Multiproduct Firms and Trade Liberalization," *The Quarterly Journal of Economics*, 126, 1271–1318. 1

Bernard, A. B. and Y. Zi (2023): "Sparse Production Networks," 55. 11

Bishop, A., J. Mateos-Garcia, and G. Richardson (2022): "Using text data to improve industrial statistics in the UK," *Working Paper*. 1

Boehm, J., S. Dhingra, and J. Morrow (2019): "The comparative advantage of firms," *Working Paper*, publisher: CEPR Discussion Paper No. DP13699. 1

Borchsenius, V., N. Malchow-MÃžller, J. R. Munch, J. R. Skaksen, and others (2010): "International trade in servicesâEvidence from Danish micro data," *NationalÃžkonomisk tidsskrift*, 148, 86–107. 4.1

Chan, J., M. Irlacher, and M. Koch (2022): "Multiproduct Mergers and the Product Mix in Domestic and Foreign Markets," *Working Paper*. 5.3

Commission of the European Communities (1992): "REGULATION (EEC) No 4064/89 MERGER PROCEDURE," . 1

Dhingra, S. (2013): "Trading Away Wide Brands for Cheap Brands," *American Economic Review*, 103, 2554–84. 7, 19

Dhyne, E., A. K. Kikkawa, M. Mogstad, and F. Tintelnot (2021): "Trade and Domestic Production Networks," *The Review of Economic Studies*, 88, 643–668. 7

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. 7

Eckel, C. and P. J. Neary (2010): "Multi-Product Firms and Flexible Manufacturing in the Global Economy," *Review of Economic Studies*, 77, 188–217. 1, 7

Escolar, E., Y. Hiraoka, M. Igami, and Y. Ozcan (2020): "Mapping Firms' Locations in Technological Space: A Topological Analysis of Patent Statistics," *SSRN Electronic Journal*. 1

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*. 7

Flach, L. and M. Irlacher (2018): "Product versus Process: Innovation Strategies of Multiproduct Firms," *American Economic Journal: Microeconomics*, 10, 236–77. 7

Girard, M. and A. Trau (2004): "Implementing the North American Industry Classification System: The Canadian Experience," Tech. rep., Statistics Canada. 1

Goldberg, P. K., A. K. Khandelwal, N. Pavcnik, and P. Topalova (2010): "Multiproduct firms and product turnover in the developing world: Evidence from india," *The Review of Economics and Statistics*, 92, 1042–1049. 1

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

Irlacher, M. (2022): "Multi-Product Firms in International Economics," *CESifo Working Paper*. 6

Jacobs, G. and C. ONeill (2003): "On the reliability (or otherwise) of SIC codes," *European Business Review*, publisher: MCB UP Ltd. 1

Kogan, L., D. Papanikolaou, L. D. Schmidt, and B. Seegmiller (2021): "Technology, Vintage-Specific Human Capital, and Labor Displacement: Evidence from Linking Patents with Occupations," Tech. Rep. w29552, National Bureau of Economic Research, Cambridge, MA. 1

Macedoni, L. (2022): "Large multiproduct exporters across rich and poor countries: Theory and evidence," *Journal of Development Economics*, 156, 102835. 7

Macedoni, L. and M. J. Xu (2022): "Flexibility and Productivity: Towards the Understanding of Firm Heterogeneity," *International Economic Review*. 7

Mayer, T., M. J. Melitz, and G. I. Ottaviano (2014a): "Market Size, Competition, and the Product Mix of Exporters," *The American Economic Review*, 104, 495–536. 3

Mayer, T., M. J. Melitz, and G. I. P. Ottaviano (2014b): "Market Size, Competition, and the Product Mix of Exporters," *American Economic Review*, 104, 495–536. 7

Melitz, M. J. (2003): "The Impact of Trade on Intra-Industry Reallocations and Aggregate Industry Productivity," *Econometrica*, 71, 1695–1725. 3

