Of Rivalry and Synergy: Nonlinear Patents and Products Similarity Effects on R&D Investment

Dario Marino

Excerpt (30 pages) from UChicago Master's Thesis

September 28, 2025

Abstract

This paper examines how the joint effect of product and technology similarity among firms shapes R&D investment. Technological similarity creates synergies by enabling improvements to be shared between related firms, within the limits of patent protection. Product similarity encourages R&D for product differentiation, however it discourages R&D for process innovation because competitors can absorb process improvements and outcompete the innovator.

I formalize this mechanism with a Cournot competition model featuring cost-reducing R&D and knowledge spillovers weighted by technological similarity and moderated by patent protection. Product similarity enters through substitutability in the demand system. Simulations of 1,000 firms with normally distributed costs and nonlinear product differentiation yield predictions for R&D and output at optimal patent transmission.

Empirically, I build a network panel dataset by combining Hoberg and Phillips product similarity embeddings with pairwise patent portfolio similarities from PatentSBERTa_V2 embeddings of all USPTO abstracts, merged with Compustat data. Consistent with prior literature, linear models fail to identify clear effects of patent and technology similarity on R&D. To capture the joint and nonlinear effects predicted by the model, I estimate a Generalized Additive Model (GAM), mapping the interaction between product and technology similarity. The results are in line with our theoretical simulations. I further validate the empirical results by using two alternative datasets, one with older patents and product similarity measures, and one with detailed semiconductor manufacturing data from the APTO NSF grant. I also estimate effective patent protection in the semiconductor sector for the period covered (1986-2024).

Keywords: Innovation, Intellectual Property, Industrial Organization,
Applied Microeconomics, Network Economics, Natural Language Processing

1 Introduction

This paper examines how the joint effect of product and technology similarity among firms shapes R&D investment decisions. Technological similarity creates synergies by allowing improvements to be shared between closely related firms, to the extent permitted by patent protection. In contrast, product similarity can discourage R&D because competitors may absorb process improvements and outcompete the investing firm. At the same time, product similarity can encourage R&D aimed at product differentiation.

Bloom, Schankerman, and Reenen 2013 have already explored the negative effect that product market spillover has on R&D investments. Their framework does not model optimal patent protection based on the interaction between product and technology similarity. Their empirical analysis focuses mostly on the aggregate effect of R&D synergy over business stealing, concluding that the societal return of synergy is higher than that of market stealing. The authors conclude that, since firms have lower private returns compared to the societal returns of R&D, innovation will be under-provided in society. The authors' results regarding R&D are not significant or ambiguous for the most part. It is tempting to attribute this to the available measures of product and patent similarity at the time, which were SIC-based or citation-based instead of NLP-based. However we are going to see that, even with more modern measures, we still find ambiguous linear results of product and technology similarity on R&D. I will argue both theoretically and empirically about why considering them jointly and nonlinearly will clarify their effect on R&D.

I formalize this mechanism using a Cournot competition model with cost-reducing R&D and knowledge spillovers weighted by technological similarity and moderated by patent protection. Product similarity is introduced through substitutability in the demand system. Simulations of 1,000 firms featuring normally distributed costs and nonlinear product differentiation yield theoretical predictions for R&D expenditures and quantity at optimal levels of patent transmission. Patent transmission positively affects welfare and R&D by making process improvements more widely available between firms, including competitors, while discouraging investment in product differentiation.

Linear and nonlinear Cournot models for R&D investment can be already found in López and Vives 2019 for example, where they focus on the effect of common ownership on R&D investment. In Antón et al. 2025 the relationship between common ownership and innovation inputs becomes less negative and can even turn positive the larger technology spillovers are. For those with relatively high technology and low product market spillovers (or similarity) the relationship is positive whereas for low technology and high product market spillovers it is negative. In these Cournot models, technological spillover effects are considered within a common ownership perspective rather than the more widespread process of inventing around all other firms' improvements as much as patent protection allows. Namely, synergy is observed through the common ownership question rather than through the idea of widespread imitation.

An analysis of a duopoly cournot model with technological collaboration has been developed in d'Aspremont and Jacquemin 1988. They observed the competition effect in the collaboration between two firms at the R&D level. They did not explore N firm's frameworks and their simulation nor empirical results, probably due to the complexity of these tasks at their time. For more modern and network-based models where synergy and rivalry is observed through an explicit firm collaboration setting it is possible to consult Goyal and Moraga-Gonzalez 2001 and König, Liu, and Zenou 2019.

Our model of innovation is based on expected value rather than arrival rate. For this reason, we do not focus on the distinction between incremental and disruptive innovation as in Byun, Oh, and Xia 2019. In this paper, they found that higher technology similarity reduces the number of superstar innovators in more technologically similar firms. I also didn't focus on the different effects of firm concentration on welfare and R&D, for this analysis it is better to look at Stepanova 2009.

Our model does not use two steps like in the classic KMZ model developed in Kamien, Muller, and Zang 1992. In our linear model quantity and R&D are independent so they can be solved separately, while in the nonlinear model, we solve them together for the optimal quantity and R&D. Therefore our model does not assume that the R&D decisions are made before engaging in production.

I simulated my Cournot model with 1000 firms, but the code is generalizable to N firms. I started with normally distributed marginal costs to observe the product and technology similarity effects on R&D and the suggested optimal patent protection in a simple environment. Finally, I included a nonlinear product differentiation investment, to take into account not only process innovation (cost reduction) but also product innovation (increase differentiation). The differentiation is nonlinear because it is exponentially related to patent protection and R&D investment. When we introduce product differentiation it is possible to see the competition incentive to invest more in R&D to differentiate its product portfolio. We were able to start from the Duopoly case and move up to 1,000 firms both for the base and the nonlinear models. To move from 10 firms to 1000 for the nonlinear model we had to restrict the connection between firms. To increase the number of firms it requires to limit the links between firms to make the product/technology similarity matrix solvable. I explicitly derive those conditions, which can be enforced through a firm exit or an explicit link formation process. Full connection between firms makes the model unstable because it can't respect our theoretically derived eigenvalue stability condition. By bounding the possible connection between firms at the product and technology similarity matrix the model becomes solvable and the output are displayed and then tested in the empirical part.

This simulation shows how the current way of assigning patent protection falls short of one dimension. Usually patent protection focuses only on technology similarity and overlooks the product similarity component, which as we can see it's fundamental in the decision of guaranteeing the right patent protection. To find the optimal patent protection we would need to move from an $\mathbb{R}^1$ decision (a line between two technologies) to an $\mathbb{R}^2$ decision (a space where two technologies also embody their firms' product portfolios positions). It is also necessary to understand the sinergy and rivalry effects to take into account the whole space of patent and technology similarity, because their effect on R&D comes jointly.

We already know from Acemoglu and Akcigit 2006 that optimal patent protection is sectoral, and it should depend on the competition and the technological gap state of the sector. In this paper I also argue that optimal patent protection should also take into account the interaction between technology and product similarity.

Encaoua, Guellec, and Martínez 2006 reminds us that technological transmission is also blocked by other forces outside of patent protection. Indeed, even with no patent protection, technological transmission cannot be increased to 1. This happens because there are different obstacles to imitating technologies between firms such as first-mover advantage and intellectual barriers to entry. To integrate these effects it will be enough to set an upper bound limit to patent transmission.

In Aghion et al. 2005 we see the concept of inverted U relationship for innovation. We see that for extremely low and extremely high levels of competition, innovation is lower, while it is higher for the intermediate levels. This could confirm the case for a patent protection regime that takes into account the level of competition of a sector before assigning a patent. My analysis cannot focus specifically on this factor because

I am not taking into account market concentration for now. However, we can have a proxy for competition with the number of firms in the market. We see in the base model that the average R&D investment increases when we move from duopoly to 10 firms, and then decrease at 1000 firms. When we add nonlinear product differentiation in the extension of our model we get the "escaping competition" mechanism also mentioned in Aghion et al. 2005.

Empirically, I build a network panel dataset using Hoberg-Phillips product similarity embeddings and by computing pairwise patent portfolios similarities using PatentSBERTa_V2 embeddings of all USPTO abstracts, merged with Compustat financials. Following previous literature we find that linear models fail in estimating unambiguous results of patent and technology similarity on R&D investment. To compute our joint and nonlinear effects observed in the model we are going to estimate a Generalized Additive Model (GAM) on our data to observe the estimated nonlinear effect of product and technology on R&D throughout the interaction grid. This nonlinear strategy will reconcile theory and data by finding unambiguous results.

2 Model

2.0.1 Market Structure and Demand

Let $i \in \mathcal{I} = \{1, \ldots, N\}$ index firms, each producing a differentiated product portfolio. Denote by q_i the quantity produced and by p_i the corresponding price. The inverse demand function is given by

$$p_i = A - q_i - \sum_{j \neq i} \delta_{ij} \, q_j,$$

where $A > 0$ is a market size parameter and $\delta_{ij} \geq 0$ measures substitutability between products i and j.

2.0.2 Technology Network

Each firm i is endowed with a technology portfolio $\mathcal{T}_i \subseteq \mathcal{T}$. For each technology $t \in \mathcal{T}_i$, the firm invests z_{it}. The total technology investment is defined as

$$z_i = \sum_{t \in \mathcal{T}_i} z_{it}.$$

Technology portfolio overlap between firms i and j is captured by ω_{ij}.

2.0.3 Cost Structure and Profit

Firms experience a cost reduction from technology spillovers. The effective marginal cost for firm i is

$$MC_i = c_i - \phi \sum_{j \in \mathcal{I}} \omega_{ij} \, z_j,$$

where c_i is the baseline marginal cost and ϕ represents the permitted patent transmission. ω_{ij} is the technological overlap, meaning that the more two firms have similar technology portfolio the more they could use the other firms improvements, weighted by the patent transmission allowed by the regulator. The technology similarity with itself is: $\omega_{ii} = \frac{1}{\phi}$, with ϕ at denominator that eliminates patent protection and sets overlap to 1 because the firm can fully use all its own developed technologies. The profit function is:

$$\pi_i = p_i \, q_i - MC_i \, q_i - \frac{\kappa_i}{2} \, z_i^2,$$

with $\kappa_i > 0$ and z_i^2 capturing the convexity of technology investment costs. κ_i can be different for every firm, depending on their R&D capabilities. We will see in the derivation how we are going to need to fix some lower bound conditions for this cost parameter κ_i.

2.0.4 Equilibrium and Welfare

Firms choose (q_i, z_i) to maximize their profits under Cournot competition. The regulator chooses ϕ to maximize overall welfare,

$$W = CS + \sum_{i \in \mathcal{I}} \pi_i,$$

where CS denotes consumer surplus.

