Spreading Out Across Expanding Idea Space*

Ina Ganguli Jeffrey Lin Vitaly Meursault Nicholas Reynolds

March 25, 2026

Abstract This paper investigates how the expanding frontier of knowledge reshapes innovation. If the space of possible inventions grows as knowledge accumulates, inventors could work on increasingly different things, facing less direct competition. However, a growing spread between an invention and its applications could force greater investment in every invention — with implications for whether collective research effort translates into growth. We find that US inventions have indeed become increasingly dissimilar over nearly two centuries (1836–2023), corroborated by 150 years of declining patent interference rates. Documenting this spreading-out requires measuring similarity correctly, as standard text approaches yield the opposite conclusion. Our validation framework, the first systematic comparison for patent text, identifies which representations are accurate and which are misleading. We develop a parsimonious model of positioning in idea space that unifies spreading out with several previously disconnected patterns. Rising research investment, invention quality, patent values, and declining research productivity all emerge from a single spatial mechanism. Other evidence — the coupling of spacing and research investment and quasi-experimental estimates of how proximity shapes knowledge spillovers — can also be explained in this framework. A calibrated decomposition attributes roughly 40% of the decline in research productivity to spatial forces, explaining why productivity growth did not accelerate through the 20th century despite an enormous expansion in aggregate R&D. Where inventors stand relative to each other matters as much for growth as how many of them there are.

JEL classification: O31, C81, L19

Keywords: Invention Similarity, Research Productivity, Natural Language Processing

---

**Author information:* Ganguli, University of Massachusetts Amherst and NBER, iganguli@umass.edu; Lin, Federal Reserve Bank of Philadelphia, jeff.lin@phil.frb.org; Meursault, Federal Reserve Bank of Philadelphia, vitaly.meursault@phil.frb.org; Reynolds, University of Essex, nicholas.reynolds@essex.ac.uk. *Disclaimer*: The views expressed in this paper are solely those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Philadelphia or the Federal Reserve System. Any errors or omissions are the responsibility of the authors. Section 3 and part of Section 4 subsume a prior working paper titled "Patent Text and Long-Run Innovation Dynamics: The Critical Role of Model Selection." First version: December 21, 2023. Acknowledgments appear on p. 52.

1 Introduction

On February 14, 1876, Alexander Graham Bell and Elisha Gray each filed for a telephone patent. Two inventors had independently (by some accounts) arrived at the same invention; years of litigation followed. Such high-stakes collisions were once routine — the US patent office recorded hundreds of so-called "interferences" annually in the nineteenth century, accounting for up to one in twenty issued patents. Over the next 150 years, the interference rate fell steadily by more than 98%. Inventors increasingly work on different things, and the collisions that once marked a crowded "idea space" nearly vanished. This paper explains why inventors spread apart in idea space, measures the spreading across two centuries of patents, and quantifies the cost to growth.

Over nearly two centuries, US inventions have become increasingly dissimilar — not just fewer collisions, but growing distance between neighboring inventions. We document this secular decline using validated neural language models applied to the full text of claims in over 11 million US patents (1836–2023), corroborating over 150 years of declining interference rates. A single spatial mechanism — inventors spreading out across expanding idea space — explains the decline and its consequences for growth. A calibrated decomposition attributes roughly 40% of the long-run decline in US research productivity (Bloom et al. 2020) to forces that arise from inventors' positioning in idea space, with traditional forces (fishing out, burden of knowledge) explaining the remainder.

In the model, inventors choose locations in a circular idea space. Adaptation costs create product differentiation: downstream firms benefit less from more distant inventions, giving spread-out inventors larger territories and pricing power. As the frontier expands and the burden of knowledge grows (Jones 2009), inventors spread out to restore profitability through higher-quality inventions. However, knowledge spillovers attenuate as spacing grows. The model's central tradeoff follows: spreading out raises individual invention quality while reducing aggregate research productivity through quality scaling, spillover attenuation, adaptation costs, and entry expansion. In this way, the model reconciles the secular increases in "breakthrough" inventions (Kelly et al. 2021) and patent values (Kogan et al. 2017) with the secular decline in research productivity.

Measuring proximity in idea space is essential to testing the model, and the choice of method is decisive. Prior work measures similarity using classifications, keywords, citations, or text[1] — each study its own measure, with no systematic comparison and no way to assess

[1]Classifications: Akcigit et al. (2017), Bloom et al. (2013), Fleming (2001), and Jaffe et al. (1993). Keywords: Arts et al. (2025) and Azoulay et al. (2019). Citations: Berkes and Gaetani (2020) and Verhoeven et al. (2016). Text: Feng (2020), Kelly et al. (2021), and Lee

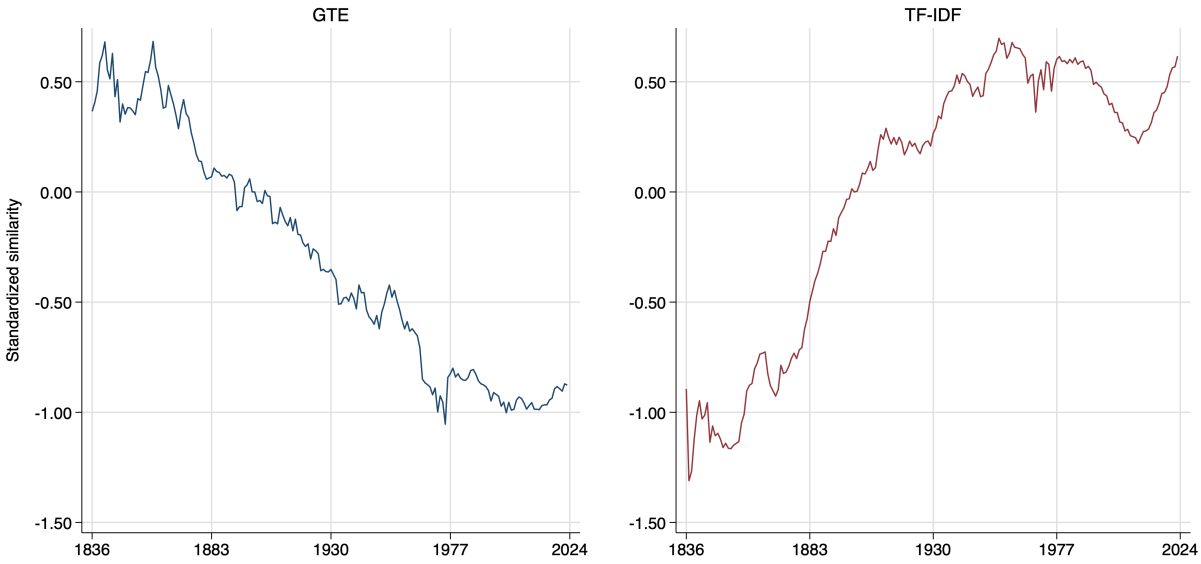


Figure 1: Similarity Trends Depend on Representation Choice

These plots show standardized average pairwise US patent claim similarity by issue year using GTE embeddings (left panel) and TF-IDF representations (right panel). For each representation, changes in similarity are standardized by the cross-sectional standard deviation and normalized to 0 in 1900. For methodological details, see Section 4. The 1.5σ-decline in similarity using the validated GTE representations contrasts sharply with the spurious 1.5σ-increase suggested by TF-IDF.

how sensitive conclusions are to the choice. The stakes are high: as Figure 1 shows, different NLP representations of the same patent text yield opposite conclusions. GTE embeddings show a historical decline in similarity; TF-IDF shows a spurious increase. We provide the first task-specific validation framework for patent text, systematically comparing leading NLP approaches using three independent ground truth tasks: patent interference cases (identical inventions identified by patent examiners (Ganguli et al. 2020)), human annotations of historical patents, and patent office classifications. These tasks span 1850–2023 and test similarity at both fine and coarse levels. GTE and PaECTER embeddings substantially outperform legacy and other modern representations; GTE demonstrates particular strength on historical text, making it our preferred measure for a 188-year analysis. This directly addresses what Ash and Hansen (2023) call the "most important" challenge for text analysis: selecting among black-box NLP models that produce varying economic measurements.[2]

and Hsiang (2019).

[2]Our validated similarity is contemporaneous (cross-sectional), but can also serve as a building block for dynamic measures of novelty, disruptiveness, and breakthroughness (Akcigit et al. 2017; Kelly et al. 2021; Park et al. 2023). Our release of patent GTE and

We test the model's most distinctive predictions. One, the spreading-out prediction is confirmed directly: validated similarity declines across nearly two centuries, at multiple spatial scales, within and between technology classes, and in independently-collected interference data. Two, the model predicts a coupling of invention quality and spacing. In the cross section, patents that are more isolated also have superior measures of quality. Over time, spreading out within technology classes is associated with increased quality. Three, the model predicts that spreading out reduces TFP growth and increases R&D spending. Time-series regressions over 1948–2015 confirm both predictions, with magnitudes consistent with independent quasi-experimental estimates from Bloom et al. (2013) and the variable cost share of R&D from NSF survey data.

Our central theoretical contribution is unification. Spatial models of R&D allocation (Dasgupta and Maskin 1987; Lamantia and Pezzino 2016) characterize location choice in idea space but lack dynamic implications or empirics. Endogenous growth models with expanding variety and rising quality (Howitt 1999; Peretto 1998, 2018) generate rich dynamics but have multiple margins that obscure or elide the core spatial mechanism.[3] Bloom et al. (2013) document that spillovers attenuate over idea space but do not model the equilibrium implications. Meanwhile, several previously disconnected empirical patterns — rising R&D investment per inventor (Hirschey et al. 2012), increasing patent rents (Bessen et al. 2018), and a narrowing wedge between private and social returns to R&D (Lucking et al. 2019) — lack a common explanation. Our framework connects these strands: a single spatial equilibrium jointly determines spacing, quality, pricing, entry, and spillovers, with comparative statics clean enough to identify each channel empirically. We then connect this equilibrium to the research productivity puzzle (Bloom et al. 2020; Jones 2009; Kortum 1997), adding four spatial forces — quality scaling, spillover attenuation, adaptation drag, and entry expansion — to traditional mechanisms.

The model also occupies an unusual position in growth theory.[4] The long-run growth

PaECTER representations for 1836–2023 can serve as a new standard for innovation research, especially useful for studies of positioning and spillovers in idea space (Clancy 2018; Ganguli et al. 2020; Jaffe et al. 1993; Murata et al. 2014; Thompson and Fox-Kean 2005).

[3]Unlike Howitt (1999) and Peretto (1998), our model predicts declining research productivity at both the aggregate and micro level, consistent with the evidence from Bloom et al. (2020). See Section 2.7 for the full argument.

[4]Our static equilibrium characterization follows Kortum (1997), isolating the spatial mechanism with sufficient clarity to calibrate structural parameters empirically; complementary dynamics are explored in Bryan and Lemus (2017), Carnehl and Schneider (2025), and Hopenhayn and Squintani (2021).

rate depends on spatial parameters — adaptation costs, R&D cost curvature, and frontier expansion efficiency — rather than on the level of research effort (Romer 1990) or the rate of population growth (Jones 1995). These parameters identify policy levers absent from both traditions: reducing adaptation frictions (through technology transfer, standardization, or open science) and lowering R&D costs raise long-run growth directly. In short, where inventors stand relative to each other in idea space matters as much for growth as how many of them there are.

2 A Theory of Invention in Idea Space

This section develops a spatial competition model in which inventors choose locations in idea space. Adaptation costs create product differentiation, giving inventors pricing power over local territories. We first characterize the static equilibrium and show that spacing, pricing, quality, and variety are jointly determined (Section 2.1–2.2). We then analyze how equilibrium responds as the knowledge frontier expands and entry costs rise: inventors spread out, invest more in R&D, and charge higher prices — while research productivity declines through six distinct forces (Section 2.3). A knowledge production function endogenizes frontier expansion (Section 2.4). The model unifies three categories of empirical evidence (Section 2.5). We close with a discussion of our theory's key modeling decisions (Section 2.6) and its positioning relative to standard growth models (Section 2.7).

2.1 Model Setup and Key Assumptions

Idea Space and Entry The market for new productivity-enhancing ideas is represented by a circle of circumference $H > 0$ (Salop 1979).[5] This *opportunity space* captures the breadth of feasible technological possibilities at the knowledge frontier; its size H is taken as given in the static analysis and endogenized in Section 2.4. A location on the circle represents a particular technological direction or approach — for instance, different ways to improve battery storage, alternative materials for semiconductors, or distinct architectures for artificial intelligence systems. Nearby locations represent similar projects that both (i) produce closer substitutes for idea consumers and (ii) generate stronger positive knowledge spillovers. In other words, proximity in idea space captures the intuition that similar problems might have similar solutions.

[5]While real-world idea space is surely high-dimensional, this geometric simplification provides analytical tractability and intuitive visualization.

There is a large pool of potential idea producers ("inventors") who make entry, location, pricing, and quality decisions in a simultaneous-move Nash equilibrium. While we call them "inventors," these could represent individuals, teams, or firms; we abstract from the team-formation margin in Jones (2009), which is instead reflected in the fixed cost of entry. ("Entry" refers to undertaking a research project at a location in idea space, not necessarily market entry by a new firm.) We will focus on symmetric equilibria where all inventors are equally spaced and make identical choices. In equilibrium, each inventor optimizes price and quality taking the spatial configuration (spacing d, number of inventors n) as given, while free entry ensures zero profits.

Entry requires a fixed cost $f > 0$. This cost captures the sunk investment needed to reach the knowledge frontier — education, equipment, team assembly. We treat f as a parameter first; entry costs may depend on the size of idea space H, a possibility we formalize in Section 2.3–2.4.

Free entry drives profits to zero, determining the equilibrium number of inventors n.

Idea Consumers Each inventor produces a *non-rival* idea, sold as a non-exclusive license to downstream firms. These firms use ideas as productivity-enhancing inputs in production of consumption goods. Downstream firms are uniformly distributed on the circle with unit mass per unit length, so total mass of potential customers is H. A firm's location represents its specific idea variety needs.

A firm that licenses from inventor i at distance h produces with technology that delivers log total factor productivity:

$$A_i(h) = Q_i - \tau h \tag{1}$$

where Q_i is the realized quality delivered by inventor i (including spillovers from neighbors, as defined below), and $\tau > 0$ measures adaptation cost intensity. Interpreting $A_i(h)$ as log TFP means that baseline productivity without licensing is $\exp(0) = 1$. This linear specification in log TFP naturally captures that firms care about proportional productivity gains and enables direct comparison to empirical TFP elasticities.

The term $-\tau h$ captures the productivity loss from *technological mismatch*: an idea developed for one application typically requires costly adaptation to be useful elsewhere. Adaptation involves organizational costs (coordinating implementation, training workers, modifying production processes) and technical costs (customization, debugging, system integration). The parameter τ measures how quickly productivity declines with technological distance. Empirically, Bloom et al. (2013) find quasi-experimental evidence that spillovers from R&D performed by others to own-firm TFP decline when moving from closely related to moderately distant technologies, consistent with substantial adaptation frictions. Arora et al.