Nocke, V. and S. Yeaple (2014): "Globalization and multiproduct firms," *International Economic Review*, 55, 993–1018. 7

Sheveleva, L. (2019): "Multi-product Exporters: Facts and Fiction," *SSRN Electronic Journal*. 11

Smeets, V., K. Ierulli, and M. Gibbs (2016): "An empirical analysis of post-merger organizational integration," *The Scandinavian Journal of Economics*, 118, 463–493, publisher: Wiley Online Library. 5.3

A Cluster Analysis

A.1 Stability of Product Distances

Table 11: Stability of Product-to-Product Distance

	Dependent Variable: Log Distance		
	(1)	(2)	(3)
Cluster-Time FE	Yes	Yes	Yes
Product i and j FE	No	Yes	No
Product Pair FE	No	No	Yes
R^2	0.84	0.88	0.93
# Obs.	935468	935468	935468

OLS of log distance on fixed effects.

A.2 Continuous Classification and CN Classification

Here we use our continuous distance measure to see to what extent the discrete CN classification system reflects our approach. First, we query whether products within the same CN code are closer than products outside the CN code. For any two products i and j in cluster-time ct, we estimate the following regression:

$$\ln \text{Distance}_{ijct} = \text{ Same CN4}_{ij} + \text{ Same CN6}_{ij} + FE_{ct} + \epsilon_{ijct} \tag{11}$$

where Same $\text{CN4}_{ij} = 1$ if products i and j are in the same CN four-digit code and zero otherwise. Similarly, Same $\text{CN6}_{ij} = 1$ if products i and j are in the same CN six-digit code and zero otherwise. Table 12 shows the results of a specification using cluster-time fixed effects, which are qualitatively similar to subsamples fixing the year or adding product or product-year fixed effects. The results indicate that products within the same CN four-digit code or the same CN six-digit code are closer to one another, which validates the usual interpretation that products with more common leading digits are more similar.

Table 12: Products within the same CN codes are closer

	Dependent Variable: Log Distance		
	(1)	(2)	(3)
Same CN4	-0.694***		-0.671***
	(0.003)		(0.004)
Same CN6		-0.737***	-0.168***
		(0.009)	(0.009)
Cluster-Time FE	Yes	Yes	Yes
R^2	0.83	0.83	0.83
# Obs.	1386345	1386345	1386345

OLS of (11). ***: significant at 99%, ** at 95%, * at 90%.

However, the discreteness of the CN classification may mask important differences between products. In common usage, all products within a certain CN four- or six-digit codes are implicitly treated as if they have the same 'distance' in the sense of having the same elasticity of substitution in preferences or production. If this was literally reflected in our distance measure, the variance of the product distance within codes should be zero, or at least less than the for products outside any particular 4 or 6 digit code. To assess this claim, we compute the standard deviation of log distance within CN six-digit codes.[22] Then, we compute the ratio of the standard deviation of each CN six-digit code relative to the standard deviation of the log distance within the cluster in which the CN six-digit code belongs to. We then take the simple average of the ratio of standard deviations within a cluster and average across years and clusters.

The results are in Table 13. The standard deviation of distances within CN six-digit codes is similar to the standard deviation within a cluster (the ratio of the two is on average 0.96). The measure is also similar to the standard deviation of distance across products that do not belong to the same CN six-digit code. This suggests that the discrete classification system is still missing substantial differences across products even within the same narrowly defined categories. For CN four-digit codes, the results are similar.

[22]We drop codes with less than 5 distances within them and clusters with less than five six-digit codes.

Table 13: Dispersion of Distance

	Standard Deviation of log distance relative to cluster	
	Within CN6	Outside CN6
Average Across Clusters	0.96	0.98
	Within CN4	Outside CN4
Average Across Clusters	0.99	0.94

The table reports the standard deviation of distance within CN6 (CN4) and outside CN6 (CN4) relative to the standard deviation of distance in a cluster.

B Product Classifications Across HS, SIC and NAICS

Table 14: 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)