2.0.5 Consumer Surplus

To compute the consumer surplus we take the area under each inverse-demand curve up to the equilibrium quantity q_i and then sum over i.

$$CS \; = \; \sum_{i=1}^{N} \int_0^{q_i} \Big[A \; - \; x \; - \; \sum_{j \neq i} \delta_{ij} \, q_j \Big] \, dx.$$

Since for each good i:

$$\int_0^{q_i} \left[A - x - \sum_{j \neq i} \delta_{ij} \, q_j \right] dx \; = \; A \, q_i \; - \; \frac{1}{2} q_i^2 \; - \; \Big(\sum_{j \neq i} \delta_{ij} \, q_j \Big) \, q_i,$$

We get the closed-form:

$$CS(q_1, \ldots, q_N) \; = \; \sum_{i=1}^{N} \Big[A \, q_i \; - \; \tfrac{1}{2} \, q_i^2 \; - \; \sum_{j \neq i} \delta_{ij} \, q_j \, q_i \Big].$$

2.1 Nonlinear General Solution with Investment in Differentiation

It is known that innovation can be broadly divided in product and process innovation. We already took into account process innovation by modeling it as a cost reducing effect. We should also take into account product innovation, and we can model it as a proportional reduction in substitutability with respect to all the other firms weighted by how much other firm can copy the product differentiation itself.

The "escape competition" (as in Aghion et al. 2005) will make a certain firm reduce its similarity with the other product portfolios and capture a different space of the market (A) with less competition. The competitors will have to fight for the remaining part of the market. We have product similarity at the numerator and the R&D investment amount (z_i) at denominator. The R&D investment amount is raised to the $(1 - \phi)$, because with no patent transmission $(\phi = 0)$ each firm takes the full escape competition reduction from its own R&D investment, because no other firm can imitate it. While with full transmission $(\phi = 1)$ there is no "escape competition" effect from R&D investment since it can be imitated by the other firms at no costs. In this case the denominator goes to 1 and we have the full substitution effect that we had in the linear case. With the other values of ϕ each firm develops its product to escape competition but also imitates the other firms to not be left out of the market, and this is also captured by the denominator, which progressively abates the effect of product differentiation the more easy it becomes to imitate.

With N number of firms we will again solve a linear system with a $N \times N$ inverse matrix. Remind that for a general firm i its profit is

$$\pi_i \;=\; \Big[A \;-\; q_i \;-\; \sum_{\substack{j=1\\ j\neq i}}^{N} \frac{\delta_{ij}}{z_i^{1-\phi}}\, q_j\Big]\, q_i \;-\; \Big[c_i \;- z_i \;-\; \phi \sum_{\substack{j=1\\ j\neq i}}^{N} \omega_{i\,j}\, z_j\Big]\, q_i \;-\; \frac{\kappa_i}{2}\, z_i^2.$$

Quantity FOC ($\partial \pi_i / \partial q_i = 0$), rearranged:

$$2\, q_i \;+\; \sum_{\substack{j=1\\ j\neq i}}^{N} \frac{\delta_{ij}}{z_i^{1-\phi}}\, q_j \;-\; z_i \;-\; \phi \sum_{\substack{j=1\\ j\neq i}}^{N} \omega_{ij}\, z_j \;=\; A - c_i.$$

R&D (investment) FOC ($\frac{\partial \pi_i}{\partial z_i} = 0$):

$$(1-\phi)\, q_i \Big(\sum_{\substack{j=1\\ j\neq i}}^{N} \delta_{ij}\, q_j\Big)\, z_i^{(\phi-2)} \;+\; q_i \;-\; \kappa_i\, z_i \;=\; 0.$$

Since we need to prove that this R&D FOC has a solution we derive this polynomial form by multiplying both sides by $-(z_i^{2-\phi})$. We are also flipping sign, which something to keep in mind for later on:

$$\kappa_i\, z_i^{3-\phi} \;-\; q_i\, z_i^{2-\phi} \;-\; (1-\phi)\, q_i \Big(\sum_{\substack{j=1\\ j\neq i}}^{N} \delta_{ij}\, q_j\Big) \;=\; 0.$$

Indeed we an see that when there is full patent transmission ($\phi = 1$) we come back to the linear case because there is no escape competition effect:

$$\kappa_i\, z_i^2 \;-\; q_i\, z_i \;=\; 0.$$

$$z_i = \frac{q_i}{\kappa_i}$$

If $\kappa_i > 0,\ \phi \in [0,1],\ q_i > 0, \Big(\sum_{\substack{j=1\\ j\neq i}}^{N} \delta_{ij}\, q_j\Big) > 0$, then $\frac{\partial \pi_i}{\partial z_i} = 0$ has at least one solution with $z \in [0, \infty[$.

At $z = 0$:

$$\frac{\partial \pi_i}{\partial z_i}(0) \;=\; \kappa_i\, 0^{\,3-\phi} \;-\; q_i\, 0^{\,2-\phi} \;-\; (1-\phi)\, q_i \Big(\sum_{\substack{j=1\\ j\neq i}}^{N} \delta_{ij}\, q_j\Big) \;=\; -\,(1-\phi)\, q_i \Big(\sum_{\substack{j=1\\ j\neq i}}^{N} \delta_{ij}\, q_j\Big) \;<\; 0.$$

As $z \to +\infty$, since $\phi \in [0,1]$ the leading term $\kappa_i z^{\,3-\phi}$ always dominates when investment goes to infinity. Hence:

$$\lim_{z \to +\infty} \frac{\partial \pi_i}{\partial z_i}(z) \to +\infty.$$

So $\frac{\partial \pi_i}{\partial z_i}(0) < 0$ and $\frac{\partial \pi_i}{\partial z_i}(z_i) \to +\infty$ as $z \to \infty$. Following Bolzano theorem we have that for $z \in [0, \mathcal{M}]$ with $\mathcal{M} >> 0$ and for the continuous function $f(0) < 0,\ f(\mathcal{M}) > 0$, there must be at least one $z > 0$ with $\frac{\partial \pi_i}{\partial z_i}(z_i) = 0$.

To prove uniqueness we have to prove that:

$$\frac{\partial \pi_i}{\partial z_i} \;=\; \kappa_i\, z_i^{3-\phi} - q_i\, z_i^{2-\phi} - (1-\phi)\, q_i \sum_{\substack{j=1 \\ j\neq i}}^{N} \delta_{ij} q_j = 0$$

can have at most one positive root.

First we have to prove that we reach the maximum with our critical point. To do so, I compute the derivative with respect to z_i of the first derivative, giving us the second derivative with flipped sign, because we flipped the sign before:

$$-\frac{\partial^2 \pi_i}{\partial z_i^2} \;=\; (3-\phi)\,\kappa_i\, z_i^{2-\phi} \;-\; (2-\phi)\, q_i\, z_i^{1-\phi}.$$

Factor out the positive term $z_i^{1-\phi}$ Since for $z_i > 0$, $z_i^{1-\phi} > 0$, we can write

$$-\frac{\partial^2 \pi_i}{\partial z_i^2} = z_i^{1-\phi}\,\Big[(3-\phi)\,\kappa_i\, z_i \;-\; (2-\phi)\, q_i\Big].$$

Define the linear factor

$$L(z_i) \;=\; (3-\phi)\,\kappa_i\, z_i - (2-\phi)\, q_i.$$

We get that $L(z_i)$ has exactly one zero at:

$$z_{i,0} = \frac{(2-\phi)\, q_i}{(3-\phi)\,\kappa_i} > 0.$$

It follows that, since z_i is connected only to the positive part in the linear factor:

When $0 < z < z_{i,0}$, $(3-\phi)\,\kappa_i\, z_i < (2-\phi)\, q_i$, $\frac{\partial^2 \pi_i}{\partial z_i^2} > 0$, so this is the part where the profit function is minimized if there is a critical point in this interval.

When $z > z_{i,0}$, $(3-\phi)\,\kappa_i\, z_i > (2-\phi)\, q_i$, $\frac{\partial^2 \pi_i}{\partial z_i^2} < 0$, so this is the part where the profit function is maximized if there is a critical point in this interval.

We see again that the model is consistent because for our full patent transmission case ($\phi = 1$) we revert back to the linear case and indeed $\frac{q_i}{\kappa_i} > \frac{q_i}{2\kappa_i}$

Now we can prove uniqueness. Coming back to the previous application of the Bolzano's Theorem:

$$\frac{\partial \pi_i}{\partial z_i}(0) = -\,\phi\, q_i \sum_{\substack{j=1 \\ j\neq i}}^{N} \delta_{ij} q_j < 0,$$

$$\frac{\partial \pi_i}{\partial z_i}(z_i) \to +\infty \text{ as } z_i \to \infty.$$

We confirm that $\frac{\partial \pi_i}{\partial z_i} = 0$ starts negative at $z = 0$. We have shown:
$\frac{\partial^2 \pi_i}{\partial z_i^2} = z^{1-\phi}\big[(3-\phi)\kappa_i\, z - (2-\phi)q_i\big]$ changes sign once at

$$z_0 = \frac{(2-\phi)q_i}{(3-\phi)\kappa_i}\,,$$

so $\frac{\partial \pi_i}{\partial z_i}$ decreases on $(0, z_0)$ and increases on (z_0, ∞). Hence we have at most one positive root z^*.

Because $\frac{\partial \pi_i}{\partial z_i}$ is still below zero at z_0 (it's been decreasing since $\frac{\partial \pi_i}{\partial z_i}(0) < 0$), the crossing $\frac{\partial \pi_i}{\partial z_i}(z^*) = 0$ must occur to the right of z_0. So we can conclude both that there is a unique root for the first derivative of profits with respect to R&D, and that this root is always in the concave part of the function where profit is maximized.

We can now move to the technology matrix and the substitution matrix both depending on technology:

$$\Gamma_{\,i\,j}(\mathbf{z}) \;=\; \begin{cases} \frac{\delta_{ij}}{z_i^{1-\phi}} & i \neq j, \\ 0, & i = j. \end{cases}$$

$$\Psi_{\,i\,j}(\mathbf{z}) \;=\; \begin{cases} \omega_{ij} z_j, & i \neq j, \\ \frac{z_i}{\phi}, & i = j. \end{cases}$$

We plug in into the final system which is now nonlinear:

$$F(\mathbf{q}) \;=\; \Big[\,2\,I + \Gamma_{\,i\,j}(\mathbf{z})\big)\Big]\,\mathbf{q} \;-\; \phi\,\Psi_{\,i\,j}(\mathbf{z})\mathbf{1} \;-\; (A\,\mathbf{1} - \mathbf{c}) \;=\; 0$$

Since it depends on $R\&D$ on both sides it can't be solved by inverting matrices. We can solve it nonlinearly with the Newton-Raphson method to find the optimal $\mathbf{q}$ and $\mathbf{z}$.

Our cost condition this time cannot be derived outside of the derivation of the quantity. Therefore we'll have to set an additional condition on a component of the quantity function for all the quantities to be positive. We can rewrite $F(\mathbf{q}^*) = 0$ as

$$M(\mathbf{q}^*)\,\mathbf{q}^* \;=\; \phi\,\Psi_{\,i\,j}(\mathbf{z}) \;+\; (A\,\mathbf{1} - \mathbf{c})$$

where

$$M(\mathbf{q}^*) \;= (2\,I + \Gamma_{\,i\,j}(\mathbf{z})\big)$$

To derive only positive quantities we need to have: $M^{-1}(\mathbf{q}^*) \geq 0$ and $\phi\,\Psi_{\,i\,j}(\mathbf{z}) \;+\; (A\,\mathbf{1} - \mathbf{c}) > 0$. The last two statements are already true because $A > c_i$ and $\phi\,\Psi_{\,i\,j}(\mathbf{z}) \geq 0$. For the inverse to be positive we need to make another assumption.