(2021) document that corporate research generates greater value when used internally versus by rivals, further supporting the importance of distance-dependent adaptation costs.[6]

The firm's net surplus after paying licensing fee p_i is:[7]

$$\text{Net Surplus}(h) = A_i(h) - p_i = Q_i - \tau h - p_i \tag{2}$$

Firms choose which inventor to license from to maximize net surplus (equation 2). This creates spatial competition among inventors, analyzed in Section 2.2. Downstream firms use licensed technologies to produce differentiated consumption goods for final consumers.[8]

R&D Technology and Spillovers Inventor i invests in R&D to produce an idea of quality q_i at cost:

$$c(q_i) = \frac{1}{2}\gamma q_i^{1+\eta} \tag{3}$$

where $\gamma > 0$ is a cost scaling parameter and $\eta > 0$ governs the curvature of R&D costs. The parameter η captures the "fishing out" mechanism (Kortum 1997): as the technological frontier advances, producing further increments requires progressively more resources. We

[6]The literature on technology transfer and adoption provides additional evidence for substantial adaptation frictions, consistent with the model's τh term being quantitatively important (Atkin et al. 2017; Hippel 1994; Teece 1977).

[7]This specification is standard in spatial competition models (Salop 1979), where consumers (here, downstream firms choosing which technology to license) have preferences linear in quality net of distance costs. We interpret $A_i(h)$ as log TFP, which naturally captures that firms care about proportional productivity gains. This reduced-form specification allows the model's predictions to be directly compared to empirical TFP elasticities (Bloom et al. 2013) and growth accounting (Bloom et al. 2020). An alternative approximate microfoundation is that each downstream firm has one unit of a fixed input ℓ (specialized capacity, entrepreneurial time, or labor) and produces output $y = e^A \cdot \ell$ where A is log TFP from the licensed technology. With output price normalized to 1 and $\ell = 1$, profit is $\pi = e^A$. Willingness to pay for technology delivering incremental log TFP A (relative to baseline productivity $e^0 = 1$) is $WTP = e^A - 1$. Since the model will be used to describe annual TFP increments ($A \approx 0.015$ per year), the first-order Taylor approximation $e^A - 1 \approx A$ is accurate to within 0.01%, yielding surplus linear in log TFP.

[8]Whether consumers have horizontal preferences over varieties (as in Salop-style models) or love-of-variety CES preferences, firms' willingness to pay for TFP improvements is linear in log TFP increments, consistent with our reduced-form specification. The distinction between horizontal preferences and love-of-variety matters for welfare analysis — entry benefits consumers through improved variety matching — but doesn't affect inventors' equilibrium choices, which depend solely on downstream firms' licensing demand.

set $\eta = 1$ (quadratic costs) as our baseline; Section 6.3 explores alternative calibrations.

However, the *realized quality* delivered to downstream firms incorporates knowledge spillovers from neighbors:

$$Q_i = q_i + \frac{1}{2}\beta\left(1 - \frac{d_c}{\lambda}\right)q_c + \frac{1}{2}\beta\left(1 - \frac{d_r}{\lambda}\right)q_r \tag{4}$$

where q_c and q_r represent R&D of the nearest clockwise and counterclockwise neighbors located at distance d_c and d_r respectively, $\beta \in (0, 1)$ measures spillover intensity, and $\lambda > 0$ governs spillover reach (spillovers vanish beyond distance λ).

Downstream firms benefit from total available knowledge — the inventor's own R&D and spillovers from neighbors. The spillover function $s(d) = 1 - \frac{d}{\lambda}$ (for $d \leq \lambda$, zero otherwise) captures the well-documented attenuation of knowledge flows with technological distance (Bloom et al. 2013; Jaffe et al. 1993). At $d = 0$ (inventors colocated), spillovers are maximized at βq. The parameter λ controls spillover reach. Linear decay ensures symmetric spacing is an equilibrium.[9]

2.2 Equilibrium Characterization

We characterize a symmetric equilibrium where n inventors enter with equal spacing $d = H/n$, and each chooses identical quality q and price p. Each inventor takes neighbors' choices and the equilibrium spacing as given when optimizing.[10]

Downstream firms choose which inventor to license from, balancing quality, price, and adaptation costs. This creates market-stealing competition: when inventor i raises quality or lowers price, they capture customers from neighbors. In symmetric equilibrium, each inventor serves a territory of firms within distance $d/2$ on either side, where the boundary firm is indifferent between neighboring inventors.

---

[9]In symmetric equilibrium, spillovers are a pure positive externality: each inventor receives identical spillovers regardless of own investment, so spillovers do not enter the private zero-profit condition. Spillovers also create no strategic complementarity ($\frac{\partial^2 Q_i}{\partial q_i \partial q_{-i}} = 0$). Bloom et al. (2013) provide quasi-experimental evidence of strategic complementarity; multiplicative spillovers (e.g., $Q_i = q_i(1 + \beta s(d) q_{-i})$) would capture this but require numerical solutions.

[10]We adopt the standard Nash equilibrium framework, where inventors optimize taking rivals' strategies as given (Fudenberg and Tirole 1991). This is the standard approach in the literature on strategic R&D with spillovers (e.g., d'Aspremont and Jacquemin 1988) and is appropriate for a setting with many non-coordinating inventors.

Optimal Pricing and Quality Inventor i chooses price p_i and quality q_i to maximize profit $\pi_i = R_i - c(q_i) - f$, taking neighbors' choices and spacing d as given. The boundary firm at distance $\tilde{h}$ from inventor i is indifferent between inventor i and the neighbor, yielding revenue $R_i = 2p_i\tilde{h}$.

Pricing. With identical realized quality Q in symmetric equilibrium, the indifference condition is:

$$Q - p_i - \tau\tilde{h} = Q - p - \tau(d - \tilde{h}) \quad \Rightarrow \quad \tilde{h} = \frac{d}{2} + \frac{p - p_i}{2\tau} - \frac{q - q_i}{2\tau} \tag{5}$$

Revenue is $R_i = 2p_i\tilde{h} = 2p_i\left[\frac{d}{2} + \frac{p-p_i}{2\tau} - \frac{q-q_i}{2\tau}\right]$. The first-order condition $\partial R_i/\partial p_i = 0$ yields:

$$d + \frac{p}{\tau} - \frac{2p_i}{\tau} = 0 \quad \Rightarrow \quad \boxed{p = \tau d} \tag{6}$$

Quality. Increasing q_i raises realized quality $Q_i = q_i + \beta(1 - d/\lambda)q$, shifting the boundary. Since $\partial Q_i/\partial q_i = 1$ and $\partial\tilde{h}/\partial Q_i = 1/(2\tau)$, the first-order condition $\partial R_i/\partial q_i = \partial c/\partial q_i$ becomes:

$$2p_i \cdot \frac{1}{2\tau} = \frac{p_i}{\tau} = \gamma q_i \quad \Rightarrow \quad \boxed{q = \frac{d}{\gamma}} \quad (\eta = 1) \tag{7}$$

Interpretation: Both price and quality are proportional to spacing. As inventors spread out, they charge higher prices (adaptation costs rise) and invest more in quality (to serve larger territories effectively).

Zero-Profit Condition Free entry drives profits to zero:

$$R - c(q) - f = 0 \tag{8}$$

Substituting revenue $R = pd = \tau d^2$, cost (3) with $\eta = 1$, and quality (7):

$$\tau d^2 - \frac{1}{2}\gamma\left(\frac{d}{\gamma}\right)^2 - f = 0 \quad \Rightarrow \quad \boxed{d^* = \sqrt{\frac{f}{\tau - \frac{1}{2\gamma}}}} \quad (\eta = 1) \tag{9}$$

Equilibrium In symmetric equilibrium, n inventors enter with equal spacing $d = H/n$, and each chooses identical quality q and price p. Equilibrium (q^*, p^*, d^*) as functions of parameters (τ, γ, f) and idea space H are characterized by equations (4), (6), (9). These equilibrium values determine the number of inventors $n^* = H/d^*$, realized quality Q^* (7), and inventor revenue $R^* = p^*d^* = \tau(d^*)^2$. This requires $\tau > \frac{1}{2\gamma}$ for a real solution, which is precisely the condition for spreading out (Proposition 2).

To ensure that this equilibrium exists, we verify in S1.1 that the second-order conditions for pricing and quality are satisfied and that inventors are not incentivized to deviate from the locational equilibrium (no spatial deviation). We also derive technical conditions on parameters ensuring that linear spillovers are active (non-negative) (S1.2) and that all downstream firms adopt from some inventor (full coverage) (S1.3). Quasi-experimental evidence that spillovers are large in recent years (Bloom et al. 2013) suggests that the spillover reach condition is satisfied. Full coverage requires that realized idea quality delivered to downstream firms is sufficiently high compared with price and adaptation costs.

Under these conditions, the zero-profit condition has a unique positive solution, ensuring that the equilibrium is unique (Online Appendix S1). For the remainder of the analysis, we assume that parameters satisfy these conditions.

Proposition 1 (Existence and Uniqueness)**.** *For parameters satisfying $\tau\gamma > 1/2$ (spreading-out condition), spillover reach and full coverage conditions in Online Appendix S1, a unique symmetric equilibrium exists. All downstream firms adopt a technology, and all inventors earn zero profits.*

Proof. See Online Appendix S1. □

Spatial Coupling With existence and uniqueness established, we summarize the equilibrium structure. Four equations characterize the static equilibrium:

$$d^* = \sqrt{\frac{f}{\tau - \frac{1}{2\gamma}}}, \quad p^* = \tau d, \quad q^* = \frac{d}{\gamma}, \quad n^* = \frac{H}{d^*} \tag{10}$$

Both horizontal features (spacing d^*, number of varieties n^*) and vertical features (quality q^*, pricing p^*) are coupled through spatial forces. Spacing depends on costs; price and quality depend on spacing; and the number of varieties depends on idea space H and costs jointly. This coupling is the model's key structural insight: increasing f, decreasing τ, and decreasing γ each independently increase equilibrium spacing, and these changes plausibly reinforce each other — accumulating knowledge raises entry costs while improving tools for technology transfer and experimentation. We focus on entry costs because their trends have the strongest empirical support: inventors train longer, specialize more narrowly, and work in larger teams (Jones 2009). This makes our focus on a single channel conservative.

2.3 Comparative Statics

Entry costs may depend on the size of idea space H — if expanding knowledge raises the cost of mastering a field, then $f(H)$ is increasing. We parameterize this relationship as $f = \phi H^{\alpha}$

with $\phi > 0$, where α governs how entry costs scale with idea space. We formalize frontier expansion in Section 2.4; here we characterize how equilibrium responds to any expansion of H. The predictions differ sharply depending on α. One, if f is constant ($\alpha = 0$), spacing is unchanged and varieties grow linearly in H. Two, if f decreases in H (e.g., an IT revolution reducing entry barriers), inventors cluster closer together and variety expands rapidly. Three, if f increases linearly in H ($\alpha = 1$, our baseline burden-of-knowledge specification), inventors spread out while variety continues to grow. Four, if f is sufficiently convex in H (a "knowledge explosion"), inventors spread out rapidly but the number of inventions may decline. Two observable patterns discipline this choice: inventions are spreading out, and the number of patents has grown steadily. Only scenario three is consistent with both.

2.3.1 Spreading Out

Spreading out $dd/dH > 0$ is the central comparative static. Rising entry costs squeeze profits as H grows; inventors restore profitability by spreading out to capture larger territories. This result follows from differentiating the zero-profit condition with respect to H:

$$\frac{dR}{dd}\frac{dd}{dH} - \frac{dc}{dd}\frac{dd}{dH} - \frac{df}{dH} = 0 \quad \Rightarrow \quad \frac{dd}{dH} = \frac{f'(H)}{\frac{dR}{dd} - \frac{dc}{dd}} \tag{11}$$

Spacing increases with H whenever entry costs rise ($f'(H) > 0$) and marginal revenue of expanding spacing exceeds marginal cost. From $R = \tau d^2$, marginal revenue is $\frac{dR}{dd} = 2\tau d$; marginal cost is $\frac{dc}{dd} = \frac{d}{\gamma}$. The condition $\frac{dR}{dd} > \frac{dc}{dd}$ yields $\tau\gamma > 1/2$, the same condition required for the equilibrium to exist (equation 9). Under the parameterization $f = \phi H^\alpha$, the derivative simplifies to:

$$\frac{dd}{dH} = \frac{\alpha}{2} \cdot \frac{d}{H} \tag{12}$$

Spreading out requires $\alpha > 0$: entry costs must rise with idea space.

The key mechanism is pricing power from differentiation. Adaptation costs create product differentiation: downstream firms face productivity losses from technological mismatch, allowing inventors to charge higher prices ($p = \tau d$) on larger territories. The revenue gain from serving more firms exceeds the cost increase from producing higher quality whenever $\tau\gamma > 1/2$.

Proposition 2 (Spreading Out)**.** *Under $f = \phi H^\alpha$ with $\alpha > 0$ and $\tau\gamma > \frac{1}{2}$, equilibrium spacing increases with idea space: $\frac{dd}{dH} > 0$.*

Other comparative statics follow immediately. R&D investment and idea quality (after spillovers) rise with idea space: $dq/dH = \frac{1}{\gamma}\frac{dd}{dH} > 0$ and $dQ/dH = \frac{1}{\gamma}\left(1 + \beta - \frac{2\beta d}{\lambda}\right)\frac{dd}{dH} > 0$.

Corollary 1. *Rising R&D Investment and Idea Quality per Inventor:* $dq/dH > 0$ *and* $dQ/dH > 0$.

Spacing and quality are jointly determined: $q = d/\gamma$. Wider spacing requires greater R&D investment to serve a larger territory, so rising quality and rising variety coexist. This mechanism differs from quality ladder models (Aghion and Howitt 1992; Grossman and Helpman 1993), where quality rises through vertical replacement on existing product lines. Here, quality growth is a consequence of spatial differentiation, and the step size is endogenous.

Corollary 2. *Spacing-Quality Comovement:* $q = d/\gamma$. *Any force that increases equilibrium spacing also increases equilibrium quality.*

The rewards to invention also rise with idea space. Patent rents grow with idea space: $\frac{dp}{dH} = \tau \frac{dd}{dH} > 0$. Inventor revenue grows quadratically with spacing, driven by both territory expansion and increased pricing power from differentiation: $\frac{dR}{dH} = 2\tau d \cdot \frac{dd}{dH} = \frac{dR}{dd} \cdot \frac{dd}{dH} > 0$.

Corollary 3. *Rising Prices and Revenue per Inventor:* $dp/dH > 0$ *and* $dR/dH > 0$.

A key comparative static is the number of inventions. Since $n = H/d$, rising variety requires idea space to expand faster than spacing. Expanding idea space is essential: without it, there is no equilibrium in which both spacing and variety increase simultaneously. This distinguishes the model from expanding-variety frameworks (Romer 1990) where new product lines appear without any implication for the distance between them. Here, the joint determination of spacing and variety means that growing H does double duty — it creates room for new entrants *and* pulls incumbents apart. The number of inventions rises with idea space: $dn/dH > 0$. This is consistent with the growing number of issued patents — from 500 per year in the 1840s to 350,000 by the 2020s — and unique patent assignees over time, as well as the findings of Hirschey et al. (2012) that the number of R&D-performing firms has grown substantially over time.