From Neumann Series Theorem we know that If A is any square matrix with $\rho(A) < 1$, then $I - A$ is invertible and $(I - A)^{-1} = \sum_{k=0}^{\infty} A^k$, with the infinite series converging in any operator norm.

For the general B with all positive entries ($B \geq 0$) we set $s > \rho(B)$. Define $M = sI - B = s\left(I - \frac{1}{s}B\right)$. Since $\rho(\frac{1}{s}B) < 1$, the Neumann Series Theorem applies to $I - \frac{1}{s}B$. We now have $\left(I - \frac{1}{s}B\right)^{-1} = \sum_{k=0}^{\infty} \left(\frac{1}{s}B\right)^k$. Because each $\left(\frac{1}{s}B\right)^k$ is entrywise nonnegative, the sum is also nonnegative. Therefore $M^{-1} = \frac{1}{s}\,(I - \frac{1}{s}B)^{-1} = \frac{1}{s}\sum_{k=0}^{\infty} \left(\frac{1}{s}B\right)^k \geq 0$ for all entries. In our case $M = 2I + \Gamma = 2\left(I - \left(-\frac{1}{2}\Gamma\right)\right)$. Set $B = -\Gamma$. Since $\Gamma \geq 0 \Rightarrow -\Gamma \leq 0$, the condition $\rho(B) < 2$ means $\rho(-\Gamma) < 2$, i.e. $\rho(\Gamma) < 2$. Then $M^{-1} = \frac{1}{2}\sum_{k=0}^{\infty} \left(\frac{1}{2}\Gamma\right)^k \geq 0$. Thus the single, strict spectral-radius bound $\rho\left(\Gamma(\mathbf{z})\right) < 2$ guarantees $M = 2I + \Gamma$ is invertible and $M^{-1} \geq 0$.

Putting them together:

$$\mathbf{q}^* \;=\; M(\mathbf{q}^*)^{-1}\,(\phi\,\Psi_{\,i\,j}(\mathbf{z}) \;+\; (A\,\mathbf{1} - \mathbf{c})) \;>\; 0$$

2.2 Optimal Patent Protection

2.2.1 Non-Linear Case

The regulator observes each firm's equilibrium (q_i^*, z_i^*) and chooses the patent-transmission parameter $\phi \in [0, 1]$ to maximize total welfare:

$$W = \sum_{i=1}^{N} \pi_i\big(q^*, z^*\big) + CS\big(q^*, z^*\big).$$

Because product innovation reduces substitutability by a factor $z_i^{1-\phi}$, by increasing the value assigned to the new product portfolio discovered through R&D, the inverse demand (price function) for product portfolio i becomes

$$p_i(q_1,\dots,q_N;z_1,\dots,z_N) = A - q_i - \sum_{j\neq i} \frac{\delta_{ij}}{z_i^{1-\phi}}\, q_j$$

Note that when $\phi = 1$ this collapses to the usual linear form $A - q_i - \sum_j \delta_{ij}\, q_j$.

By definition, consumer surplus is the area under the inverse demand curve above the market price:

$$CS = \sum_{i=1}^{N} \int_0^{q_i} \left[A - x - \sum_{j\neq i} \frac{\delta_{ij}}{z_i^{1-\phi}}\, q_j \right] dx.$$

Carrying out the integral for each i gives

$$\int_0^{q_i} \left[A - x - \sum_{j\neq i} \frac{\delta_{ij}}{z_i^{1-\phi}}\, q_j \right] dx = A\, q_i - \frac{1}{2}\, q_i^2 - \left(\sum_{j\neq i} \frac{\delta_{ij}}{z_i^{1-\phi}}\, q_j \right) q_i.$$

Hence the closed form consumer surplus is

$$CS(q,z) = \sum_{i=1}^{N} \left[A\, q_i - \tfrac{1}{2}\, q_i^2 - \left(\sum_{j\neq i} \frac{\delta_{ij}}{z_i^{1-\phi}}\, q_j \right) q_i \right].$$

Putting profits and consumer surplus together at the equilibrium $\big(q^*(\phi), z^*(\phi)\big)$,

$$W(\phi) = \sum_{i=1}^{N} \pi_i\big(q^*, z^*\big) + CS\big(q^*, z^*\big).$$

The social planner chooses

$$\phi^* = \arg\max_{\phi\in[0,1]} W(\phi).$$

2.3 Simulation NonLinear General Solution with Investment in Differentiation

Now we are going to simulate the general nonlinear solution with investment in product innovation. We start with 10 firms, the model takes the following credible parameters:

1. $N = 10$ **Number of Firms**
2. $A = 1000$ **Demand Intercept**
3. $c \sim \mathcal{N}(25, 5)$ **Marginal Costs Distribution**
4. $\kappa \sim \mathcal{N}(5, 1)$ **R&D Costs Distribution**

To move from 10 firms to 100 and 1000 we need to respect the Neumann Series based stability conditions that we characterized before. If we allow full connection between firms the total similarity matrix (product + technology similarity) becomes unsolvable for positive quantities.

For this reason we are going to increase the number of firms but limit their links at the product and technology similarity level. For each firm we draw from a normal distribution with average 10 links and standard deviation of 5, meaning that every firm will have an average positive similarity at the technology and product similarity level with 10 firms on average, while there will be no connection with the remaining firms both at the product and technology similarity level. For the 1000 firms case we take the following values:

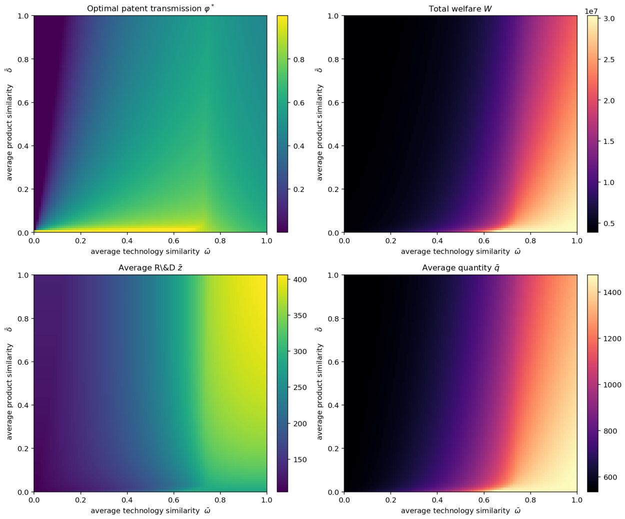


Figure 1: 10 Firms Nonlinear Model - A=1000

1. $N = 1000$ **Number of Firms**

2. $A = 100000$ **Demand Intercept**

3. $c \sim \mathcal{N}(25, 5)$ **Marginal Costs Distribution**

4. $\kappa \sim \mathcal{N}(5, 1)$ **R&D Costs Distribution**

5. $E \sim \mathcal{N}(10, 5)$ **Edges Distribution for Product and Technology Similarity**

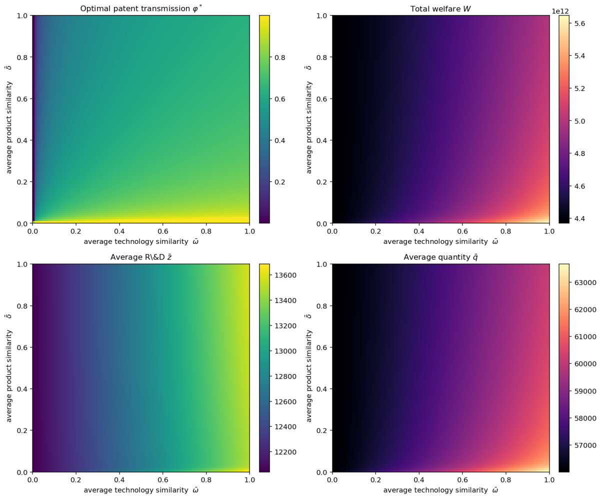


Figure 2: 1000 Firms Nonlinear model

This is our most refined and extended version of our model, which we will test empirically. We will test the amount of R&D invested by each firm to see if it follows the pattern observed in the simulation, with R&D concentrated in the right part of the graph. What we see is an increase in investment purely for competitive reasons. Firms could imitate the improvements of the competitors at no costs, but they decide to invest more in the highest part of technology similarity because other competitors can similarly copy each other improvements, and thus they have to adapt to resist competition. This process happens with high (but not full) values of patent transmission, which as we showed in the model make product differentiation possible.

Regarding product similarity, we are now taking into account the investment in product differentiation. Most of the graph sets an optimal intermediate patent protection, which can be considered as the most realistic framework. What we are seeing in the R&D investment graph is both an investment for product differentiation and for process efficiency. The zone with the highest investment is the extremely high technology similarity zone, where technology similarity gives a neck-to-neck competition in the investment side that requires all firms to invest more. This happens in interaction with product similarity, showing that the combination of the two changes smoothly and jointly.

2.4 Simulation NonLinear General With Separate Investment

It is also possible to compute a model with separated investment for differentiation s and cost reduction z, and have two separate optimal choices for the two types of investment. This framework would take into account more fully informed firms who can expect what type of improvements their investment will yield. Expectedly, the result is really similar, but the distinction between product differentiation and cost reduction can show us more precisely the mechanism that leads to invest more in cost reduction when product similarity is low and technology similarity is high (because of low competition pressure from technology stealing and high synergy) and expectedly for low levels of product similarity there is no need for investing in product differentiation. We see the small change to the theoretical model and then the simulations here:

For a general firm i its profit is

$$\Big[A \;-\; q_i \;-\; \sum_{\substack{j=1 \\ j\neq i}}^{N} \frac{\delta_{ij}}{s_i^{1-\phi}}\, q_j\Big]\, q_i \;-\; \Big[c_i \;-\; z_i \;-\; \phi \sum_{\substack{j=1 \\ j\neq i}}^{N} \omega_{i\,j}\, z_j\Big]\, q_i \;-\; \frac{\kappa_i}{2}\,(z_i^2 + s_i^2).$$

Quantity FOC ($\partial\pi_i/\partial q_i = 0$), rearranged:

$$2\, q_i \;+\; \sum_{\substack{j=1 \\ j\neq i}}^{N} \frac{\delta_{ij}}{s_i^{1-\phi}}\, q_j \;-\; z_i \;-\; \phi \sum_{\substack{j=1 \\ j\neq i}}^{N} \omega_{ij}\, z_j \;=\; A - c_i.$$

The process R&D remains as in the linear case:

$$\frac{\partial\pi_i}{\partial z_i}: \qquad q_i - \kappa_i z_i \;=\; 0 \;\Longrightarrow\; z_i \;=\; \frac{q_i}{\kappa_i}.$$

Differentiation satisfies

$$\frac{\partial\pi_i}{\partial s_i} = (1-\phi)\, q_i \left(\sum_{j\neq i} \delta_{ij} q_j \right) s_i^{-(2-\phi)} - \kappa_i s_i = 0$$

$$\kappa_i s_i^{3-\phi} = (1-\phi)\, q_i \sum_{j\neq i} \delta_{ij} q_j,$$

with the closed form

$$s_i = \left(\frac{(1-\phi)\, q_i \sum_{j\neq i} \delta_{ij} q_j}{\kappa_i} \right)^{1/(3-\phi)}$$

$$\text{when } (1-\phi)\, q_i \sum_{j\neq i} \delta_{ij} q_j > 0, \quad \text{and } s_i = 0 \text{ otherwise.}$$

Because the map $s \mapsto \kappa_i s^{3-\phi}$ is strictly increasing on $(0,\infty)$, the solution for s_i (given q) is unique whenever the right-hand side is positive.

Define the product-interaction matrix

$$\Gamma(s) = \begin{cases} \delta_{ij}/s_i^{1-\phi}, & i \neq j, \\ 0, & i = j, \end{cases}$$

and write Ωz for the spillover vector. We'll have the main equation to solve:

$$\big(2I + \Gamma(s)\big)\, q \;=\; (A\mathbf{1} - c) \;+\; z \;+\; \phi\, \Omega z. \tag{1}$$

From the process FOC, $z_i = q_i/\kappa_i$. In computation we proceed by block iteration as we did before: given (z, s), solve the main equation for q; then update

$$z_i \;\leftarrow\; \frac{q_i}{\kappa_i}, \qquad s_i \;\leftarrow\; \left(\frac{(1-\phi)\, q_i \sum_{j\neq i} \delta_{ij} q_j}{\kappa_i}\right)^{1/(3-\phi)}.$$

Rebuild $\Gamma(s)$ and repeat until convergence. The same argument for the spectral radius smaller than two for the similarity matrix holds here too.

1. $N = 100$ **Number of Firms**
2. $A = 50000$ **Demand Intercept**
3. $c \sim \mathcal{N}(25, 5)$ **Marginal Costs Distribution**
4. $\kappa \sim \mathcal{N}(5, 1)$ **R&D Costs Distribution**
5. $E \sim \mathcal{N}(5, 2)$ **Edges Distribution for Product and Technology Similarity**

3 Data

3.1 Product Similarity

To quantify product similarity I have used the ETNIC dataset from the forthcoming paper Hoberg and G. M. Phillips 2025. The dataset computes product similarity by creating Word2Vec embeddings of product descriptions from 10-K statements filed yearly with the Securities and Exchange Commission. The authors use Word2Vec to avoid look-ahead bias that context conscious embedding models like BERT instead could exhibit. The authors measure the pairwise cosine similarity distance of all firms' 10-K statements, giving a product similarity value for each pair of firm available in the EDGAR website. The **dataset** spans from 1988 to 2023. Companies with more than $10 million in assets and a class of equity securities that is held by more than 2000 owners must file annual and other periodic reports, regardless of whether the securities are publicly or privately traded. The authors have around 50,000 firms fillings available for pairwise cosine similarity every year.

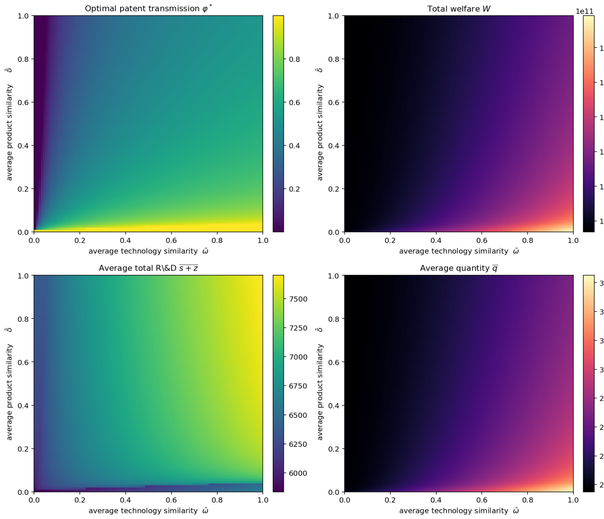


Figure 3: 100 Firms Nonlinear Model with Separated Investment

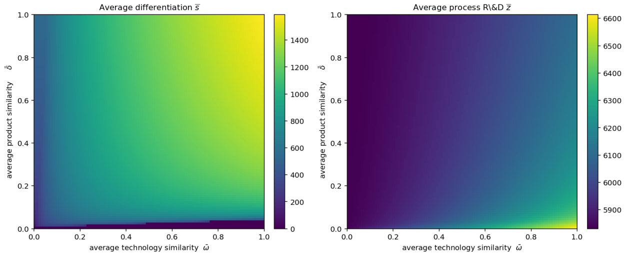


Figure 4: 100 Firms Separated Investments

Using SIC or NAICS for my research purpose would have several limitations. As highlighted in Hoberg and G. Phillips 2010 neither NAICS or SIC adjusts significantly over time as product markets evolve, and neither can easily accommodate innovations that create entirely new product markets.

More generally, fixed classifications like SIC and NAICS have at least four shortcomings: they only rarely re-classify firms that move into different industries, they do not allow for the industries themselves to evolve over time, and they impose transitivity even though two firms that are rivals to a third firm may not compete against each other. Lastly, they do not provide continuous measures of similarity both within and across industries.

The authors observed partial overlap between SIC classification and product similarity scores. I decided to test the dataset's external validity differently by observing the 4 dummy variables effect on pairwise similarity related to SIC of the pair in question. The first dummy variable represents when the pair of firms has the same SIC scores, the second dummy variable when the pair of firms has 3 digits out of 4 in common, the third when the pair of firms has 2 digits out of 4 in common and the fourth when the pair of firms has only one digit in common. As expected we find that, after considering time and industry fixed effects, the dummy variable coefficient is stronger for same SIC code, and then progressively decreases for the 3,2 and 1 code dummy.

Regarding the texts embedded by the authors, they extract the business description section from all the 10-K available in the EDGAR website. The business description is usually called Item 1 or Item 1A in most

10-K. These business descriptions are legally required to be accurate, as Item 101 of Regulation S-K legally requires that firms describe the significant products they offer to the market, and these descriptions must also be updated and representative of the current fiscal year of the 10-K. This recency requirement gives an available document for each year from which is possible to extract a valid annual indicator of each firm product space. A web crawling approach is used to extract only the business section from each 10-K filled in the SEC Edgar site. You can see an example of Nvidia 10K 2025 annual 10-K report at this **link**. In the image below I show the business description of NVIDIA for the year 2025:

3.2 Technology Similarity

I downloaded the abstracts of all patents granted in the US from 1976 to 2025 from the **PatentsView database**. I then used **PatentSBERTa_V2** to assign word embeddings for each patent. This embedding model has been trained on both claims and abstracts for more than 100 million patents and it uses 768 tokens.

Before computing technology similarity between these embeddings, we have to match Compustat firm names to PatentsView assignee names. We use a two step process. We first use regex to strip punctuations and common words in firm names (Incorporated, International etc.) and punctuations. Then we use the sentence encoder **All-mpnet-base-v2** to embed the firm names. In this way we match more firm than previous attempts to match Compustat firms and PatentsView assignees, such as Arora, Belenzon, and Sheer 2021. We match more firms for every year available and we also cover 10 years more since Arora, Belenzon, and Sheer 2021 stops at 2015. The performance of our 2 step matching for firms with available R&D expenditures in Compustat with respect to the previous approach is available in Appendix I. If possible I am going to make public the matching table for future use and manual verification.

After we have the match between Compustat and PatentsView we can create the patent portfolio of each firm for each year. We do so by selecting all the patents granted to a certain firm up to 5 years before the fiscal year for which we have the R&D expenditure amount for that firm. For each firm we then average the abstract embeddings in its patent portfolio to create an object which represents the technological space of the patent portfolio of that firm for that year. The similarity between patent portfolios is the similarity between these two vectors in that year. Each link shares the same similarity for that year, since it is the similarity of the two patent portfolios. For this reason each link appears twice, to observe the effect on both companies considered in our fixed effect regression.

For financial data I have used Compustat. For each firm I have ID, fiscal year, total assets, R&D expenses, SIC code. I have used total assets to compute the HHI index at 4 digit SIC level for every year. The dataset can be considered a weighted network between firms where we have financial data for both firms and the weight of their relationship at the product and technological level. We are going to use two measures of R&D expenditures, namely millions of R&D Expenditures (xrd_start), and $\log(1 + \text{xrd\_start})$.

In the dataset each link appears twice, therefore we also observe the link from the perspective of the NVIDIA competitors, where conm1 and conm2 are inverted. This is because our regression considers xrd_start as their dependent variable, for this reason these links are looking at the effect on NVIDIA of technology and product similarity but we'll also have to see the same effect but for Silicon Image and Red Hat.

We can see how NVIDIA and Red Hat are similar in their technology portfolio but not in the product

gvkey1	gvkey2	year	conm1	conm2	tech_simil	product_simil	at_start	emp_start	xrd_start	sich_start	at_receive	xrd_receive	sich_receive
117768	124599	2007	NVIDIA CORP	SILICON IMAGE INC	0.641619	0.4341	3747.671	4.985	691.637	3674	412.948	77.994	3674
117768	122841	2006	NVIDIA CORP	RED HAT INC	0.640112	0.1548	2675.263	4.083	553.467	3674	1785.854	71.038	7372

Table 1: Firms Similarity Network Dataset

portfolio. This probably happens because Red Hat is an IT consulting company while NVIDIA produces chips, so at the product level they are different. However NVIDIA has also multiple computer science patents for its PyCuda GPU environment, and Red Hat has almost all patents which are computer science based.

4 Results

4.1 Fixed Effect Regressions

We estimate

$$\mathrm{R\&D}_{i,t,j} \;=\; \beta_0\,\mathrm{Tech}_{i,j,t} \;+\; \beta_1\,\mathrm{Prod}_{i,j,t} \;+\; \Gamma_{i,t} \;+\; \zeta_{\mathrm{SIC}(i)} \;+\; \Delta_t \;+\; \epsilon_{i,t,j}. \tag{2}$$

$\mathrm{R\&D}_{i,t,j}$ is firm i's R&D at time t observed on link (i,j); $\mathrm{Tech}_{i,j,t}$ and $\mathrm{Prod}_{i,j,t}$ are pairwise cosine similarities (patents and products). $\Gamma_{i,t}$ collects firm-level controls (assets, employees, concentration, EBITDA) for both endpoints; $\zeta_{\mathrm{SIC}(i)}$ are 4-digit SIC fixed effects; Δ_t are year fixed effects.

Relative to Bloom, Schankerman, and Reenen 2013, we focus on same-year effects (similarity is persistent). With controls only, signs mirror the literature. Once time and sector FE are added, coefficients attenuate toward zero, consistent with the theoretical ambiguity and suggesting latent nonlinearity.