Corollary 4. *Rising Number of Inventions. Under the conditions of Proposition 2 with* $\alpha < 2$*, the number of inventions rises with idea space:* $dn/dH > 0$.

Spreading out, rising quality, and rising prices hold for any entry cost function with $f'(H) > 0$. Rising variety additionally requires $\alpha < 2$: if entry costs are sufficiently convex in H, the number of inventions may decline even as spacing grows.

2.3.2 Declining R&D Productivity

A central question in innovation economics is why R&D productivity has declined dramatically. Bloom et al. (2020) document that research effort has risen more than 20-fold since 1930 while total factor productivity growth has slowed — implying a more than 95% decline in research productivity. Our model delivers this result — unsurprisingly, given multiple sources of diminishing returns. The value is not the prediction itself but the decomposition: it reveals new spatial forces (spillover attenuation, adaptation drag, quality scaling, entry expansion) alongside established mechanisms (fishing out, burden of knowledge). Section 6.3 quantifies the relative contributions.

We define two research productivity concepts to capture distinct economic forces. *Per-inventor productivity* ρ measures private returns, determining entry incentives and market structure. *Aggregate productivity* Π measures social returns, determining economy-wide TFP growth per R&D dollar — the measure in Bloom et al. (2020). Both decline with H, but through different mechanisms.

First, define *per-inventor* research productivity ρ as own idea output per own R&D input. Idea output is measured including spillovers. This measures the private return to R&D investment for an individual inventor: the quality they deliver to downstream firms (benefiting from others' research) relative to their own costs.

$$\rho = \frac{Q}{\frac{1}{2}\gamma q^2 + f} \tag{13}$$

Second, define aggregate research productivity Π: aggregate TFP growth delivered to downstream firms per aggregate R&D input. (This corresponds to the measure in Bloom et al. (2020), who compute TFP growth relative to total effective research employment.) Define aggregate R&D spending and aggregate TFP growth[11]:

$$\text{Agg R\&D} = n \cdot [c(q) + f] = n \cdot \left[\frac{1}{2}\gamma q^2 + f\right] \tag{15}$$

$$\text{Agg TFP growth} = Q - \frac{\tau d}{4} \tag{16}$$

[11]The TFP growth delivered to a firm at distance h from its inventor is $Q - \tau h$, where Q is realized quality (including spillovers). The average TFP growth over an inventor's territory of length d is:

$$\text{Average TFP growth} = \frac{1}{d}\int_{-d/2}^{d/2}(Q - \tau|h|)\,dh = \frac{1}{d}\left(Qd - \frac{\tau d^2}{4}\right) = Q - \frac{\tau d}{4} \tag{14}$$

Then:

$$\Pi \equiv \frac{\text{Agg TFP growth}}{\text{Agg R\&D}} = \frac{Q - \frac{\tau d}{4}}{n \cdot [\frac{1}{2}\gamma q^2 + f]} \tag{17}$$

This quantity measures the *social* return to R&D investment. Aggregate productivity Π is lower than per-inventor productivity ρ for two reasons.[12] First, *adaptation costs* reduce effective TFP from Q to $Q - \frac{\tau d}{4}$: downstream firms farther from their idea supplier incur productivity losses from technological mismatch. Second, *entry dilutes aggregate productivity*: total R&D spending scales with the number of inventors n, but average TFP (the intensive margin) does not — entry expands territorial coverage (the extensive margin) without improving productivity at each location. This mirrors the standard monopolistic competition result (Dixit and Stiglitz 1977).[13]

Both productivity measures decline as opportunity space expands. Per-inventor productivity ρ declines because output grows sublinearly (limited by weakening spillovers) while costs — variable R&D plus knowledge burden — grow faster.[14]

The aggregate productivity decline reveals distinct forces. We decompose the derivatives

[12]The relationship is $\Pi = \frac{1}{n} \cdot \frac{Q - \tau d/4}{Q} \cdot \rho$, showing aggregate productivity equals per-inventor productivity adjusted for entry scaling and adaptation costs.

[13]The productivity measures ρ and Π capture *average* returns. Marginal private returns equal marginal costs in equilibrium ($\partial R/\partial q = 0$), but marginal social returns exceed marginal costs due to positive spillovers: when inventor i raises quality q_i, neighbors benefit through $\beta(1 - d/\lambda)q_i$. The marginal social benefit exceeds marginal private benefit by $2\beta(1 - d/\lambda)$, where the factor of 2 reflects spillovers to both neighbors. This classic positive externality implies equilibrium R&D investment is below the social optimum. As inventors spread out (d increases), the spillover externality shrinks — the wedge between private and social returns narrows, but this reflects weakening knowledge flows rather than improved efficiency.

[14]To see this formally, use the zero-profit condition to substitute $\frac{1}{2}\gamma q^2 + f = \tau d^2$:

$$\rho = \frac{Q}{\tau d^2} = \frac{q(1 + \beta - \beta d/\lambda)}{\tau d^2} = \frac{1}{\gamma \tau d}\left(1 + \beta - \frac{\beta d}{\lambda}\right)$$

Since d^* increases with H whenever $f'(H) > 0$ (Proposition 2), and $\partial\rho/\partial d < 0$ from both the $1/d$ term and the declining spillover factor $(1 + \beta - \beta d/\lambda)$:

$$\frac{d\rho}{dH} = \frac{d\rho}{dd} \cdot \frac{dd}{dH} < 0$$

of equations (15) and (16) with respect to H:

$$\frac{d(\text{Agg R\&D})}{dH} = \underbrace{n \cdot c'(q)}_{\text{(1) Fishing out}} \cdot \underbrace{\frac{dq}{dH}}_{\text{(2) Quality scaling}} + \underbrace{n \cdot f'(H)}_{\text{(3) Burden of knowledge}} + \underbrace{\frac{dn}{dH} \cdot [c(q) + f(H)]}_{\text{(4) Entry expansion}} \tag{18}$$

$$\frac{d(\text{Agg TFP growth})}{dH} = \underbrace{\frac{dq}{dH}\left[1 + \beta\left(1 - \frac{d}{\lambda}\right)\right]}_{\text{Quality investment with spillovers}} - \underbrace{\frac{\beta q}{\lambda}\frac{dd}{dH}}_{\text{(5) Spillover attenuation}} - \underbrace{\frac{\tau}{4}\frac{dd}{dH}}_{\text{(6) Adaptation drag}} \tag{19}$$

Six forces drive the decline, four raising R&D costs and two reducing TFP growth.

Forces raising R&D costs:

(1) Fishing out: Convex R&D costs ($c'(q)$, governed by η) mean that each quality increment costs more than the last.

(2) Quality scaling: Spreading out forces inventors to serve larger territories, raising quality investment ($dq/dH > 0$). This is a spatial force.

(3) Burden of knowledge: Rising fixed costs ($n \cdot f'(H)$) from expanding idea space.

(4) Entry expansion in differentiated idea space. The term $\frac{dn}{dH} \cdot [c(q) + f(H)]$ captures entry costs as more inventors cover additional territory. With $n = H/d$ and $\frac{dn}{dH} = \frac{1}{2d} > 0$, each new entrant incurs fixed and variable costs but serves a distinct market niche — horizontal differentiation means new inventors don't generate productivity spillovers to existing territories. This directs R&D spending toward the extensive margin (territorial coverage) rather than the intensive margin (productivity per location), reducing aggregate productivity per R&D dollar.[15]

Forces reducing TFP growth:

(5) Spillover attenuation: $-\frac{\beta q}{\lambda}\frac{dd}{dH}$ — as inventors spread apart, knowledge flows between neighbors weaken.

(6) Adaptation drag. The term $-\frac{\tau}{4}\frac{dd}{dH}$ reflects rising productivity losses as territories expand. Downstream firms located farther from their assigned inventor face larger TFP losses from technological mismatch. With $\frac{dd}{dH} > 0$, the average adaptation cost $\frac{\tau d}{4}$ grows as spacing increases, directly reducing aggregate TFP growth delivered to downstream firms.

[15]This extensive-intensive trade-off mirrors the classic monopolistic competition result (Dixit and Stiglitz 1977): entry increases variety (here, idea-space coverage) but not average quality (here, TFP growth per R&D dollar). Technology adoption involves discrete choice (firms select one production process), so entry creates territorial coverage but no direct variety gains through aggregation as in CES models.

Proposition 3 (Declining Research Productivity)**.** *Both per-inventor and aggregate productivity decline as opportunity space expands:*

$$\frac{d\rho}{dH} < 0 \quad \textit{and} \quad \frac{d\Pi}{dH} < 0 \tag{20}$$

Proof of $d\Pi/dH < 0$*:* Aggregate productivity declines when TFP growth is slower than productivity-adjusted R&D cost growth. Substituting equations (18) and (19) and simplifying using the equilibrium relationships yields the condition $\frac{1}{4} < \frac{1}{\gamma\tau}[1+\beta-\beta d/(2\lambda)]$. Under the full coverage condition (S1.11), $\frac{1}{\gamma\tau}[1+\beta-\beta d/\lambda] \geq 3/2$. Since $1+\beta-\beta d/(2\lambda) > 1+\beta-\beta d/\lambda$, the right-hand side exceeds $3/2 > 1/4$, establishing the result. □

2.4 From Static Model to Dynamics

We now endogenize frontier expansion through a knowledge production function, closing the model.

Innovation expands the frontier: each discovery opens adjacent research directions, and sequential innovation pushes the boundary outward (Scotchmer 1991). (To fix ideas, consider recombinant growth a la Weitzman (1998): new ideas mechanically increase the number of potential recombinations.) Electricity enabled electronics, which enabled computing, which enabled AI — each generation of ideas expanding the space of what could be invented next. We formalize this as a knowledge production function. Each invention's net quality contribution — realized quality Q (including spillovers) minus average adaptation losses $\tau d/4$ — expands the local segment of the idea-space circle it occupies. The sum of these local expansions is total frontier growth:

$$\dot{H} = \delta \cdot n \cdot (Q - \tau d/4) \tag{21}$$

The parameter δ captures the reduced-form efficiency of innovation in opening new research directions — through recombination, demand expansion, institutional creation, or any combination of these channels. The knowledge production function uses social quality Q (including spillovers) rather than private effort q, since spillovers are precisely the mechanism by which individual R&D expands the collective frontier. Only the "usable" part of quality — net of adaptation losses — opens new directions.[16]

[16]This knowledge production function expands the horizontal frontier H (variety of possible research directions) rather than the vertical frontier A (quality of ideas). This horizontal-vs-vertical distinction is the paper's contribution: coupling positioning in idea space with quality investment. In symmetric equilibrium, $n(Q - \tau d/4) = H(1/\gamma - \tau/4)$,

The knowledge production function and entry cost function together generate a self-reinforcing cycle: quality growth expands the frontier, which raises entry costs and spreads inventors apart, which raises quality further.

The combined system yields the growth rate of idea space. In symmetric equilibrium with $n = H/d$, $q = d/\gamma$, and $Q = q(1 + \beta(1 - d/\lambda))$:

$$g_H^* = \delta \left(\frac{1}{\gamma} - \frac{\tau}{4} \right) \tag{22}$$

on the long-run balanced growth path (where $d \geq \lambda$ and spillovers have vanished). This is positive whenever $\tau\gamma < 4$, compatible with the spreading-out condition $\tau\gamma > 1/2$. Spillovers govern the transition to balanced growth, not the long-run rate: during the transition phase ($d < \lambda$, spillovers active), $g_H = g_H^* + \frac{\delta\beta}{\gamma}(1 - d/\lambda)$ exceeds the long-run level and declines as spreading out weakens spillovers. On the balanced growth path, all endogenous variables grow at common rates:

$$g_d = g_q = \frac{\alpha}{2} g_H, \quad g_n = \left(1 - \frac{\alpha}{2}\right) g_H \tag{23}$$

where $g_q = g_d$ follows from $q = d/\gamma$, and $g_n = g_H - g_d$ follows from $n = H/d$. Under the baseline $\alpha = 1$, these simplify to $g_d = g_q = g_n = \frac{1}{2} g_H$. This balanced growth path holds for any admissible parameter values (τ, γ, ϕ) satisfying $\tau\gamma > 1/2$ — this linearity emerges from the model's structure rather than requiring a specific parameter choice.

The long-run growth rate depends on spatial parameters alone. It contains no population growth term, unlike semi-endogenous models (Jones 1995). It is not determined by the level of research effort, unlike fully endogenous models (Romer 1990) — a policy that changes the stock of researchers leaves g_H^* unchanged because spatial crowding adjusts endogenously: $n \times d = H$ and $q = d/\gamma$ absorb the shock through repositioning and quality adjustment, not through the growth rate. Growth is pinned by the organization of research — adaptation cost intensity τ, R&D cost curvature γ, and frontier expansion efficiency δ. The macro structure shares the standard $\dot{H} \propto H$ linearity, but the rate constant is determined by how inventors position themselves relative to each other.

These growth rates map directly to TFP and R&D dynamics. Section 6 tests the model's predictions for TFP growth and R&D spending growth using time-series data and calibrates structural parameters from the model's equilibrium identity and external data.

---

so $\dot{H} = \delta H(1/\gamma - \tau/4)$ — i.e., $\dot{H} \propto H$, the standard Romer (1990) linearity that delivers constant growth rates on a balanced growth path. The macro behavior is conventional; the novelty is in how spatial parameters (τ, γ, spillovers through Q) determine the rate of frontier expansion.

2.5 Unification of Empirical Results

The model's central contribution is unification. Prior work has documented three categories of empirical regularities in isolation; the spatial structure connects them through a single mechanism. The key is the coupling of horizontal and vertical margins: spreading out (horizontal positioning) and rising quality (vertical investment) are joint consequences of spatial competition, not independent phenomena.

One, on positioning and variety (horizontal margins: $dd/dH > 0$, $dn/dH > 0$): spreading out over time (Section 4, Chiopris (2024)), more inventions, more firms, and expanding idea space (Hirschey et al. (2012), Section 5). Two, on quality and returns (vertical margins: $dq/dH > 0$, $dp/dH > 0$): rising R&D investment per firm (Hirschey et al. 2012), rising gross returns to patents (Bessen et al. 2018; Kogan et al. 2017), and rising patent quality (Hall et al. 2005). Three, on productivity decline: decelerating TFP growth and declining research productivity (Bloom et al. 2020).

The horizontal-vertical coupling also reinterprets existing findings. One, Kelly et al. (2021) document a rising rate of "breakthrough" patents — inventions dissimilar from predecessors but similar to successors — consistent with both rising quality (vertical) and expanding reach into new territory (horizontal). (We replicate Kelly et al. (2021)'s breakthrough analysis using our validated GTE measure, with detailed results in Appendix A.) Two, Lucking et al. (2019) find that private returns to R&D rose while social returns declined over 1985–2015, narrowing the wedge between them. The model predicts this: as inventors spread out, spillover reach attenuates ($\beta(1 - d/\lambda)$ falls), shrinking the externality that separates private from social returns (footnote 13).