	Dependent Variable: xrd_ start			
	(1)	(2)	(3)	(4)
at_start	0.0072***	0.0068**	0.0080	0.0077
	(1.39e-5)	(0.0019)	(0.0070)	(0.0070)
emp_start	9.328***	9.415**	12.60	12.66
	(0.0049)	(2.999)	(8.632)	(8.626)
at_recv	0.0003***	0.0001*	0.0001	6.48e-5
	(9.43e-6)	(4.56e-5)	(9.92e-5)	(5.55e-5)
emp_recv	-0.0420***	0.0608***	-0.0273	0.0315
	(0.0049)	(0.0147)	(0.0370)	(0.0306)
xrd_recv	0.0012***	-0.0019***	0.0009	-0.0011$^{\cdot}$
	(0.0001)	(0.0005)	(0.0008)	(0.0006)
concentration_start	-293.1***	-276.2***	-136.2	-129.5
	(0.5220)	(50.62)	(90.45)	(83.41)
concentration_recv	-6.205***	9.578**	-4.951	3.458
	(0.5225)	(3.366)	(5.604)	(3.233)
ebitda_start	0.1693***	0.1701***	0.1717***	0.1721***
	(7.58e-5)	(0.0206)	(0.0324)	(0.0323)
product_similarity	323.2***	279.4***	115.6	99.93
	(1.265)	(67.83)	(82.20)	(80.58)
technology_similarity	-143.5***	-115.1*	-81.06	-72.31
	(0.6458)	(50.18)	(93.86)	(91.27)
Fixed Effects				
Year FE	No	Yes	No	Yes
Sector FE	No	No	Yes	Yes
Standard Errors	IID	Year	Sector	Sector
Observations	39,341,495	39,341,495	39,341,495	39,341,495
R-squared	0.592	0.594	0.672	0.673

Note: Robust standard errors in parentheses. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.10$.

Table 2: Similarities on R&D Investment — Levels with FE.

4.1.1 Interaction and Nonlinearity

To capture the joint mechanism emphasized in the model, we add an interaction:

$$\text{R\&D}_{i,t,j} = \beta_0\,\text{Tech}_{i,j,t} + \beta_1\,\text{Prod}_{i,j,t} + \beta_2\left(\text{Tech}_{i,j,t} \times \text{Prod}_{i,j,t}\right) + \Gamma_{i,t} + \zeta_{\text{SIC}} + \Delta_t + \epsilon_{i,t,j}. \tag{3}$$

With FE, Prod is negative, Tech turns positive in semilog, and the interaction is positive and significant, indicating a nonlinear surface consistent with theory, with more recent studies such as König, Liu, and Zenou 2019 and with my two external validity datasets mentioned in the abstract that I have no space to show.

	Dependent Variable: xrd_ start			
	(1)	(2)	(3)	(4)
product_similarity	-68.37***	-54.53	-93.89*	-93.13*
	(2.785)	(34.87)	(40.67)	(40.17)
technology_similarity	-213.7***	-176.3**	-120.6	-109.2
	(0.7842)	(57.81)	(106.9)	(103.8)
product_sim × technology_sim	876.3***	750.2***	473.7**	437.3**
	(5.554)	(93.48)	(178.3)	(168.3)
Fixed Effects				
Year FE	No	Yes	No	Yes
Sector FE	No	No	Yes	Yes
Standard Errors	IID	Year	Sector	Sector
Observations	39,341,495	39,341,495	39,341,495	39,341,495
R-squared	0.59269	0.59442	0.67175	0.67260

Note: Standard errors in parentheses. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.10$.

Table 3: Levels with interaction: product×technology.

	Dependent Variable: log(1+_ xrd_ start)			
	(1)	(2)	(3)	(4)
product_similarity	0.0026	0.0152	-0.0752	-0.0723
	(0.0044)	(0.0556)	(0.0645)	(0.0666)
technology_similarity	0.0888***	0.2298***	0.0483	0.1265**
	(0.0013)	(0.0458)	(0.0493)	(0.0485)
product_sim × technology_sim	1.915***	1.629***	0.7520***	0.6272***
	(0.0087)	(0.1252)	(0.1453)	(0.1465)
Fixed Effects				
Year FE	No	Yes	No	Yes
Sector FE	No	No	Yes	Yes
Standard Errors	IID	Year	Sector	Sector
Observations	39,341,495	39,341,495	39,341,495	39,341,495
R-squared	0.65836	0.66358	0.83314	0.83519

Note: Standard errors in parentheses. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.10$.

Table 4: Semilog with interaction: product×technology.

4.1.2 GLM/PPML Robustness

To handle heteroskedastic $R\&D$, we fit a PPML for levels with log link and a Gaussian GLM for log(1+R&D):

$$\log \mathbb{E}[\mathrm{R\&D}_{i,t,j}] = \beta_0\,\mathrm{Prod}_{i,j,t} + \beta_1\,\mathrm{Tech}_{i,j,t} + \beta_2\,(\mathrm{Prod} \times \mathrm{Tech}) + \Gamma_{i,t} + \zeta_{\mathrm{SIC}(i)} + \Delta_t,$$

$$\mathbb{E}[\log(1 + \mathrm{R\&D}_{i,t,j})] = \beta_0\,\mathrm{Prod}_{i,j,t} + \beta_1\,\mathrm{Tech}_{i,j,t} + \beta_2\,(\mathrm{Prod} \times \mathrm{Tech}) + \Gamma_{i,t} + \zeta_{\mathrm{SIC}(i)} + \Delta_t.$$

Results (two-way clustered SEs) confirm: product similarity ↓, technology similarity ↑, and a positive interaction in the semilog.

	Dependent Variable: xrd_start			
	(1)	(2)	(3)	(4)
product_similarity	-1.418***	-1.315***	-1.084**	-1.093***
	(0.354)	(0.347)	(0.330)	(0.329)
technology_similarity	1.855***	1.264***	1.116***	1.317***
	(0.210)	(0.214)	(0.226)	(0.196)
prod_sim × tech_sim	1.781*	0.528	0.158	-0.424
	(0.733)	(0.722)	(0.775)	(0.752)
Fixed Effects				
Year FE	No	Yes	No	Yes
Sector FE	No	No	Yes	Yes
Std. Errors	Clustered	Clustered	Clustered	Clustered
Observations	38,677,321	38,677,321	38,677,321	38,677,321
Pseudo R^2	0.452	0.485	0.574	0.599

Note: Two-way clustered standard errors (industry × year) in parentheses. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $\cdot p < 0.10$.

Table 5: R&D — PPML (levels) with FE (two-way clustered SEs).

	Dependent Variable: log(1+ xrd_start)			
	(1)	(2)	(3)	(4)
product_similarity	-1.516***	-1.251***	-1.058***	-0.932***
	(0.195)	(0.129)	(0.175)	(0.114)
technology_similarity	0.506	1.072***	0.765***	1.114***
	(0.252)	(0.200)	(0.125)	(0.116)
prod_sim × tech_sim	4.163***	2.218***	2.079***	0.980**
	(0.619)	(0.317)	(0.373)	(0.296)
Fixed Effects				
Year FE	No	Yes	No	Yes
Sector FE	No	No	Yes	Yes
Std. Errors	Clustered	Clustered	Clustered	Clustered
Observations	38,677,321	38,677,321	38,677,321	38,677,321
R-squared	0.230	0.299	0.366	0.415

Note: Two-way clustered standard errors (industry × year) in parentheses. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $\cdot p < 0.10$.

Table 6: R&D — Gaussian on $\log(1 + Y)$ with FE (two-way clustered SEs).

Across FE specifications, we see a general trend of product similarity dampening R&D, technology similarity boosting it, and their interaction being positive. There are some cases where the trends are not significant, probably due to the linear restriction or the different distributional assumptions between Gaussian Log and PPML. This pattern hints to more widespread nonlinearities, and thus motivates for the nonparametric GAM surfaces used later to map the full $(\text{Tech}, \text{Prod})$ $[0,1] \times [0,1]$ space.

4.2 Generalized Additive Model (GAM)

It is clear from the interaction model that the effect of product and technology similarity on R&D is joint and nonlinear. For this reason we are going to use GAM (Generalized Additive Model), a nonparametric technique which will allow use to estimate from the data a nonlinear distribution in the whole space of product and patent similarity.

The model is best explained in S. N. Wood 2017, but I am going to explore it here to show how it is actually perfectly suited to connect my theoretical model with the data. Following the textbook I cited I will start by describing univariate smooth to get the basic idea, then multivariate smooth to cover our case of interacting variables, and then tensor product smooth which will be helpful to add controls and fixed effects.

4.2.1 Univariate smooth

Our dependent variable y (R&D or log(1+R&D) in our case) behaves in relation to the independent variables (product and technology similarity) following an unknown function f plus an error term ε. The model is:

$$y = f(x) + \varepsilon$$

We are going to fit this model with our data in order to identify the functional form of f. We start by identifying basis functions, which will represent function f in its entirety covered by the data. A very simple example is a linear regression, in which the basis functions are just $\beta_2 x$ and β_1. Applying the basis expansion, we have

$$y = \beta_1 + \beta_2 x + \varepsilon$$

In matrix form, we would have:

$$Y = XB + \varepsilon$$

Where Y is an $N \times 1$ column vector with our dependent variable values, X is an $N \times 2$ dependent variable matrix with a column on ones for the constant β_1, B is a 2×1 column vector of model coefficients, and ε is an $N \times 1$ errors column vector.

The same principle applies for basis expansion in the R package *mgcv* (Mixed GAM Computation Vehicle with Automatic Smoothness Estimation). Obviously the basis functions will be more sophisticated than the linear ones we mentioned before. The full model is represented as the sum of the basis functions, each of which is evaluated at every value of the independent variable. It is possible for some of these basis functions to take on a value of zero outside of a given interval and thus do not contribute to the basis expansion outside of that interval.

To be explicit, a basis expansion of dimension $i - 2$ could look like:

$$y = \beta_1 + \beta_2 x + \beta_3 f_1(x) + \beta_4 f_2(x) + ... + \beta_i f_{i-2}(x) + \varepsilon$$

where each function f could be a cubic function of the independent variable x.

The matrix equation $Y = X\beta + \varepsilon$ can still be used to represent our model. The only difference is that X is now an $N \times 1$ matrix, where we include the constant and linear term plus $i-2$ nonlinear basis functions. Since the process of basis expansion has allowed us to represent the model in the form of a matrix equation, we can use linear least squares to fit the model and find the coefficients β. Plugging in the different independent variable x values would then give us the prediction for the dependent variable y. This is how we are going to represent the empirical heatmap, by plugging values in the functions estimated on the area covered by our data. It is an area and not an interval because we are using 2 variables. We will see in the next subsection how it is possible to use GAM also for a combination of variables.

We only saw unpenalized regressions for now. One of the main strengths of *mgcv* is its smoothness estimation via a penalty matrix and smoothing parameter. In other words, instead of:

$$\beta = (X^T X)^{-1} X^T Y$$

we have:

$$\beta = (X^T X + \lambda S)^{-1} X^T Y$$

where S is a quadratic $i \times i$ penalty matrix and λ is a scalar smoothing parameter. λ controls smoothness, with $\lambda = 0$ we come back to the linear OLS. Larger λ penalizes sudden changes so increases smoothness. The S penalty matrix is quadratic in the sense that it is derived from a quadratic penalty based on the independent variable and its basis expansion.