Natural experiments validate the model's building blocks. Railroad expansion in 19th century Germany — an exogenous expansion of idea markets — led to intellectual divergence rather than convergence (Chiopris 2024). University endowment shocks generate stronger spillovers to technologically similar firms (Kantor and Whalley 2014). Bloom et al. (2013), using R&D tax credit shocks to generate exogenous variation in rivals' R&D, provide quasi-experimental identification of localized spillovers that decay with technological distance — a core mechanism our model assumes.

2.6 Discussion

The most important alternative explanation is heterogeneous idea space. If some regions are more fertile than others, declining similarity could reflect exhaustion of "low-hanging fruit" rather than spatial competition — inventors spread out because productive territory shrinks, not because competitive pressure pushes them apart. We address this empirically. Section 4

shows that similarity declines *within* technology classes, not only between them, ruling out composition effects from the birth of new fields. Interference rate declines spanning 150+ years provide independent corroboration that the pattern is not an artifact of changing patent language. Moreover, neural language models handle vocabulary evolution better than word-counting approaches like TF-IDF, which conflate linguistic change with conceptual change (Online Appendix S7).

The model assumes single-patent firms, but multi-patent firms account for a growing share of patents. Portfolio incentives might work against the model's mechanism: firms internalize cannibalization and diversify their patent portfolios, mechanically reducing measured similarity even absent spatial competition (Champsaur and Rochet 1989; Klemperer and Padilla 1997). This confound potentially explains declining invention similarity. We address it by sampling one patent per firm, isolating spacing between independent entities. The correction *strengthens* the measured decline (Section 4) — the opposite of what one would expect if portfolio diversification drove the raw trend. This suggests that multi-patent firm growth was attenuating, not generating, the decline in similarity. A structural model of portfolio-level spatial competition that fully resolves this puzzle is beyond our scope.

We adopt symmetric equilibrium as a design choice to isolate the spatial mechanism. In symmetric equilibrium, all adjacent distances equal d^*; average pairwise distance is proportional to d^*. The empirical counterpart is average pairwise similarity across patents. The model could apply at both the sector and aggregate levels; the empirics confirm spreading-out patterns within technology classes, between classes, and in aggregate (Section 4). The framework also accommodates time-varying cost parameters: a decline in adaptation costs τ or R&D costs γ from AI-assisted research would reinforce spreading out and raise growth through the same equilibrium formulae.

Specific functional forms ($f = \phi H^{\alpha}$, linear adaptation costs, linear spillover decay) yield closed-form solutions; alternative specifications would preserve qualitative results but could require numerical methods. Linear spillover decay merits comment. In the locational equilibrium, an inventor considering a deviation toward one neighbor gains spillovers from that neighbor but loses spillovers from the other. Linear decay ensures exact cancellation (S1.1). More standard functional forms — exponential decay, for instance — would create a net incentive to deviate in the baseline simultaneous-move game, because convex spillovers reward proximity more than they penalize distance. Under sequential timing (where prices adjust after location choices), adaptation cost curvature can offset this incentive by intensifying price competition between clustered inventors. The qualitative predictions are robust to this substitution.

2.7 Relation to Growth Theory

The model adds the horizontal margin that Jones (2009) lacks. Jones (2009) explains vertical phenomena — why researchers specialize more, train longer, and form larger teams — but does not explain spreading out. Our model connects the two: burden of knowledge drives both horizontal spreading and vertical quality scaling through the equilibrium coupling described above.

Howitt (1999) and Peretto (1998) predict that research productivity should remain constant *within* fields even as it declines in aggregate, because their composition channel operates only at the extensive margin: new varieties enter at average quality, diluting aggregate productivity without affecting incumbents. Bloom et al. (2020) decisively rejected this prediction. Research productivity declines sharply within semiconductors, crops, diseases, and firms. Our model reconciles these micro-level findings with extensive margin expansion. The model applies at the sector level: within any "arc" of idea space corresponding to a technology field, expanding opportunity space causes inventors to spread out, weakening spillovers and generating declining productivity within the field (Section 2.6). The extensive margin then operates as an additional aggregate channel, directing resources toward territorial coverage rather than productivity improvement.

The model's growth rate occupies a third position between existing frameworks (equation (22)). In Howitt (1999) and Peretto (1998), steady-state growth depends on innovation parameters governing quality improvement; in Jones (1995), it depends on the rate of population growth. Our model introduces a third determinant: the spatial organization of research, captured by parameters (δ, γ, τ) that govern how inventors position themselves. The level of research effort adjusts endogenously through free entry and does not appear in equation (22).

The spatial parameters identify policy levers absent from both the endogenous and semi-endogenous growth traditions. Reducing adaptation costs τ — through technology transfer infrastructure, open science mandates, or standardization — raises the long-run growth rate directly. Reducing R&D cost curvature γ — through shared equipment or subsidized research inputs — has a similar effect. During the transition ($d < \lambda$, spillovers active), extending spillover reach λ or increasing spillover intensity β also raises growth, but these levers lose force as spreading out attenuates knowledge flows. Spillover policy is most valuable early; cost and efficiency parameters dominate as the frontier matures.

3 Measuring Distance in Idea Space

3.1 The Measurement Challenge

The theory in Section 2 makes predictions about distance in idea space, but ideas are not directly observable. In this section, we establish how to measure distance in idea space.

Patents are the closest available record of inventive output, and their text — particularly the legal claims defining an invention's scope — provides the richest description of each idea's location. The choice of how to map patent text to a similarity space is consequential. Figure 1 illustrates: using GTE embeddings, we observe the predicted decline in similarity over nearly two centuries, consistent with spreading out. Using TF-IDF representations on identical patent text, we find a dramatic *increase* in similarity, contradicting the theory.

This divergence motivates systematic validation-based model selection. Researchers face an abundance of NLP options, from traditional TF-IDF to sophisticated neural network models. Unlike structural economic models where we can examine functional forms, these models often operate as "black boxes" with complex engineering choices that make *a priori* evaluation difficult. Our solution evaluates multiple NLP approaches against external ground truth measures designed to capture different aspects of technological similarity, moving beyond arbitrary choice to evidence-based selection.

3.2 Data and Representations

We compare multiple approaches for mapping patent text to numerical representations. We denote the mapping of patent text p_i to a location in idea space as $m(p_i) \equiv C_i^m$, where C_i^m represents the coordinate vector based on method m.

A traditional mapping uses patent office technology classifications (Jaffe 1986), which treats all patents within a class as equally similar and all patents across classes as equally dissimilar. NLP methods offer finer granularity.[17] We evaluate traditional frequency-based approaches (TF-IDF) and modern neural embeddings (Doc2vec, USE, S-BERT, GTE, PaECTER, OpenAI).

Frequency-Based Representations The workhorse model TF-IDF (Sparck Jones 1972) represents patents based on word frequency, weighted by inverse document frequency. This approach captures which words are distinctive to each patent but treats words as independent

[17]See Bochkay et al. (2023), Dell (2024), Gentzkow et al. (2019), and Grimmer et al. (2022) for reviews of NLP methods in economics.

and ignores semantic relationships. Kelly et al. (2021) use a variant of TF-IDF to measure "breakthrough" inventions.

Neural Network Embeddings Modern NLP methods produce distributed embeddings that capture semantic relationships. Doc2vec (Le and Mikolov 2014; Mikolov et al. 2013) extends word embeddings to documents. Universal Sentence Encoder (USE) (Cer et al. 2018) produces sentence-level embeddings. S-BERT (Reimers and Gurevych 2019) adapts BERT for sentence similarity. GTE (Li et al. 2023) uses contrastive learning to separate similar and dissimilar texts. PaECTER (Ghosh et al. 2024) is trained on patent data using citation relationships as similarity signals. The engineering choices underlying these models significantly impact performance but are often poorly documented (see Online Appendix S2), making empirical validation essential.

3.3 Validation Framework

Our framework evaluates multiple NLP representations using three complementary tasks: patent interference cases (patent office determinations that independent inventors made identical discoveries), non-expert human similarity judgments on historical patents, and patent office classifications. These tasks complement each other across key dimensions: they span 1850–2023, capture different similarity levels (identical inventions to broad categories), and incorporate different expertise types (patent examiners, lay annotators, institutional classifications). For more discussion, see Online Appendix S3.

For each validation task j, we evaluate how well similarity measures from representation m align with ground truth:

$$V^j(m) = S^j\left(1 - d^m(\mathbf{p}), g^j(\mathbf{p})\right) \tag{24}$$

where $1-d^m(\mathbf{p})$ measures similarities using representation m (via cosine similarity $\frac{C_i^m \cdot C_j^m}{||C_i^m|| ||C_j^m||}$), $g^j(\mathbf{p})$ provides ground truth, and S^j quantifies correspondence using, e.g., ROC AUC. No single ground truth exists for "similarity." In the theory, distance in idea space governs both competitive pressure and knowledge spillovers — a useful measure must capture multiple dimensions of technological proximity. A model performing well on one task could fail on another, so multi-task evaluation across the dimensions in our framework is essential.

3.4 Validation Results and Model Selection

Interferences Patent interferences — US patent office proceedings determining that independent applications describe identical inventions — provide the most granular ground truth: exact identity in idea space. A specialized examiner initiated an interference upon encountering applications containing the "same patentable invention." We use 215 cases decided between 2001 and 2014 (Ganguli et al. 2020), producing 96,580 application pairs of which 322 are interfering pairs. GTE and PaECTER dominate: PR AUC of 0.64 and 0.65, respectively, with nearly identical F10 scores of 0.90. TF-IDF trails at 0.45 PR AUC — 30% worse than the top models. The top models also generate 2.8–4.7 times fewer false positives than TF-IDF while maintaining high detection rates. Detailed metrics, tables, and threshold analysis appear in S4.1.

Human judgment Non-expert annotators made relative similarity judgments on historical patents (1850–1975, oversampling 1880–1920), testing temporal robustness on text spanning nearly two centuries. Annotators compared two patent pairs and judged which pair was more similar — a relative task, because humans struggle to place similarity on absolute scales. GTE demonstrates the strongest alignment with human judgment ($\beta_1 = 0.62$), substantially outperforming PaECTER (0.51), S-BERT (0.54), and TF-IDF (0.35). GTE's 22% advantage over PaECTER on historical text is particularly important for our 188-year analysis. Detailed task design and regression results appear in S4.2.

Classifications and why task variation matters Different tasks reward different model strengths: interferences test fine-grained identity, human judgments test continuous similarity on historical text, classifications test categorical boundaries. Classifications (Cooperative Patent Classification or CPC assignments, 1850–2023, coarse and fine granularity) illustrate this: S-BERT leads on classification despite trailing on the other two tasks, likely because classifications emphasize administrative utility rather than semantic proximity. A model excelling at classification may fail at continuous proximity — the concept our theory requires. Only models performing consistently across all dimensions are reliable for our application. Detailed classification results appear in S4.3.

Note on LLM-Based Validation We explored using Large Language Models (Claude 3.5 Sonnet and GPT-4) as a scalable alternative to human annotation. However, LLMs showed notable disagreement with human annotators and with each other: Claude selected GTE as best-performing, while GPT-4 chose S-BERT. See Appendix S6 for detailed results.

Table 1: Validation Results Summary: Model Performance Across All Tasks

Model	Interferences		Human	Classifications	
	PR AUC	F10	Agreement[a]	Section[b]	Class[b]
GTE	0.640	**0.899**	**0.62**	0.596	0.656
	(2)	(1)	(1)	(2)	(3)
PaECTER	**0.654**	0.897	0.51	0.590	**0.672**
	(1)	(2)	(3)	(3)	(1)
S-BERT	0.517	0.816	0.54	**0.600**	0.671
	(3)	(3)	(2)	(1)	(2)
TF-IDF	0.448	0.765	0.35	0.514	0.525
	(4)	(4)	(4)	(4)	(4)

[a] Coefficient from regression of human judgments on model rankings (higher = better agreement)

[b] ROC AUC for predicting same top-level section or same three-character class

Note: Rankings in parentheses. Bold indicates best performance for that metric. Interference task uses 2001–2014 data, human annotations use 1850–1975 data, classifications use 1850–2023 data.

Results Summary Table 1 summarizes performance of four models across all validation tasks. This summary excludes models that performed poorly on the interferences task (Doc2vec, USE) as well as OpenAI, which performed similarly compared with non-proprietary GTE and PaECTER. Complete results are reported in Online Appendix S4

The race between GTE and PaECTER is close on interferences (PR AUC 0.64 vs. 0.65, F10 0.90 vs. 0.90) and classifications (GTE ranks second or third). GTE's decisive advantage is temporal robustness: its 22% outperformance on historical text ($\beta_1 = 0.62$ vs. 0.51) is uniquely important for a 188-year analysis. PaECTER yields qualitatively similar results for the post-1900 period — both show declining similarity — which strengthens this conclusion. TF-IDF ranks last on every task, with 20–40% lower performance; it would lead to opposite conclusions about similarity trends.

We select GTE as our primary representation for three reasons: temporal robustness on historical text, near-best performance on interferences, and consistent performance across all tasks. We use PaECTER and S-BERT as robustness checks. We explicitly avoid TF-IDF despite its transparency and widespread use.

An important methodological point is that single-model validation would not detect TF-IDF's failure. Every model we evaluate — including TF-IDF — beats random chance on every task. A researcher validating TF-IDF alone would conclude it "works." Only comparative evaluation across multiple candidates reveals that TF-IDF systematically underperforms and would produce opposite conclusions about similarity trends.

TF-IDF fails because it treats words as independent tokens with no semantic structure. A 19th-century "velocipede" patent and a 21st-century "bicycle" patent describe the same invention, but TF-IDF assigns them zero similarity because they share no vocabulary (Online Appendix S7). More generally, TF-IDF overweights period-specific language: the unigrams characteristic of high-TF-IDF patent pairs are more volatile over time than those characteristic of high-neural-embedding pairs (Online Appendix S7 documents this using S-BERT, which shares relevant properties with GTE as both are neural embedding models), meaning TF-IDF measures when patents were written rather than what they describe. Neural embeddings avoid this failure by assigning similar vectors to words that appear in similar contexts, capturing semantic relationships across vocabulary and time. Visualizations of the resulting similarity spaces confirm this: in projected embeddings, semiconductor patents are positioned between materials science and electrical engineering clusters, reflecting their hybrid nature, while TF-IDF produces diffuse, less structured representations (Online Appendix S8).

4 Inventors are Spreading Out across Idea Space

Proposition 2 predicts that equilibrium spacing increases as idea space expands ($dd^*/dH > 0$). In symmetric equilibrium, average pairwise distance is monotonically increasing in d^*. Declining average pairwise similarity is the empirical counterpart of increasing d^*, under the maintained assumption that cosine similarity in the embedding space is a monotone transformation of idea-space distance. The spatial-scale analysis in Section 4.3 further validates this mapping by examining similarity at quantiles of the distance distribution, not just the global average.

We use the full text of claims in all US utility patents issued 1836–2023. For historical patents (1836–1975), we use digitized patent text from the Patents Core database by ProQuest. For modern patents (1976–2023), we use patent text from PatentsView (U.S. Patent and Trademark Office 2023). We focus on patent claims rather than abstracts or descriptions because claims define the precise boundaries of what each patent covers, making them most relevant for measuring technological similarity. We measure similarity using GTE embeddings, selected in Section 3 for superior performance across all three validation tasks. We show robustness using PaECTER and S-BERT embeddings, which also performed well.

For each year, we compute average pairwise cosine similarity across all patents.[18] To

[18]We use an efficient computational method that reduces complexity from $O(N^2)$ to $O(N)$ for unit-normalized vectors, detailed in Online Appendix S9. To estimate cross-sectional standard deviations, we subsample up to 5,000 patents per year; in years with fewer patents,

enhance comparability across representations, we standardize similarity measures by dividing by the cross-sectional standard deviation in each year. This standardization is important because different NLP representations have unknown scaling with no easily interpretable economic meaning (Bergeaud et al. 2025); standardizing by the cross-sectional standard deviation allows us to compare magnitudes of change across different embedding spaces. The cross-sectional standard deviations prove relatively stable over time for each representation, and using a time-invariant global standard deviation yields nearly identical quantitative results.

Using GTE, we document substantial secular decline in patent similarity from 1836 to 2023. We then examine robustness to multi-patent entities, spatial scales, and within versus between technology class comparisons. Finally, we corroborate these findings using an entirely independent data source: patent interference rates spanning 1836–2014.

4.1 Main Finding: Declining Similarity

Figure 2 shows average annual pairwise patent similarity using GTE, PaECTER, S-BERT, and TF-IDF (each series is indexed to 0 in 1900).[19] Our best validated embeddings, GTE, exhibit a clear and consistent secular decline in patent similarity from 1841 through the late 20th century. The trend is gradual but substantial: minimum similarity is approximately 1.5 standard deviations (σ) below the historical maximum, indicating that contemporary patents have become markedly less similar to each other over nearly two centuries.

Our main empirical finding is clear: average pairwise patent similarity declined substantially and consistently from the early 19th century through the late 20th century. This pattern provides strong support for our theoretical prediction that inventors spread out over an expanding knowledge frontier as the burden of knowledge increases. Average pairwise similarity declined about 1.5σ 1836–1980. The rate of decline moderated after 1980. GTE similarity reached its minimum around 1999 before partially retracing — a pattern we attribute to multi-patent entity growth (Section 4.2). Section 6.3 shows that this spreading out accounts for roughly 40% of the long-run decline in US research productivity documented by Bloom et al. (2020).

In Figure 2, PaECTER suggests declining similarity for nearly a century with a partial retracing after 1999, S-BERT shows a more consistent decline from 1900 to 2023, and TF-IDF exhibits a strikingly different pattern that contradicts our theoretical predictions. GTE,

---

we use all available patents.

[19]There is evidence of a slight discontinuity coincident with the change between the ProQuest corpus (pre-1976) and the PatentsView corpus (1976–2023).

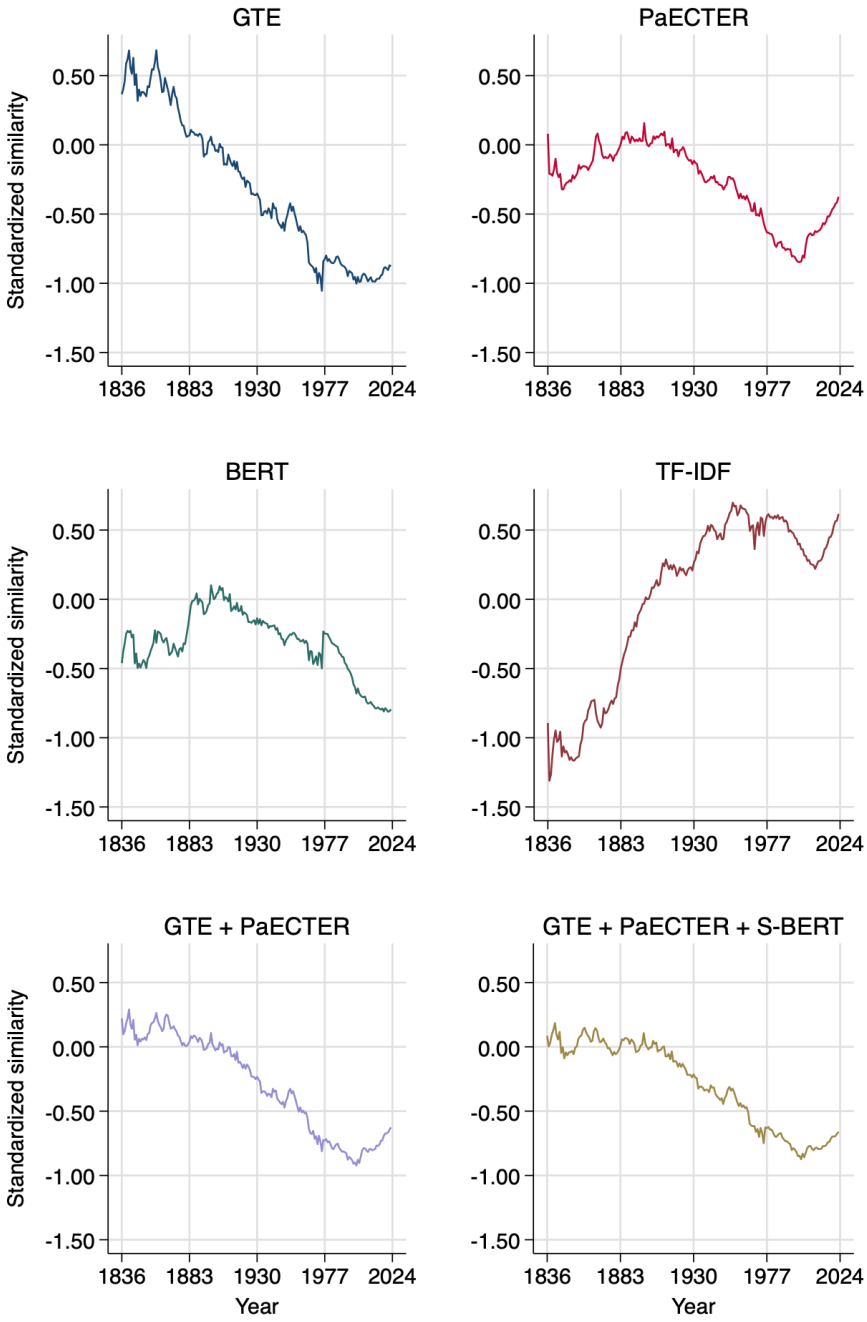


Figure 2: Similarity by Year and by Representation

These plots show standardized average pairwise US patent claim similarity by issue year and by representation. For each representation, changes in similarity are standardized by the cross-sectional standard deviation and normalized to 0 in 1900. See Appendix S9 for methodological details and additional results. The top left panel (GTE) shows our main finding of secular decline. Other models (PaECTER, S-BERT, TF-IDF) show various patterns, discussed in Section 3. The bottom panels show ensemble estimates, discussed in Section 3.4.

PaECTER, and S-BERT all show consistent declines from 1900 to 2000 of about 0.8–1.0σ. This can be seen in the ensemble measures — averaging across models — in the bottom panels.

4.2 Accounting for Multi-Patent Entities

GTE representations show a notable pattern: declining similarity arrests around 1999, with a slight retracing beginning 2013–2014. Intriguingly, this timing coincides with well-documented phenomena in patent economics: the surge in business method patents following the 1998 State Street Bank decision (Hall 2009), the proliferation of non-practicing entities ("patent trolls") in the early 2000s (Cohen et al. 2019), and the rise of defensive patenting (Hall and Ziedonis 2001). These developments led to rapid growth in the number of patents per entity, raising a concern for our analysis: if single entities are filing many similar patents, our measure of contemporaneous similarity may conflate within-entity and between-entity similarity.

Our theoretical framework focuses on strategic positioning choices by independent inventors facing competition and seeking spillovers from others. Similarity among patents filed by the same entity likely reflects different economic forces. Multi-product firms occupy intervals of technology space rather than points, reflecting defensive fencing around core inventions (Hall and Ziedonis 2001) and local market segmentation — portfolio strategies rather than independent competitive positioning. The one-patent-per-entity correction recovers the relevant single-inventor margin predicted by our model.

Methodology We implement two complementary approaches to address multi-patent entities. Our primary approach uses the PatentsView disambiguation algorithm (Monath et al. 2021), which assigns consistent identifiers to patent assignees and individual inventors 1976–2023.[20] Each disambiguated assignee or individual inventor represents an "entity."

Figure 3 reveals the scale of the issue: after 1999, the number of entities grew far more slowly than the number of issued patents, indicating a substantial increase in patents per entity. This divergence is precisely when GTE shows arrested decline in similarity.

To isolate between-entity similarity, we randomly sample one patent per entity per year and recompute average pairwise similarity. This ensures that each similarity calculation represents the distance between independent inventors rather than multiple patents from the same entity. We repeat this sampling procedure multiple times to ensure robustness.

[20]We use the 2025Q1 vintage. For unassigned patents, we assign identifiers based on individual inventors.

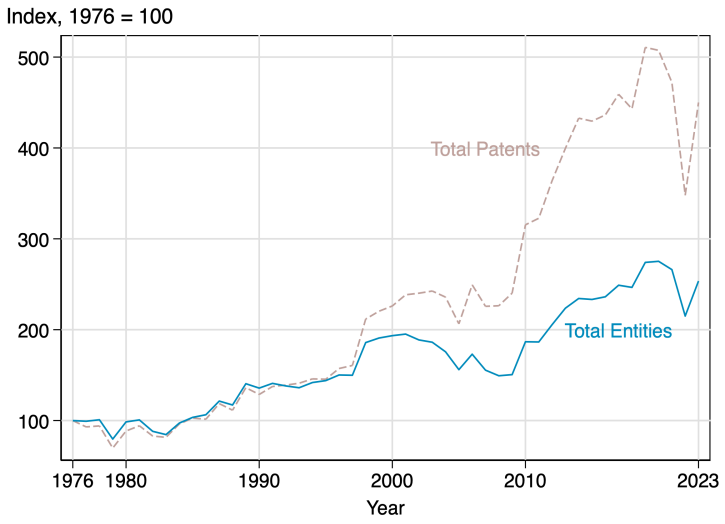


Figure 3: Growth in Patents and Patenting Entities

This figure shows the number of issued utility patents and unique patenting entities per year. The divergence after 1999 indicates substantial growth in patents per entity, likely driven by business method patents and non-practicing entities.

Our secondary approach uses the KPSS (Kogan et al. 2017) disambiguation of patents issued to publicly-traded firms, updated through 2023. While this covers only public firms, it extends back to 1926, providing a longer historical perspective than the PatentsView data (which begins in 1976).

Results Figure 4 shows similarity trends after correcting for multi-patent entities. The correction based on PatentsView disambiguated entities reduces the arrest in declining similarity around 1999. When we account for the fact that individual entities are filing multiple similar patents, the underlying trend of inventors spreading out over idea space continues more consistently through the 2000s and 2010s.

Our theory predicts spreading-out among independent inventors, not within entities pursuing portfolio strategies. The fact that entity-corrected measures show continued decline validates our theoretical mechanism.

Importantly, the divergence between patent counts and entity counts emerges sharply around 1999 and accelerates through the 2000s. This timing is well within the PatentsView disambiguation sample period, meaning our entity corrections directly address the period when multi-patent strategies became most prevalent. The fact that entity corrections meaningfully affect similarity trends precisely when and where we would expect — in the post-1999 period — provides confidence that the disambiguation is capturing real economic phenomena rather than measurement artifacts.

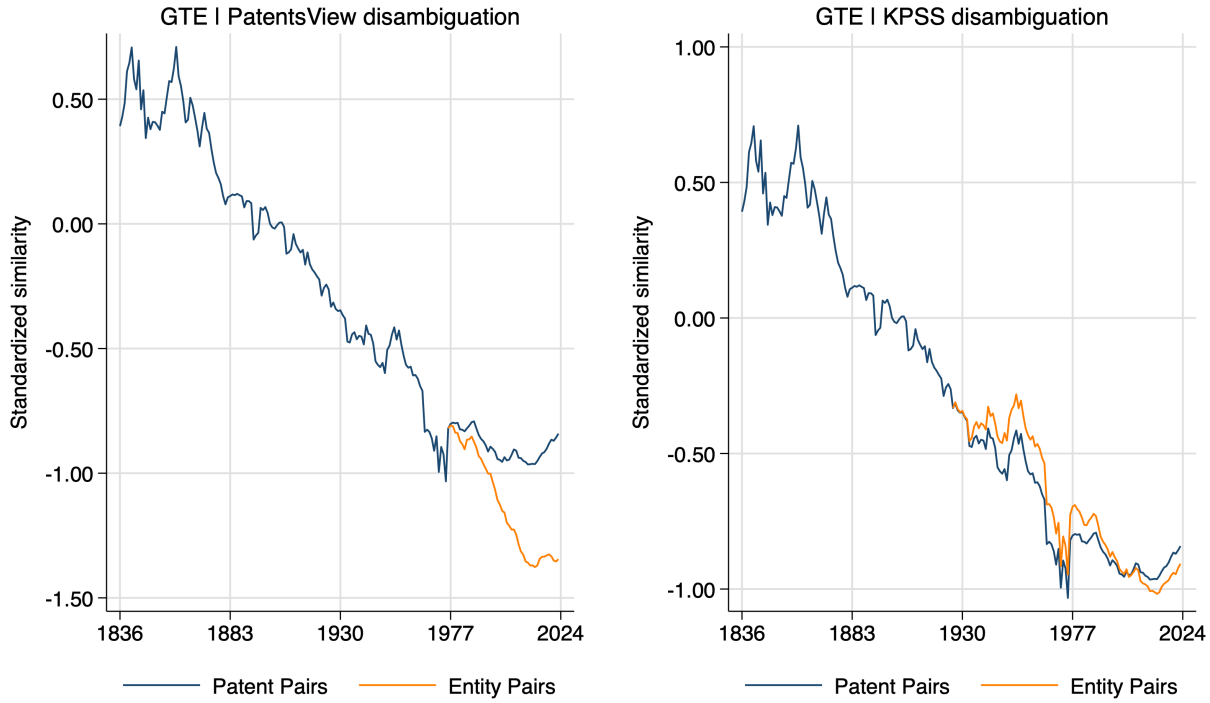


Figure 4: Similarity Correcting for Multi-Patent Entities

These plots show GTE similarity trends when sampling one patent per entity per year. The PatentsView disambiguation (left panel, 1976–2023) shows that correcting for multi-patent entities reduces the post-1999 arrest in declining similarity, revealing continued spreading-out of independent inventors. The KPSS public firms disambiguation (right panel, 1926–2023) shows little difference from baseline results, perhaps because it does not account for private firms or individual inventors that are issued multiple patents.

A secondary approach using KPSS public-firm disambiguation (Kogan et al. 2017) shows smaller effects, likely because it excludes multi-patent strategies by private firms, individual inventors, and non-practicing entities. The PatentsView approach provides more comprehensive entity identification.