All the smooths covered in this explanation are based on splines. Here's the basic idea following again S. Wood 2010:

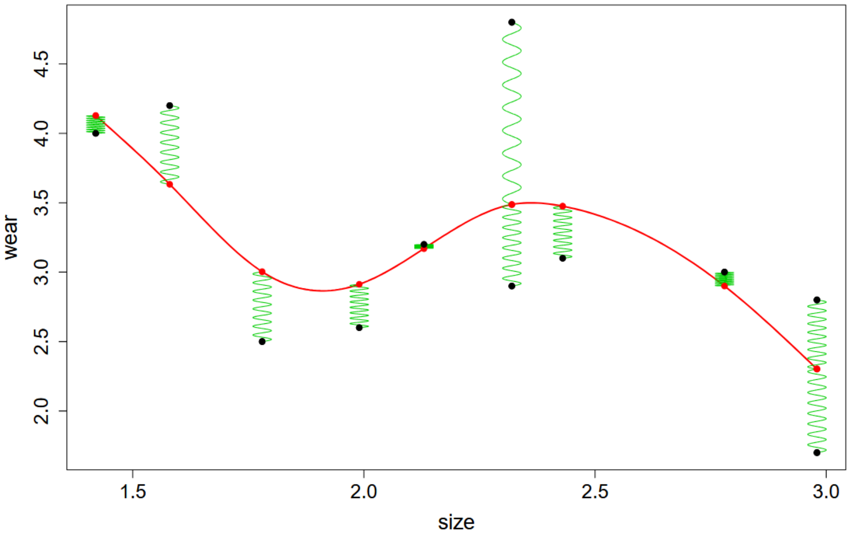


Figure 5: Cubic Spline

Mathematically the red curve is the function minimizing

$$\sum_i (y_i - f(x_i))^2 + \lambda \int f''(x)^2 dx.$$

Splines have variable stiffness. Varying the flexibility of the strip (i.e. varying λ) changes the spline function curve.

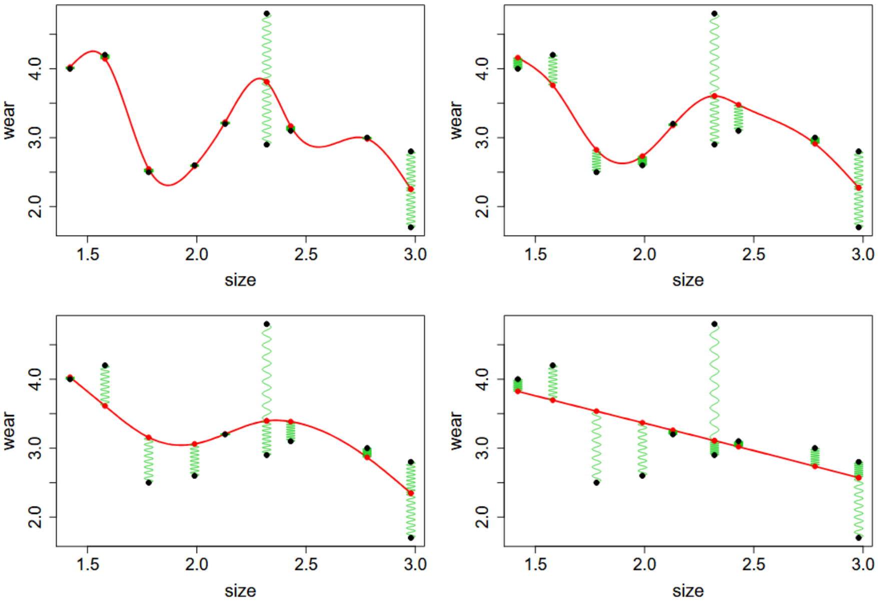


Figure 6: Spline functions with different λ

But irrespective of λ the spline functions always have the same basis. We can produce splines for a variety of penalties, including for functions of several variables, for example:

$$\int f'''(x)^2 dx \quad \text{or} \quad \iint f_{xx}(x,z)^2 + 2f_{xz}(x,z)^2 + f_{zz}(x,z)^2 dxdz$$

. Splines always have an N dimensions basis quadratic penalty representation. If $y_i = g(x_i)$ and f is the cubic spline interpolating x_i, y_i then

$$\max|f-g| \leq \frac{5}{384}\max(x_{i+1}-x_i)^4 \max(g'''')$$

4.2.2 Multivariate smooth

The above explanation can be generalized to multiple dimensions:

$$y = f(x,z) + \varepsilon$$

It should be obvious that at least one of these basis functions must be functions of both x and z (if this was not the case, then implicitly f would be separable such that $f(x,z) = f_x(x) + f_z(z)$). The following image is a visual illustration of multidimensional spline basis can be found here. The sum of these multidimensional spline basis with our values plugged in will be our heatmap:

A full two dimensional basis expansion of dimension $i-3$ could looks like:

$$y = \beta_1 + \beta_2 x + \beta_3 z + \beta_4 f_1(x,z) + ... + \beta_i f_{i-3}(x,z) + \varepsilon$$

Since we have two variables that interact nonlinearly we are going to estimate these multidimensional functions on the available data. In our case x could be the technology similarity and z the product similarity (or viceversa).

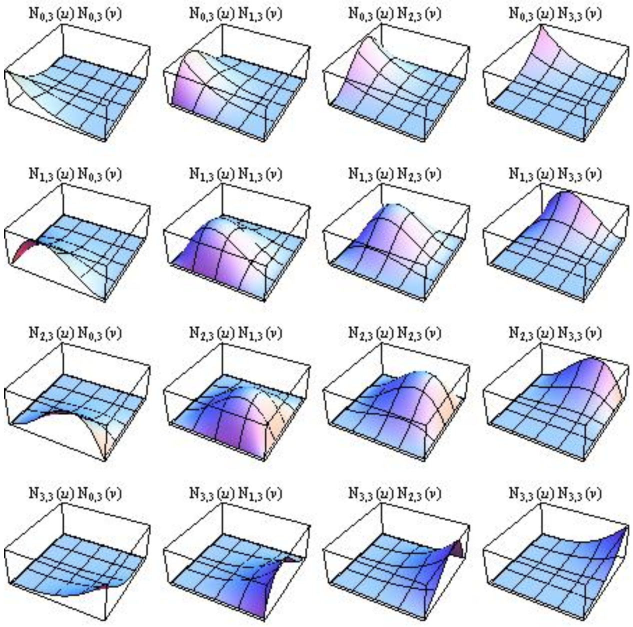


Figure 7: Multidimensional Splines Generated with Wolfram Mathematica

We can still represent this in matrix form with:

$$Y = X\beta + \varepsilon$$

by simply evaluating each basis function at every unique combination of x and z. The solution is still:

$$\beta = (X^T X)^{-1} X^T Y$$

Computing the second derivative penalty matrix is very much the same as in the univariate case, except that instead of integrating the second derivative of each basis function with respect to a single variable, we integrate the sum of all second derivatives (including partials) with respect to all independent variables. This is to say that we can still construct penalty matrix S and use the same method to get the optimal value of smoothing parameter λ, and given that smoothing parameter, the vector of coefficients is still:

$$\beta = (X^T X + \lambda S)^{-1} X^T Y$$

This two-dimensional smooth has an isotropic penalty, this means that a single value of λ applies in both directions. This works when both x and z are on approximately the same scale, such as our cosine similarity similarities for both technology and product similarity, where most combinations lie in the $[0, 1] \times [0, 1]$ interval. But if we have a temporal variable t then its units could be much larger or smaller than the units of x, and this can throw off the integration of our second derivatives because some of those derivatives will contribute disproportionately to the overall integration. The implication here is that multivariate smooths must be constructed from bases supporting multiple variables. Tensor product smooths support construction of multivariate bases from univariate marginal bases.

4.2.3 Tensor Product Smooths

Tensor product smooths provide a sophisticated framework for modeling responses that depend on interactions between multiple inputs measured in different units. Consider a response variable y that is functionally dependent on both a spatial variable x and a temporal variable t, such that the underlying relationship may be expressed as:

$$y = f(x, t) + \varepsilon$$

The objective is to construct a two-dimensional basis that effectively captures the joint dependence on variables x and t. This construction becomes considerably more tractable when the function f can be represented in separable form:

$$f(x, t) = f_x(x) f_t(t)$$

While such separability is not universally attainable, the methodology exploits the fact that within the domain defined by the support of x and t, any non-separable function $f(x, t)$ may be approximated through the interaction of simpler univariate functions $f_x(x)$ and $f_t(t)$. This approximation becomes increasingly accurate as the basis dimensions are chosen to ensure sufficiently fine partitioning of the variable intervals.

The resulting basis expansion, utilizing an i-dimensional basis for x and a j-dimensional basis for t, takes the following form:

$$\begin{aligned} y = {} & \beta_1 + \beta_2 x + \beta_3 f_{x1}(x) + \beta_4 f_{x2}(x) + ... + \beta_i f_{x(i-3)}(x) \\ & + \beta_{i+1} t + \beta_{i+2} tx + \beta_{i+3} t f_{x1}(x) + \beta_{i+4} t f_{x2}(x) + ... + \beta_{2i} t f_{x(i-3)}(x) \\ & + \beta_{2i+1} f_{t1}(t) + \beta_{2i+2} f_{t1}(t) x + \beta_{2i+3} f_{t1}(t) f_{x1}(x) + \beta_{i+4} f_{t1}(t) f_{x2}(x) + ... + \beta_{2i} f_{t1}(t) f_{x(i-3)}(x) \\ & + \cdots + \beta_{ij} f_{t(j-3)}(t) f_{x(i-3)}(x) + \varepsilon \end{aligned}$$

This expansion admits interpretation as a tensor product construction. Consider the evaluation of each basis function over the respective domains of x and t, yielding model matrices X of dimension $N \times i$ and T of dimension $N \times j$. The tensor product $X \otimes T$ produces an $N^2 \times ij$ matrix that, when appropriately reorganized into columns, generates a unique combination for each ij pairing. The marginal model matrices possess i and j columns respectively, corresponding to their basis dimensions, while the bivariate basis maintains dimension ij with an equivalent number of columns in its associated model matrix.

The model may thus be compactly represented as:

$$y = \beta_1 + \beta_2 x + \beta_3 t + \beta_4 f_1(x, t) + \beta_5 f_2(x, t) + ... + \beta_{ij-i-j+1} f_{ij-i-j-2}(x, t) + \varepsilon$$

where each multivariate basis function f represents the product of corresponding marginal basis functions for x and t. This construction preserves the standard matrix formulation:

$$Y = X\beta + \varepsilon$$

with the familiar least squares solution:

$$\beta = (X^T X)^{-1} X^T Y$$

where the model matrix X contains $ij - i - j + 1$ columns.

The penalty structure requires separate construction for each independent variable. The penalty matrices J_x and J_t are formulated as:

$$J_x = \beta^T (I_j \otimes S_x)\beta$$

and

$$J_t = \beta^T (S_t \otimes I_i)\beta$$

This formulation enables anisotropic penalization, namely that penalties applied to the second derivative with respect to x are aggregated across all knots along the t axis, and vice versa. The smoothing parameters λ_x and λ_t may be estimated using extensions of the methodology employed for univariate and multivariate smooths. A key advantage of this tensor product approach is that the overall functional form remains invariant to rescaling of the independent variables, ensuring robust estimation across different measurement scales.