4.3 Robustness to Spatial Scale

A potential concern is that global average similarity may not capture the competitive and spillover dynamics emphasized by our theory. Our model focuses on the trade-off between competition and spillovers from nearby inventors in idea space, suggesting that local similarity — the distance to near neighbors — may be more relevant than average similarity across all patents. Conversely, as discussed in Section 2.6, local measures might be confounded by clustering around "low-hanging fruit" or other project-specific factors our model abstracts from.

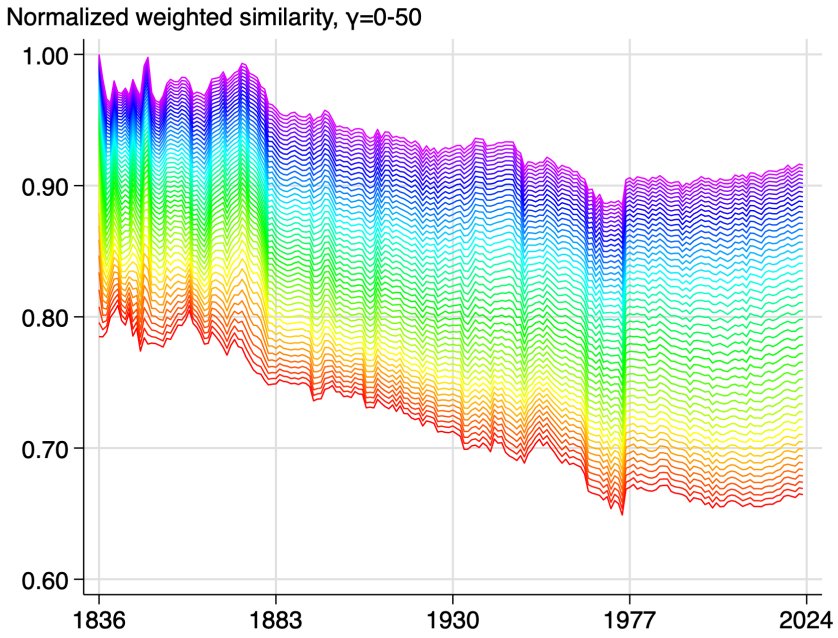


Figure 5: Similarity at Multiple Spatial Scales

This figure shows normalized weighted GTE similarity trends using different spatial scales, with weights γ ranging from 0 (red line, lower envelope, global average) to 50 (magenta line, upper envelope, emphasizing nearest neighbors). The secular decline in similarity is robust across all spatial scales, appearing at both local and global levels of idea space. The slight increase after 1999 is slightly faster for at local scales, indicating clustering. Each γ-similarity has a different natural scale, so dividing all of them by a common cross-sectional standard deviation distorts comparisons across series. Instead, normalizing by the global max preserves both the shape and the relative levels.

We address this concern by computing similarity at multiple spatial scales. Rather than simple average pairwise similarity, we compute weighted average similarity where the weight decays with distance:

$$\text{Weighted Similarity} = \frac{1}{n}\sum_{i=1}^{n}\frac{\sum_{j\neq i}(1-d_{ij})e^{-\gamma d_{ij}}}{\sum_{j\neq i}e^{-\gamma d_{ij}}} \tag{25}$$

where d_{ij} is the cosine distance between patent vectors i and j, and $(1-d_{ij})$ is cosine similarity. The parameter $\gamma \geq 0$ characterizes how rapidly the weight decays with distance in idea space. When $\gamma = 0$, this reduces to unweighted average similarity — our baseline measure. As γ increases, the measure increasingly emphasizes near neighbors: high γ approximates nearest-neighbor similarity, while low γ emphasizes global average similarity.

Figure 5 shows similarity trends for γ values ranging from 0 to 50. The red line (lower envelope) represents $\gamma = 0$ (global average), while the purple line (upper envelope) represents

$\gamma = 50$ (emphasizing very near neighbors). Several patterns emerge. First, as expected, near-neighbor similarity (high γ) is consistently higher than global average similarity (low γ) — patents are more similar to their closest neighbors than to random other patents. Second, and more importantly, the secular decline in similarity is robust across all spatial scales. Whether we emphasize very local proximity or global average distance, inventors are spreading out over time. Third, after 1999, the increase in similarity is slightly faster at local scales, indicating clustering.

This robustness to spatial scale strengthens our interpretation that the declining similarity pattern reflects the fundamental mechanism emphasized by our theory rather than artifacts of how we measure similarity or confounding factors like low-hanging fruit clustering. The spreading-out of inventors appears at multiple levels of idea space, from the more global distribution to the immediate neighborhood (relevant for spillovers and for competitive pressure).

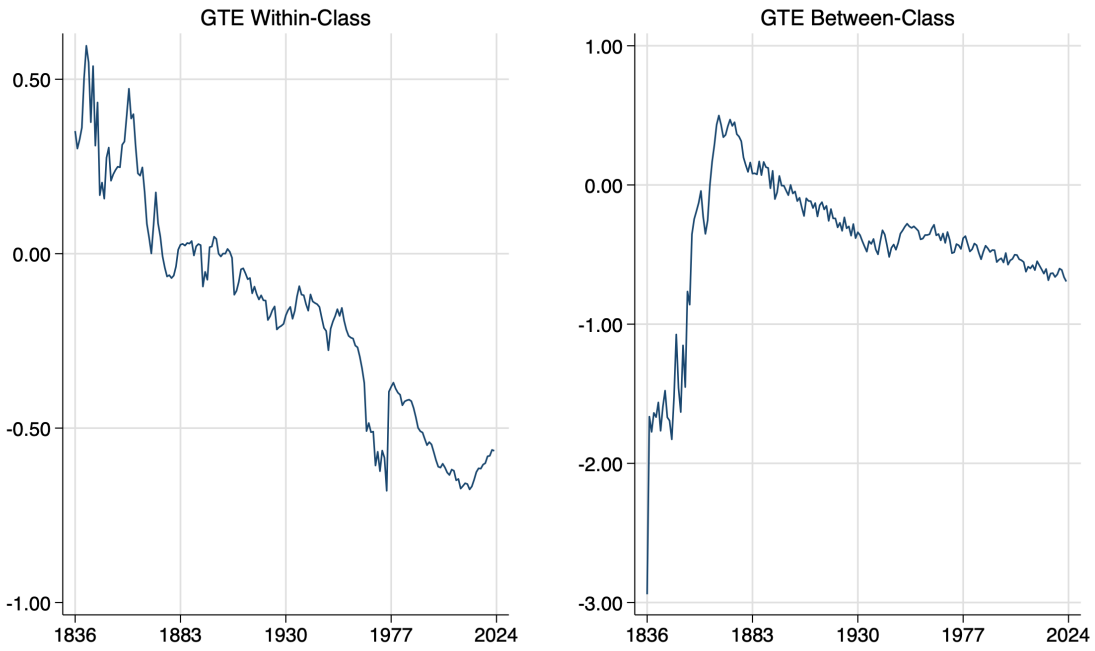
A complementary analysis examines quantiles of the pairwise similarity distribution (Online Appendix S9.5). The secular decline is robust across the entire distribution, not driven by outliers or particular quantiles. The post-1999 increase is slightly faster at higher quantiles, consistent with the local clustering evident at high γ in Figure 5.

4.4 Robustness: Within Versus Between Technology Classes

A related concern is whether spreading-out reflects inventors moving across broad technological field boundaries or also occurs within established fields. This distinction matters both theoretically and empirically. Theoretically, our model emphasizes local competition and spillovers, suggesting that within-field spreading-out may be particularly important. Empirically, if spreading-out only occurs between fields, it might simply reflect shifts in the industrial composition of innovation rather than the fundamental mechanism our theory emphasizes.

Figure 6 decomposes average pairwise similarity into within-class and between-class components using three-character CPC technology classes. (The CPC has eight top-level sections subdivided into over 120 three-character classes.) For each year, we separately compute average similarity for patent pairs in the same class versus pairs in different classes.

Both within-class and between-class similarity decline substantially throughout the sample period, closely tracking the overall trend. The within-class decline is the more important result for identification: it rules out composition effects from the birth of new fields and addresses concerns about "low-hanging fruit" exhaustion. If inventors were simply moving to distant fields after exhausting opportunities in their original fields, we would observe de-



(a) Within three-character CPC classes **(b)** Between three-character CPC classes

Figure 6: Similarity Within and Between Technology Classes

These plots show average pairwise standardized patent similarity using GTE representations, decomposed into within-class (Panel A) and between-class (Panel B) components using three-character CPC technology classifications. The number of CPC classes grows from 42 in 1836 to 123 in 2023; 90% of classes appear by 1891. Both components exhibit secular decline, closely mirroring the overall trend from Figure 2.

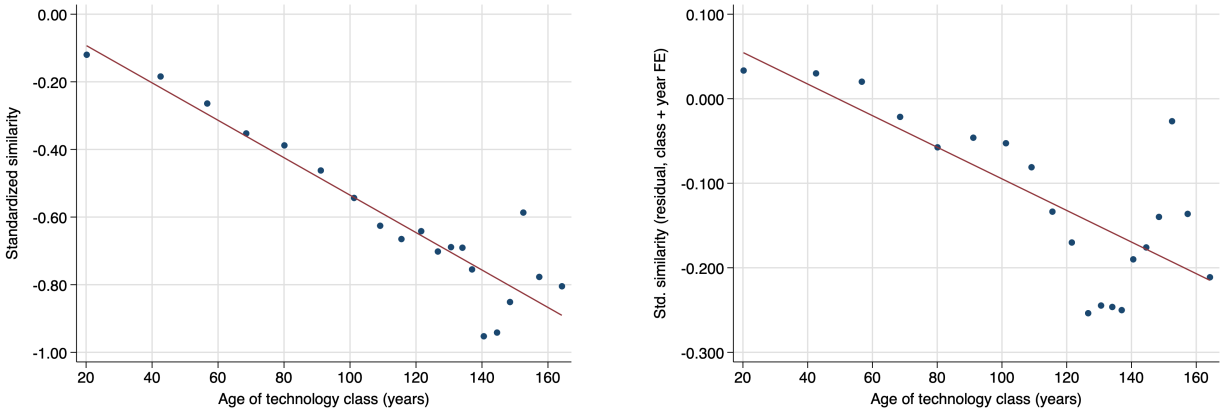
clining between-class similarity but stable or increasing within-class similarity. Instead, both decline in parallel, with within-class decline at least as robust as between-class. This within-field decline raises a further question: does it reflect calendar time — changing language, policy, or measurement — or something about how fields themselves develop?

4.5 Robustness: Similarity Trends by Technology Class Age

Figure 7 separates these explanations by examining similarity trends within technology classes as they age. We define class "age" as years since a class became substantively active (first calendar year with at least 50 issued patents). Classes are born at different historical moments — from A01 ("Agriculture") in 1843 to C40 ("Combinatorial chemistry") in 2001 — allowing us to distinguish field maturation from calendar time.

As classes mature, within-class similarity steadily declines. Mature classes display lower within-class similarity than younger classes, even within the same calendar year. This pattern points to endogenous field evolution — consistent with the model's equilibrium forces

— rather than calendar-time confounds such as changing patent language or examination standards. The interference rate analysis in Section 4.6 provides further evidence against the examination-standards explanation, since interference determinations operated through an institutionally separate process.



(a) Raw similarity **(b)** Residualized on class and year FEs

Figure 7: Similarity Within Class by Class Age

Binscatter plots show average within-class patent similarity for technology classes as they age, using GTE representations. "Class age" is defined as years since reaching at least 50 patents issued in the class. Panel (a) shows raw standardized similarity; Sim ilarity is indexed to 0 at class age of 0. Panel (b) residualizes similarity on CPC three-character class and calendar year fixed effects. The steady decline in within-class similarity with increasing class age survives these controls, indicating that spreading out is tied to endogenous field evolution rather than calendar-time confounds or class-specific levels.

4.6 Corroboration from Declining Interference Rates

We corroborate the declining similarity pattern using an entirely independent data source: patent interference rates from 1836 to 2014. Pre-2001 interferences were not used in our validation process, making this a true out-of-sample test. While interference practices themselves evolved over time — most notably in 1870 and 2014 — the secular decline spans over 150 years across multiple institutional regimes, making it unlikely that any single policy change drives the pattern.

Patent interferences occurred when the US patent office determined that two or more independent parties claimed the same invention. The interference rate — the probability that an issued patent was involved in an interference — thus provides a direct measure of how often inventors independently arrived at identical or nearly identical inventions. Declining interference rates would indicate that inventors are less likely to be working on the same

ideas, consistent with spreading out in idea space.

Data Sources We construct a time series of interference rates from five distinct sources spanning 177 years. One, we digitized the finding guide to Patent Interference Case Files at the National Archives (Butler 1993), covering 1838–1900. Part I (1838–1869) yields 29 case files per year. Part II (1870–1900) identifies 2,682 surviving case files out of 27,271 sequentially-numbered cases; surviving case files underestimate the true number, while numbered cases overestimate it, bounding the true number between 87 and 880 cases per year. Two, we purpose-digitized the US patent office's *Registers of Interferences* from National Archives records for 1864–1900, documenting 19,388 interference cases with an average of 504 annual terminations.[21] Three, summary statistics from Di Simone et al. (1963) report an average of 640 annual interferences for 1950–1962. Four, data from Calvert and Sofocleous (1982, 1986, 1989, 1992, 1995) show an average of 237 annual interferences for 1980–1994. Five, Ganguli et al. (2020) document an average of 76 annual interferences for 1998–2014.[22]

Results Figure 8 reveals a striking and consistent decline in interference rates over more than a century and a half. The average interference rate fell from 2.71% in 1864–1901 (with an upper bound of 4.91% based on the finding guide) to 1.43% in 1950–1962, then to 0.30% in 1980–1994, and finally to 0.05% in 1998–2014 — a decline of more than 98% over the full period.[23]

This dramatic and steady reduction might be explained by changes in patent examination procedures alone. However, if the US patent office's capacity to identify potential interferences improved over time, early rates are potentially understated. Moreover, the consistent decline both within and across four independent data sources and 150 years provides evidence against discrete changes in patent policy. Inventors seem genuinely less likely to be working on identical inventions.

The temporal pattern of interference rate decline also resembles the validated GTE similarity trends. This correspondence is remarkable given that the interference data are com-

[21]See Online Appendix S10 for an example Register page.

[22]This likely slightly undercounts actual interferences, as some were terminated before reaching the Board of Patent Interferences.

[23]The interference rate was at least 1.04% during 1838–1869 based on the finding guide data. The Patent Act of 1870 both reformed interference procedures and required inventors to "distinctly claim" their inventions, marking a shift towards more precise delineation of patent claims (Nard 2010), making interference rates before and after 1870 not directly comparable.

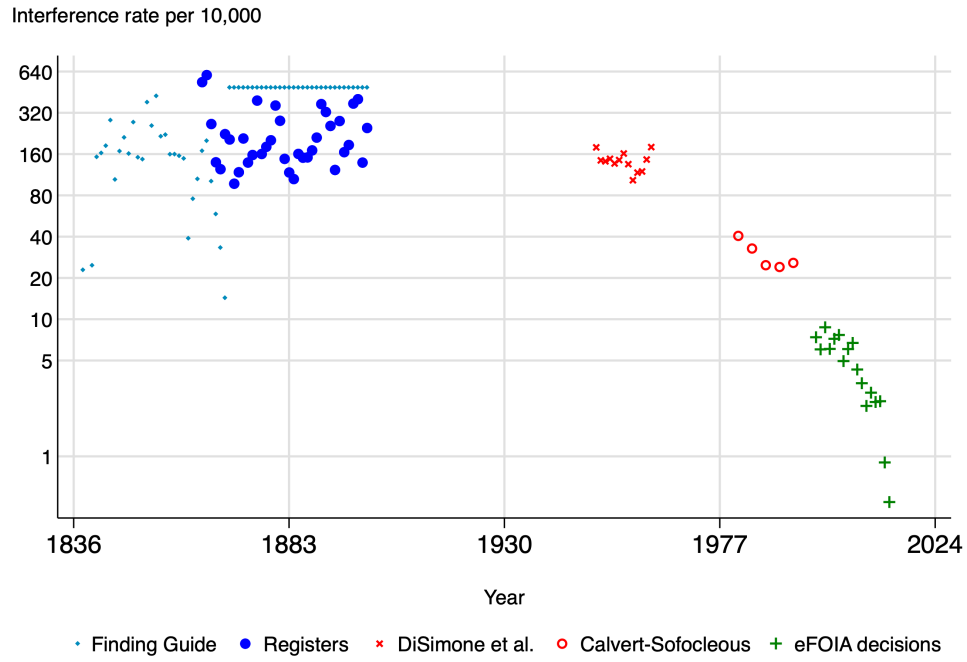


Figure 8: Interference Rate, 1838–2014

This plot shows the estimated interference rate per 10,000 issued utility patents across 177 years. Different markers indicate data from different sources. The interference rate is shown on a log scale to facilitate visualization of the decline. The secular decline in interference rates provides independent confirmation of declining invention similarity. The pattern of greater variability in the 19th century followed by steady decline from the mid-20th century closely resembles the similarity trends measured using validated text representations.

pletely independent of our text-based similarity measures for the pre-2001 period. Moreover, the interference rate series provides a particularly clean control for multi-patent entities: because the US patent office explicitly verifies that interference participants are independent parties, this measure avoids the potential disambiguation errors that could affect our earlier entity-corrected analysis in Section 4.2. The consistency of declining patterns across both approaches — entity-corrected patent similarity and independently-verified interference rates — strengthens confidence that spreading-out reflects genuine changes in inventor positioning rather than artifacts of patent portfolio strategies.

4.7 Expanding Convex Hull of Idea Space

Is idea space H expanding over time? We provide direct evidence by measuring the volume of the convex hull of patent embeddings in each year.

GTE embeddings lie in a 1024-dimensional space, making direct volume computation

infeasible. We reduce dimensionality by projecting embeddings onto their first 7 principal components and compute the volume of the convex hull of projected patents in each year. Figure 9 shows that the convex hull volume grows at approximately 0.5% per year. This growth rate provides an independent estimate of g_H — the rate of idea space expansion — that can be compared against the structural estimate from the R&D regression in Section 6.2.

Dimensionality reduction likely causes this estimate to understate the true rate of expansion of idea space, since projecting onto a low-dimensional subspace compresses variation in the remaining dimensions. (We estimate a slower growth rate of 0.4%/year using 6 principal components.) The convex hull estimate should therefore be interpreted as a lower bound on the growth rate of idea space.

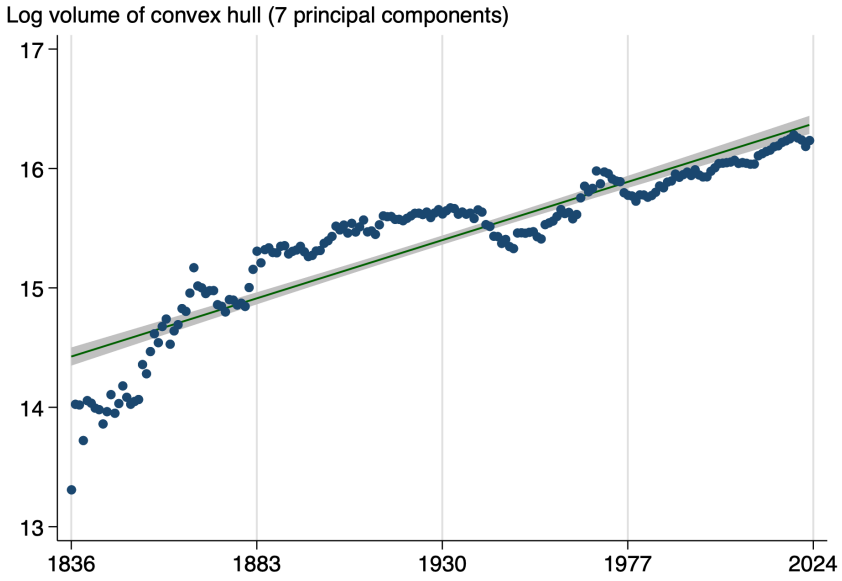


Figure 9: Convex Hull Volume of Patent Embeddings

This figure shows the log volume of the convex hull of patent GTE embeddings projected onto 7 principal components, by year. The convex hull expands at approximately 0.5%/year, providing direct evidence that idea space is growing over time.

4.8 Interpretation and Implications

The convergence of evidence from multiple sources provides strong support for our theoretical prediction that inventors are spreading out over an expanding knowledge frontier. This pattern appears consistently across nearly two centuries of patent text (1836–2023). It strengthens in recent decades when correcting for multi-patent entities. It appears at multiple spatial scales, from the global to the very local. It appears between and within technology

class boundaries and within technology classes as they age, at different points in calendar time. It is also corroborated by independent evidence from falling interference rates (1838–2014).

The robustness of declining similarity across these diverse approaches strengthens confidence in our finding. While uncertainty remains about precise magnitudes, the qualitative conclusion is clear: contemporary invention similarity has declined substantially over the long run of American innovation history.

5 Spreading Out and Quality

The model predicts that spacing and quality are jointly determined: $q = d/\gamma$ (Corollary 2). Patents farther from competitors in idea space require greater R&D investment to serve larger technological territories. We test this prediction using cross-sectional variation across approximately 220,000 patents from 1976 to 2019.

We measure spacing using patent-level similarity to contemporaneous inventions, computed from both GTE and PaECTER embeddings at two scales: the mean similarity (our baseline global measure) and the 98th percentile (capturing extremely local distance, approximately nearest-neighbor proximity). We measure quality using three patent-level indicators. Two capture R&D input intensity: the number of co-inventors and whether the patent is assigned to a firm. Firm-assigned patents reflect organized R&D with greater resource inputs compared with independent inventors. The third is an output measure: forward citations received within five years.

We regress each quality measure on each similarity measure, including year and CPC technology class fixed effects so that identification comes from within-field and within-year variation. Standard errors are clustered by year and CPC class. Table 2 reports the results; Figure 10 displays the relationships visually.

The R&D input measures are strongly consistent with the comovement prediction. All eight R&D input coefficients are negative, and seven are statistically significant. Including the forward citation measure, ten of twelve coefficients are negative overall. Patents more similar to their contemporaries — closer to neighbors in idea space — have fewer co-inventors and are less likely to be assigned to a firm. The pattern holds for both GTE and PaECTER, and at both the mean and 98th-percentile scales, though PaECTER's 98th-percentile estimate for co-inventors is imprecise.

Forward citations tell a less clear story. GTE citation coefficients are positive but statistically indistinguishable from zero. PaECTER citation coefficients are negative and significant. This divergence is unsurprising: forward citations are an output measure subject to iden-

Table 2: Cross-Sectional Evidence: Patent Similarity and Quality

	Co-Inventors	Firm Assignment	Citations (5-yr)
Panel A: Mean Similarity			
GTE	−2.545***	−0.659***	6.347
	(0.679)	(0.180)	(4.480)
PaECTER	−4.928***	−1.723***	−36.681***
	(1.002)	(0.424)	(8.383)
Panel B: 98th Percentile Similarity			
GTE	−1.477**	−0.253*	4.788
	(0.670)	(0.131)	(3.220)
PaECTER	−1.067	−1.174***	−19.404***
	(1.592)	(0.431)	(6.443)
Observations		219,772	
Fixed Effects		Year + CPC Class	

Notes: Each cell reports the coefficient from a separate bivariate regression of the column variable on the row similarity measure. Sample: 5,000 randomly selected patents per year, 1976–2019. All specifications include year and CPC technology class fixed effects. Standard errors in parentheses, clustered by year and CPC class. *** — $p < 0.01$, ** — $p < 0.05$, * — $p < 0.1$.

tification concerns that do not affect the input measures. Citation counts reflect not only invention quality but also the density of the local citation network, examiner practices, and strategic citing behavior. The input measures — co-inventors and firm assignment — more directly capture the R&D investment the model predicts, and these are unambiguous.

Time-series evidence reinforces these cross-sectional patterns. In Online Appendix S11, we aggregate to the CPC-class-year level and regress changes in quality on changes in similarity, with CPC class and year fixed effects. Point estimates are predominantly negative (10 of 12 specifications), consistent with the prediction that spreading out within a technology class accompanies rising quality. The modern-patent estimates are generally imprecise, with only two of twelve coefficients statistically detectable — reflecting limited power in class-level annual changes. Extending the co-inventor analysis to 1836–2023 using the CUSP dataset (Berkes 2016) yields 15,813 class-year observations and statistically significant negative coefficients for both GTE and PaECTER similarity.

6 Spreading Out and Research Productivity

Research productivity is declining. Figure 11 displays the central fact documented by Bloom et al. (2020): aggregate R&D spending grew at a roughly constant 4.0%/year over 1948–

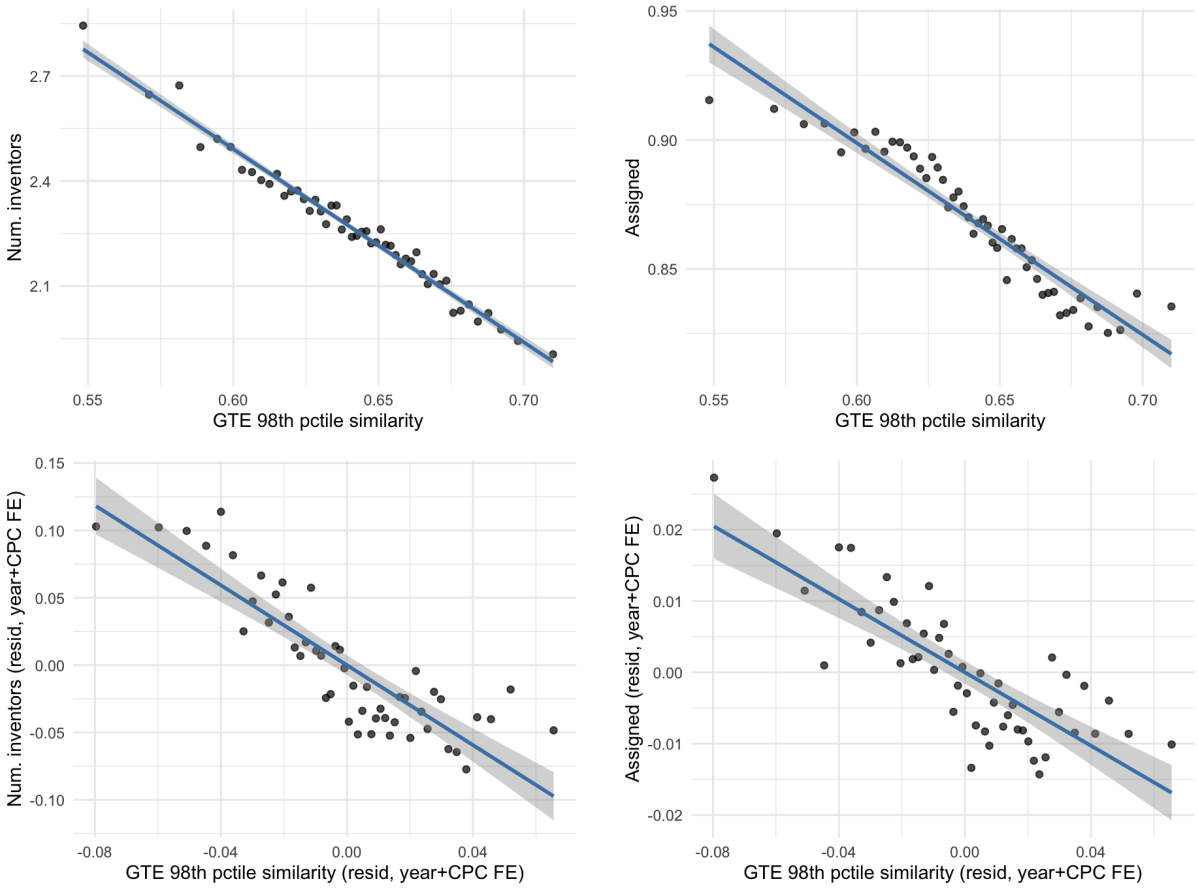


Figure 10: Patent Similarity and Quality: Binned Scatterplots

Notes: Binned scatterplots (50 quantile bins) with fitted lines. Top row: raw data. Bottom row: residualized on year and CPC class fixed effects. Left column: number of co-inventors. Right column: firm assignment (binary). Similarity measured using GTE 98th percentile. Negative slopes indicate that patents farther from neighbors in idea space have higher R&D inputs, consistent with the spacing-quality comovement prediction.

2015, representing a 14-fold increase in research effort, yet TFP growth fell from 2.1%/year to 0.7%/year — a decline by a factor of 3. Proposition 3 predicts exactly this: declining research productivity as H grows.

We evaluate these predictions using annual data from Bloom et al. (2020) for TFP growth and aggregate R&D real input growth and changes in our GTE-PaECTER ensemble similarity measure. Figure 12 illustrates the time series correlation. Declining similarity (spreading out) associates with lower TFP growth and faster R&D growth, consistent with Proposition 3.