4.2.4 Empirical Strategy for GAM

Let y be either $R\&D$ or $\log(1+R\&D)$. I fit a bivariate smooth $f(\text{Tech}, \text{Prod})$ with controls and FE handled via tensor-product smooths and/or parametric terms. Penalized regression splines with automatic smoothness selection (*mgcv*) avoid functional-form misspecification while respecting the common $[0,1] \times [0,1]$ scaling of both similarities.

4.3 Heatmaps (Core)

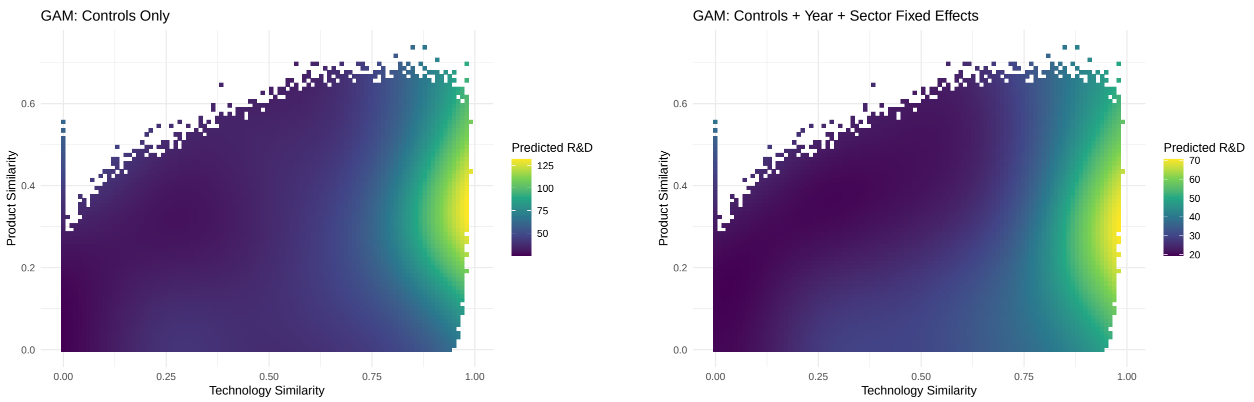


Figure 8: GAM surfaces for $R\&D$ across controls/time/sector FE.

We see that our empirical estimation closely follows the patterns predicted by theory in the nonlinear model, with R&D concentrated in the right part of the graph. The zone with the highest investment is the extremely high technology similarity zone, where technology similarity gives a neck-to-neck competition in the investment side that requires all firms to invest more. This happens in interaction with product similarity, showing that the combination of the two changes smoothly and jointly. Firms could imitate the improvements of the competitors at no costs, but they decide to invest more in the highest part of technology similarity because other competitors can similarly copy each other improvements, and thus they have to adapt to resist competition. This process happens with high (but not full) values of patent transmission, which as we showed in the model make product differentiation possible. We are also going to see in the next subsection how estimated sales also follow closely the quantity pattern predicted by the theoretical model. But first I will show the XgBoost prediction which closely follow the GAM results and my theoretical model:

4.4 Gradient-Boosted Trees (Sanity Check)

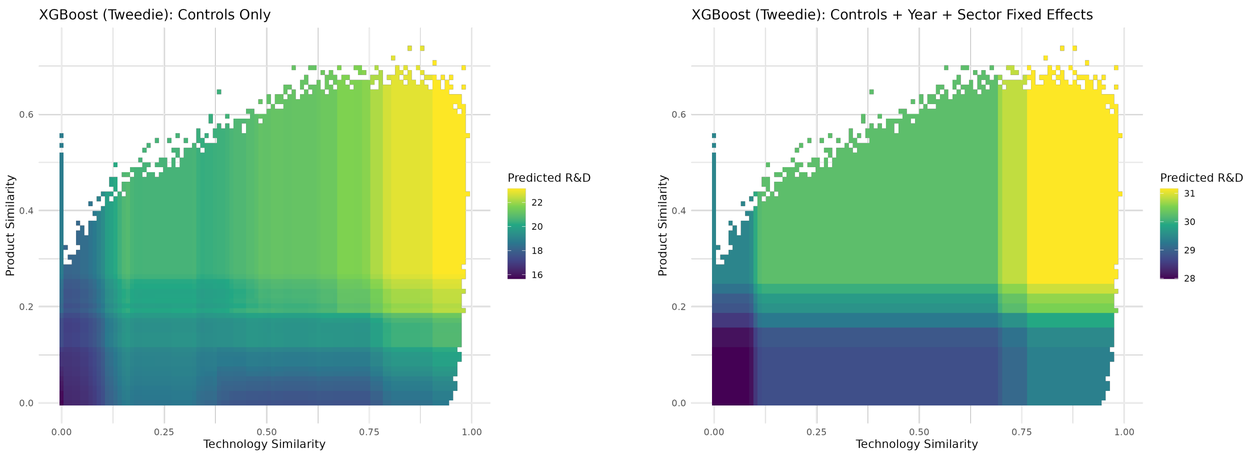


Figure 9: XGBoost partial heatmaps: consistent nonlinear interaction.

4.5 Case Study - Semiconductor Data

I had access to the NSF grant APTO dataset on semiconductor products. This dataset contains around 200,000 semiconductor products with manufacturer names, release year and product description. To capture the position in the product space of each manufacturer for each year I have embedded all the product descriptions with **PatentSBERTa_V2**. For each manufacturer and year I then averaged all the embeddings of the product description produced by that manufacturer up to 5 year before the year in question. The resulting vector could be considered as their product portfolio embedding. I then computed all the pairwise similarities between each manufacturer-year tuple. In this way I have created myself the product similarity measure for this dataset instead of relying upon and external measure. I then matched each manufacturer with the Compustat financial data and their patent portfolio and pairwise technology similarity values. Since I had only around 200 semiconductor manufacturers I was able to run the same machine learning pipeline that I have used for the previous match. For this occasion I set an extremely low threshold for cosine similarity, because I was able to check manually each match between manufacturer name in the APTO dataset and company name in Compustat. The results are available below. We have less observations compared to my previous dataset (around 6000) but we were able to compute the GAM, albeit with less coverage then before. The trend is still holding:

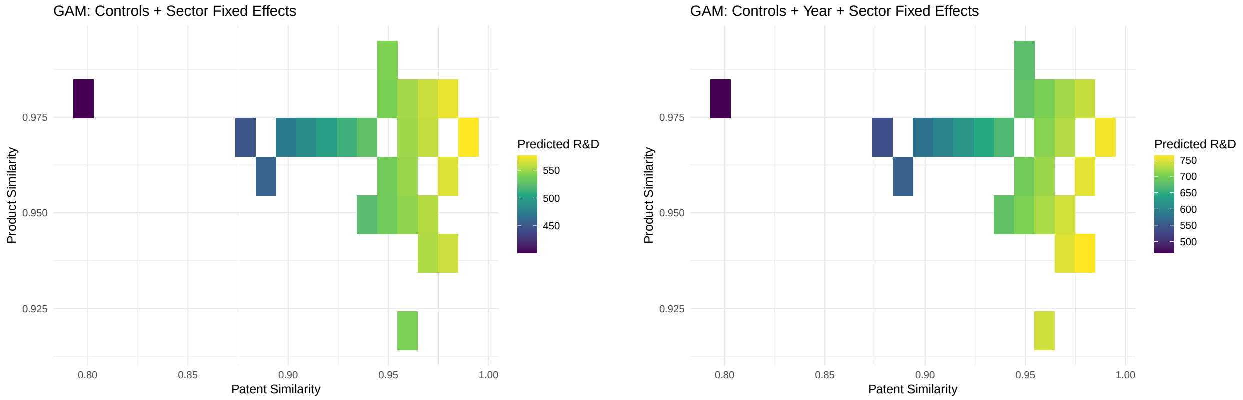


Figure 10: GAM R&D Semiconductor - Levels

5 Estimating Patent Transmission by Sector

Having access to semiconductor data we can estimate the optimal patent transmission precisely for this sector. We have data on quantities (sales), R&D, product and technology similarity. The only unknowns are market size, marginal costs and research costs. Different studies have approached this sector in its early stage, for example Hall and Ziedonis 2001 and Irwin and Klenow 1994.

For each year t I build the firm network and take raw sales $\tilde{q}_{it}$ and raw R&D $\tilde{z}_{it}$ for firms i. Sales and R&D are rescaled once using two global scales $s_q, s_z > 0$. This makes the two measures comparable in the firm choice setting as characterized in the model:

$$q_{it} \equiv \frac{\tilde{q}_{it}}{s_q}, \qquad z_{it} \equiv \frac{\tilde{z}_{it}}{s_z}.$$

The previously estimated Product similarity and Patent similarity give the pairwise matrices $\Delta_t = (\delta_{ij,t})$ and $\Omega_t = (\omega_{ij,t})$ for that year.

Given $\phi \in [0,1]$, product rivalry is attenuated by R&D via

$$\Gamma_{ij,t}(z;\phi) \equiv \frac{\delta_{ij,t}}{z_{it}^{1-\phi}} \quad (i \neq j),$$

and technology spillovers is weighted by patent transmission with $\phi\,\Omega_t z$.

The quantity FOCs can be written as

$$(2I + \Gamma(z;\phi))\,q \;-\; z \;-\; \phi\,\Omega z \;=\; A\mathbf{1} - c.$$

Define the left-hand side as

$$D(\phi) \;\equiv\; (2I + \Gamma(z;\phi))q - z - \phi\,\Omega z.$$

At the true parameters this vector satisfies $D(\phi) = A\mathbf{1} - c$. On the right-hand side we have only the difference $A - c$ and it is identified empirically since we have all the parameters empirically.

From the previous simulation I impose the mean of marginal costs to be 25. This allows me to identify the market scope parameter A in each year as

$$\widehat{A}_t(\phi) \;=\; 25 + \widehat{D_t(\phi)},$$

and the implied firm-level costs as

$$\widehat{c}_{it}(\phi) \;=\; \widehat{A}_t(\phi) - \widehat{D}_{it}(\phi).$$

This procedure delivers the cost distribution consistent with the model and ensures that its mean matches the simulation target.

The R&D FOC implies that now the only unknown is κ_i, and if we keep the assumption of normally distributed research costs we can estimate patent transmission by deriving the values so that the implied distributions of research and marginal costs match the theoretical ones:

$$\widehat{\kappa}_i(\phi) \;=\; \frac{q_i\, z_i^{2-\phi} + (1-\phi)\, q_i \sum_{j\neq i} \delta_{ij} q_j}{z_i^{3-\phi}}.$$

Following the simulations, I target

- $c \sim \mathcal{N}(25, 5)$ **Marginal Costs Distribution**
- $\kappa \sim \mathcal{N}(5, 1)$ **R&D Costs Distribution**

In the first stage I set the weights that I will use throughout the model. I Fix $\phi_{\text{init}} = 0.7$ and choose (s_q, s_z) using only the initial 8 years $(\mathcal{Y})$ to find the weights that would make the two measures approaching the distribution of research and marginal costs:

$$(\hat{s}_q, \hat{s}_z) \;=\; \arg\min_{s_q, s_z > 0} \sum_{t\in\mathcal{Y}} \Big[\big(\overline{\kappa_t(\phi_{\text{init}})} - 5\big)^2 + \big(\text{sd}(\kappa_t(\phi_{\text{init}})) - 1\big)^2 + \big(\text{sd}(\widehat{c}_t(\phi_{\text{init}})) - 5\big)^2 \Big].$$

This step fixes the units once and for all and it does not estimate ϕ_t.

With $(\hat{s}_q, \hat{s}_z)$ fixed, I estimate ϕ_t year by year by matching the same distributions:

$$\hat{\phi}_t \;=\; \arg\min_{\phi\in[0,1]} \Big[\big(\widehat{\kappa}_t(\phi) - 5\big)^2 + \big(\text{sd}(\widehat{\kappa}_t(\phi)) - 1\big)^2 + \big(\text{sd}(\widehat{c}_t(\phi)) - 5\big)^2 \Big].$$

I search on a dense grid over $[0, 1]$ and refine with bounded scalar optimization. At $\hat{\phi}_t$ I also report $\widehat{A}_t$ from the mean restriction above.

The estimator returns a yearly series $\{\hat{\phi}_t\}$, the implied $\widehat{A}_t$, number of firms used, average q and z per year:

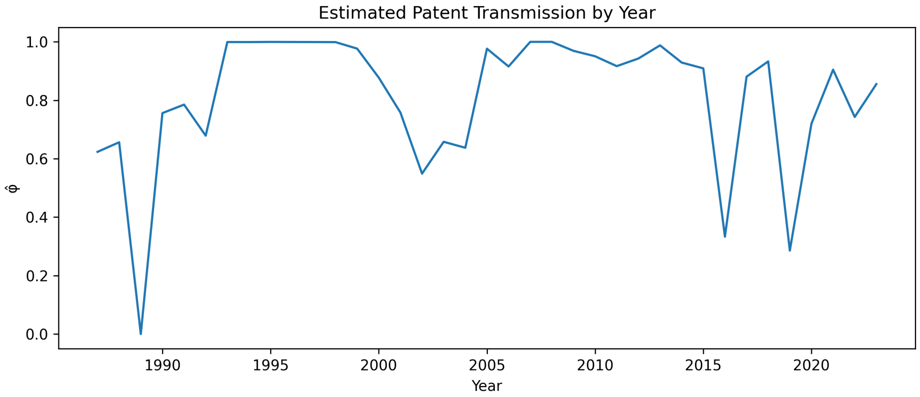
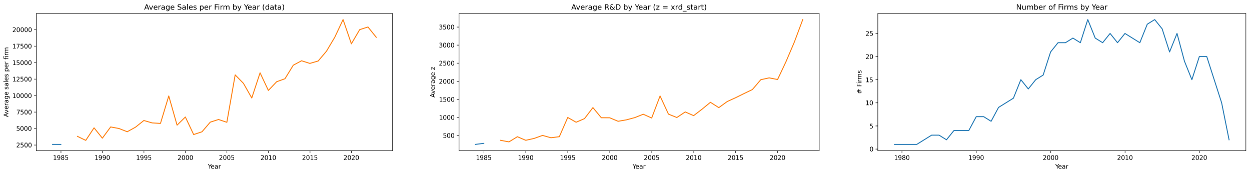


Figure 11: Estimated Patent Transmission By Year



(a) Average Sales by Year

(b) Average R&D by Year

(c) Number of firms by year

The high transmission is in line with the low usage of patents found in the "Yale" and "Carnegie Mellon surveys conducted in 1983 and 1994. We can also observe the reduction of patent transmission happened in the dot-com bubble aftermath, consistent with firms reducing R&D spending growth, enforcing patents more aggressively, and limiting the effective technology transmission. But the general trend remains one of high technological transmission in the semiconductor sector. Further expansion of this study could match the estimated patent transmission of different sector with proxies for sectoral intellectual property enforcement, to externally validate the measure.

6 Discussion

This work wanted to investigate how product and technology similarity act jointly in shaping R&D incentives. The analysis began with a Cournot competition framework where R&D reduced marginal costs through knowledge spillovers weighted by technological similarity and moderated by the regulator's optimal choice of patent transmission. Product similarity entered through the demand system as a measure of substitutability, representing direct rivalry. In the linear baseline, without product differentiation investment, the first order conditions yielded closed form expressions for optimal quantities, R&D investment, and welfare as functions of technology and product similarity.

The framework was then extended to incorporate product innovation on top of the already considered linear process innovation. Product differentiation entered nonlinearly, because its effectiveness in reducing substitutability scaled positively with own R&D and negatively with patent transmission (because with higher patent transmission every firm cannot successfully differentiate its products). This "escape competition" mechanism heightened a component that in the linear model was minor, namely the incentive to invest when product similarity is high to outcompete the other firms. The resulting nonlinear system, with R&D affecting both costs and substitutability, required numerical solution via Newton-Raphson methods. The existence and uniqueness of the investment solution were established, and stability required bounding the spectral radius of the adjusted product/technology similarity matrix. The model generalized from the duopoly to an 1000 firm setting, showing that stability depended on the eigenvalues of the combined similarity matrices. Full connectivity in large N cases could not satisfy the derived spectral radius condition, making bounded link structures a practical requirement for solvable and meaningful equilibria. Every firm draws the number of connections with other firms from a normal distribution with mean 10 and standard deviation of 5.

The dataset was based on Hoberg and Phillips product similarity measures and patent portfolio similarities computed from PatentSBERTa_V2 embeddings of all USPTO abstracts, matched to Compustat financials. Each firm-year observation was connected to its pairwise product and technology similarities with all other firms, along with firm-level R&D, size, profitability, and industry concentration measures.

Linear fixed effects regressions, replicating the structure used in Bloom, Schankerman, and Reenen 2013, produced coefficients on the effect of technology and patent similarity on R&D that became insignificant once time and industry effects were included. An extension with GLM shows a positive effect of technology similarity and a negative effect of product similarity, following the theory. The interaction term was positive and significant only in the Gaussian GLM. To represent fully the interaction in the joint space of product and technology similarity we moved to GAM (Generalized Additive Model).

The Generalized Additive Model addressed this directly by estimating a smooth, bivariate function of product and technology similarity, with controls and fixed effects. The fitted surfaces closely matched the nonlinear model's simulation outputs. We see that our empirical estimation closely follows the patterns predicted by theory in the nonlinear model, with R&D concentrated in the right part (highest technology similarity) of the graph, meaning that in total the positive effect of synergy outweighs the disincentive given by rivalry on R&D. The zone with the highest investment is the extremely high technology similarity zone, where technology similarity gives a neck-to-neck competition in the investment side that requires all firms to invest more. This happens in interaction with product similarity, showing that the combination of the two changes smoothly and jointly. Firms could imitate the improvements of the competitors at no costs, but they decide to invest more in the highest part of technology similarity because other competitors can similarly copy each other improvements, and thus they have to adapt to resist competition. This process happens with high (but not full) values of patent transmission, which as we showed in the model make product differentiation possible. We are also going to see in the next subsection how estimated sales also follow closely the quantity pattern predicted by the theoretical model.

GAM performs a regression throughout the joint space of all variables, including controls and fixed effects. The resulting multivariate smooth can be considered as a $\mathbb{R}^2$ regression over the whole space created by product and technology similarity controlled for the idiosyncratic characteristics of the observed edges in our firm network. The GAM regression behaves like the standard regressions but it can be used to observe the predicted dependent variable throughout the joint space of the two independent variables instead of a line created by one independent variable effect on the dependent variable.

Since our measure of product and technology similarity can be contextual to the measure used for similarities and the subset of firms analyzed, I decided to assure external validity by using two additional datasets. The first one uses an older measure of product similarity based on textual analysis of rare words, a different set of firms matched with patents without machine learning methods, and a previous measure of technology similarity based on Word2Vec embeddings. The results of this dataset follow my main analysis and are available in Appendix II. Another interesting dataset that I have used was the semiconductor manufacturers dataset that I had access through the APTO NSF grant. This dataset was important because it gave us a sectoral analysis and also gave us the occasion to compute product similarity ourselves instead of relying on an external measure. Even if this dataset was smaller, the sectoral case confirmed the concentration of investment in high technology similarity regions where product similarity was sufficient to keep rivals in direct competition but not so high as to eliminate differentiation incentives. The joint effect of product and technology similarity was visible throughout my study, both at the theoretical and empirical level. We can conclude that synergy in the end outweighs rivalry, because technology similarity transmits improvements between different firms that incentivize competition, and also product similarity incentivize to invest more in product differentiation. To avoid killing the incentive to invest in product differentiation our model predicts an upper moderate patent protection of around 0.8 which is enough to distribute technological improvements throughout the economy while also keeping the incentive to invest in product differentiation. The negative rivalry effect of investing less to avoid helping competitors is not large enough to require high level of patent protection in almost all the zones of the economy, excluded the high product similarity low technology similarity zones where product differentiation is still important but the synergy are minor because firms cannot make use of each other improvement.

Finally I also estimated effective patent protection in the semiconductor sector for the period covered (1986-2024). We saw an overall trend of high technological transmission in the semiconductor sector. A future expansion could validate this empirical exercise by correlating it with intellectual property enforcement levels in various sectors.

7 Conclusion

This paper studied how the interaction between product and technology similarity influences firms' R&D investment decisions, extending earlier an earlier framework from Bloom, Schankerman, and Reenen 2013 by modeling their joint and nonlinear effects. I modeled this in a Cournot competition setting with cost-reducing R&D and technology spillovers weighted by technological similarity and moderated by patent transmission. I then extended the model to include nonlinear product differentiation, where the effectiveness of differentiation depends on the level of patent transmission, introducing an "escape competition" mechanism.

The simulations confirmed that ignoring product similarity when setting patent protection omits a key part of the decision space. With product differentiation we saw that high technology similarity zones foster high R&D even with higher permission for imitation from the regulator, because competitive pressure forces firms to invest both in process and in product innovation.

On the empirical side I built a network panel dataset by combining Hoberg and Phillips product similarity measures with patent portfolio similarities computed from USPTO abstracts using PatentSBERTa_V2, matched to Compustat financial data. Linear fixed effects regressions replicated the ambiguous coefficients found in previous literature and highlighted the fact that they are insignificant once sector and time fixed effects are included. GLM regressions indicated that the two similarities act jointly, but parametric forms could not recover the shape of the relationship. The Generalized Additive Model estimated a smooth surface for the interaction between product and technology similarity that matched the simulation results: R&D peaks at high technology similarity and in moderate to high product similarity. I externally validated the results by using two alternative datasets, one with a different subset of firms and older measures of product and technology similarity, and one with a detailed semiconductor dataset from the APTO NSF grant.

The main idea of this paper is that patent protection should be set in a two dimensional space defined by product and technology similarity. A one dimensional approach based only on technological proximity cannot capture the regions where the combination of rivalry and synergy produces the strongest incentives to invest in R&D.