The regressions provide reduced-form evidence: predicted signs (spreading out reduces TFP growth, increases R&D spending), predicted ratios ($g_d/g_{R\&D} \approx 1/3$), and cross-validation against Bloom et al. (2013) quasi-experimental spillover elasticities. Combined

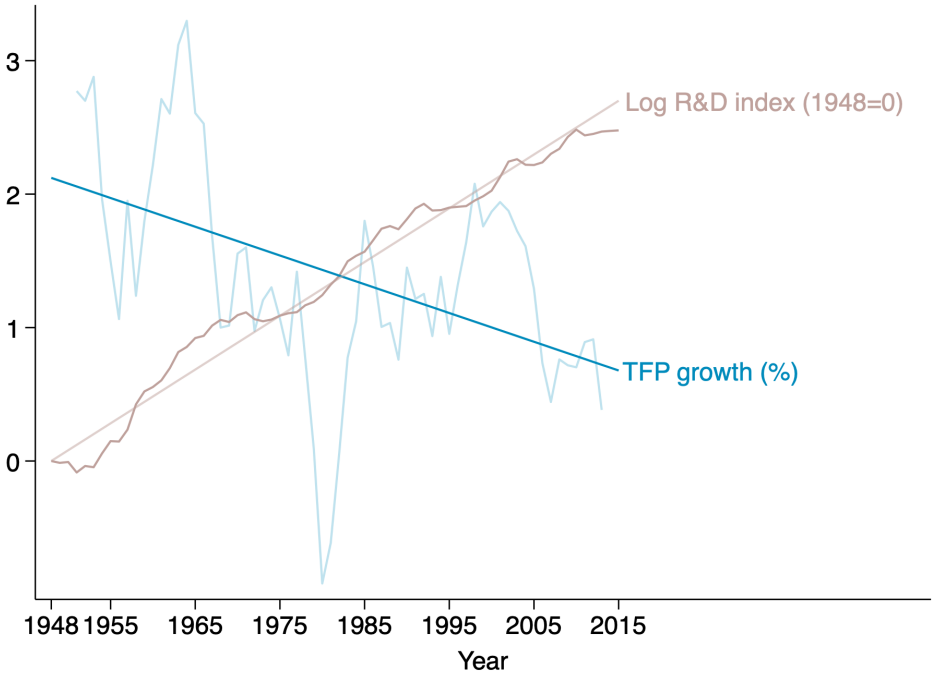


Figure 11: TFP Growth and Real R&D Growth

Notes: Annual TFP growth rate and log real R&D spending (indexed, 1948=0) from Bloom et al. (2020), 1948–2015. R&D spending grew at a nearly constant 4.0%/year, while TFP growth declined from 2.1%/year in 1948 to 0.7%/year in 2015 (average=1.5%), illustrating the research productivity decline.

with the model's unification of previously disconnected empirical patterns (Section 2.5), this ensemble of validated predictions builds confidence in the spatial mechanism. Section 6.2 also calibrates the model's key structural parameters — g_H from the equilibrium identity $R\&D = \tau H d$ and θ from National Center for Science and Engineering Statistics (2025) survey data — and Section 6.3 uses them to decompose the aggregate research productivity decline into spatial and non-spatial forces.

6.1 TFP Growth

We begin by examining how spreading out affects TFP growth, the output of the innovation process. We estimate the TFP–similarity relationship in regression form. Consider the regression:

$$\Delta \log(\text{TFP})_t = b_0 + b_1 \cdot \Delta(-1 \times \text{Similarity})_t + b_3 \cdot t + \epsilon_t. \tag{26}$$

This equation relates TFP growth to observed (standardized) changes in technological distance $\Delta(-1 \times \text{Similarity})$ between inventors. If spreading out leads to declines in TFP growth

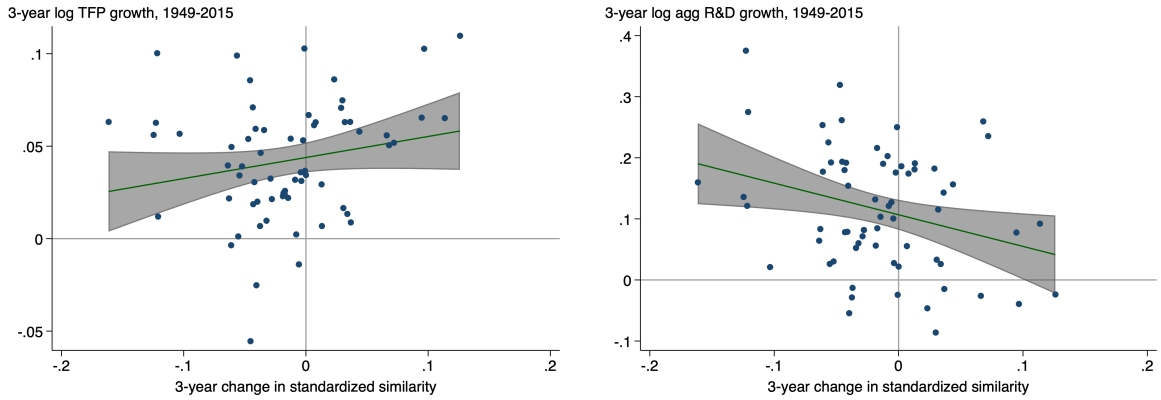


Figure 12: TFP Growth and Aggregate R&D Growth versus Changes in Idea Similarity
This figure plots 3-year log TFP growth and log aggregate R&D input growth from Bloom et al. (2020), 1948–2015, against 3-year changes in our GTE-PaECTER ensemble patent similarity measure. Declining similarity (spreading out) associates with lower TFP growth and faster R&D growth.

(e.g., from spillover attenuation and adaptation costs) then $b_1 < 0$. The time trend b_3 (we reserve b_2 for the interaction term introduced below) absorbs smooth trend components of TFP growth — including the transition dynamics predicted by the model as spillovers attenuate — so that b_1 isolates the marginal effect of year-to-year changes in similarity.

This yields estimates $b_1 = -0.169$ (s.e. 0.057, $p < 0.01$) and $b_3 = -2.67 \times 10^{-4}$ (s.e. 9.90×10^{-5}, $p < 0.05$) (Table 3, Column 1). Spreading out is associated with slower TFP growth, consistent with Proposition 3.

Structural interpretation We derive a structural equation that yields a sharper prediction. Average log TFP is $A \equiv Q - \frac{\tau d}{4}$, where realized quality $Q = q(1 + \beta - \beta d/\lambda)$ includes spillover benefits. Differentiating with respect to time yields a structural equation that holds along any equilibrium path, not only on the balanced growth path:

$$g_{TFP} = \frac{dq}{dt}\left(1 + \beta - \frac{\beta d}{\lambda}\right) - \left(\frac{\beta q}{\lambda} + \frac{\tau}{4}\right)\frac{dd}{dt} \tag{27}$$

The first term captures quality growth scaled by the spillover factor. The second term represents combined spatial drag: spillover attenuation ($\frac{\beta q}{\lambda}\frac{dd}{dt}$) and adaptation costs ($\frac{\tau}{4}\frac{dd}{dt}$). Both forces reduce TFP growth as inventors spread out.

Substituting the equilibrium relationships $q = d/\gamma$ and $dq/dt = (1/\gamma)(dd/dt)$ eliminates unobserved quality in favor of spacing changes Δd, which we proxy with changes in measured

Table 3: Time Series Evidence: Technological Proximity and TFP Growth

	Annual		Multi-Year	
			3-Year	5-Year
	(1)	(2)	(3)	(4)
$b_1 : -1 \times \Delta_t \text{Sim}$	-0.169***	-0.171**	-0.278***	-0.269***
	(0.057)	(0.083)	(0.095)	(0.098)
$b_2 : (-1 \times \Delta_t \text{Sim}) \times (-1 \times \text{Sim}_{t-1})$		-0.015	-0.408	-0.571*
		(0.342)	(0.320)	(0.312)
b_3: Year	-2.67e-4**	-2.62e-4**	-8.19e-4***	-1.19e-3***
	(9.90e-5)	(9.99e-5)	(1.91e-4)	(2.70e-4)
R-squared	0.174	0.174	0.280	0.293
Observations	67	67	65	63
Implied annualized TFP drag from spreading out (pp/year):				
In 1991 (Sim = 0):	-0.084	-0.085	-0.139	-0.156
In 1948 (Sim = 0.347):		-0.082	-0.068	-0.041
In 2015 (Sim = 0.060):		-0.084	-0.127	-0.136

Notes: Each column reports estimates from a separate regression. Dependent variable is log TFP growth over the specified horizon. Columns 2–3 use three-year and five-year differences ($\Delta_3 \log$ and $\Delta_5 \log$) to smooth through high-frequency measurement error. Similarity is standardized (standard deviation = 1) and indexed to 0 in 1991, so the main effect can be compared directly with the Bloom et al. (2013) cross-sectional elasticity. All specifications include a constant (not reported). Standard errors in parentheses. TFP from Bloom et al. (2020), 1948–2015. Average change in standardized similarity is -0.005, -0.015, and -0.029 over annual, 3-year, and 5-year horizons, respectively. *** — $p < 0.01$, ** — $p < 0.05$, * — $p < 0.1$.

similarity. (We validate this proxy later.)

$$g_{TFP} = \underbrace{\left[\frac{1}{\gamma}\left(1 + \beta - \frac{2\beta d}{\lambda}\right)\right]}_{\text{Spillover-adjusted quality effect}} \cdot \Delta d - \underbrace{\frac{\tau}{4}}_{\text{Adaptation drag}} \cdot \Delta d \tag{28}$$

The reduced-form coefficient b_1 reflects a combination of negative and positive contributions to TFP from spreading out. Spillover attenuation and adaptation costs contribute negatively, but gross quality improvements contribute positively $(1 + \beta)$, creating an offsetting force. The structural equation also predicts an interaction term $d \cdot \Delta d$ with coefficient $-\frac{2\beta}{\gamma\lambda}$: as inventors spread farther apart, marginal R&D spending has reduced impact on TFP because of spillover attenuation.

We include this interaction term (with coefficient b_2) in Column 2, where we have multiplied both similarity and changes in similarity by -1 to facilitate interpretation as technological distance, aligning with equation (28). (We index similarity to 0 in 1991 so that

the main effect can be compared directly with the Bloom et al. (2013) cross-sectional elasticity.)[24] The coefficient estimate on the interaction is negative, as predicted by theory, but it is statistically insignificant. Columns 3–4 smooth through high-frequency measurement error using three-year and five-year differences. The mechanism emerges clearly. Using five-year differences (Column 4), all coefficients are statistically significant: increasing distance reduces TFP growth (-0.269, $p < 0.01$), and crucially, the interaction term is large and significant at the 10% level (-0.571, $p < 0.10$).

The interaction coefficient indicates that as average distance increases, spreading out reduces TFP more. To illustrate this, we compute the implied annualized TFP drag from spreading out, holding fixed the average annual rate of similarity (-0.005 standardized units/year) but varying the average similarity level. At 1948 similarity levels, strong spillovers mean that quality improvements partially offset attenuation and adaptation costs, yielding a modest expected net drag of -0.041pp/year. At 2015 similarity levels, spreading out generates a net negative contribution to TFP growth of -0.136pp/year — spillover attenuation and adaptation costs dominate.

The regression coefficient on Δd (Column 4: $\hat{b}_1 = -0.269$) captures the combined effect of spillover-adjusted quality growth $\frac{1+\beta}{\gamma}$ net of adaptation drag $-\frac{\tau}{4}$. The interaction coefficient (Column 4: $\hat{b}_2 = -0.571$) estimates the spillover attenuation rate $-\frac{2\beta}{\gamma\lambda}$. This quantifies $\frac{\beta}{\gamma\lambda} = 0.286$, the rate at which spillover benefits decay with distance (scaled by the cost parameter γ).

Validation with Cross-Sectional Elasticity Estimates Bloom et al. (2013) and Lucking et al. (2019) estimate how firm-level TFP responds to its spillover pool, defined as the sum of neighboring firms' R&D weighted by (inverse) idea-space distance. The spillover pool is instrumented using state R&D tax credit shocks — completely independent of our time-series approach. Bloom et al. (2013) report a direct TFP elasticity of 0.206. This elasticity captures the full effect on productivity: distant neighbors provide weaker spillover benefits, and any costs of absorbing or adapting distant knowledge are in principle reflected in the reduced measured TFP response.[25]

[24]Because the interaction model is $g_{TFP} = b_1\Delta d + b_2(d \cdot \Delta d)$, the marginal effect of Δd is $b_1 + b_2 d$. Centering d at 1991 means $\hat{b}_1$ is the marginal effect evaluated at the 1991 baseline, not the unconditional main effect.

[25]Our similarity data show that technological proximity was relatively stable during the Bloom et al. (2013) sample period (1981–2001), declining sharply pre-1980 and stabilizing thereafter with retracing after 2000. This stability favors the Bloom et al. (2013) elasticity over the Lucking et al. (2019) estimate because it strengthens the causal interpretation of the

Our patent similarity data show technological proximity declining by 0.144 standard deviations over 1981–2001 (Bloom et al. (2013)'s sample period). Combining this with Bloom et al. (2013)'s elasticity estimate and the cross-sectional standard deviation of log(SPILLTECH) = 1.04:

$$\Delta \log(\text{SPILLTECH}) = -0.144 \times 1.04 = -0.150 \tag{29}$$

$$\Delta \log(\text{TFP}) = 0.206 \times (-0.150) = -0.031 \tag{30}$$

This predicts a 3.1pp cumulative decline in TFP due to spreading over 1981–2001, or approximately -0.16pp/year. If instead we use the -0.287 decline in standardized similarity over 1948–2015 and the Lucking et al. (2019) elasticity estimates of 0.287 with the 1981–2015 standard deviation of log(SPILLTECH) = 1.17, this yields an annualized TFP drag of $(-0.287 \times 1.17 \times 0.287)/(2015-1948) = -0.14$pp/year. The fact that two very different identification strategies — our time-series regressions (yielding -0.084pp/year to -0.156pp/year) and Bloom et al. (2013)'s or Lucking et al. (2019)'s cross-sectional IV (yielding -0.14pp/year to -0.16pp/year) — yield nearly identical magnitudes is strong evidence that confounds are not driving the time-series relationship. The near-exact alignment between our time-series estimate and independent quasi-experimental evidence is remarkable.

The range of time-series estimates are slightly smaller compared with the cross-sectional estimates. This is consistent with equation (28): the cross-sectional elasticity captures the full effect of distance on downstream TFP (spillover attenuation plus adaptation frictions), while the time-series coefficient also reflects the offsetting quality improvements that spreading out induces through larger territories. The cross-sectional elasticity does not include these equilibrium adjustments — increases in spacing, entry, or quality — so it should exceed the net time-series effect.

6.2 R&D Spending Growth

We turn next to the supply side, examining how spreading out affects aggregate R&D spending. Aggregate R&D spending $\text{R\&D}(t) = n(t) \cdot [c(q(t)) + f(H(t))]$ grows at rate:

$$g_{R\&D} = \underbrace{g_n}_{\text{Entry}} + \underbrace{\theta(1+\eta)g_q}_{\text{Quality (incl. fishing out)}} + \underbrace{(1-\theta)g_f}_{\text{Rising fixed costs}} \tag{31}$$

cross-sectional estimates: the identifying variation comes from differences in initial proximity across firms rather than from active repositioning, reducing concerns about endogenous sorting. Lucking et al. (2019) extend the sample to 2015 and find that the elasticity is 0.287.

where $\theta \equiv \frac{c(q)}{c(q)+f}$ is the variable cost share and η governs R&D cost curvature (Section 2.1). Substituting the equilibrium relationships $g_q = g_d$ (from $q = d/\gamma$), $g_n = g_H - g_d$ (from $n = H/d$), and $g_f = \alpha g_H$ (from $f = \phi H^\alpha$):

$$g_{R\&D} = \underbrace{[1 + \alpha(1-\theta)]}_{\text{idea space expansion}} g_H + \underbrace{[\theta(1+\eta) - 1]}_{\text{spreading out}} g_d \tag{32}$$

R&D growth depends on two forces: expansion of idea space g_H, reflecting both entry growth and rising fixed costs; and spreading out g_d, which increases variable costs via quality scaling but slows entry. When $\theta(1+\eta) > 1$ — variable costs sufficiently dominate — spreading out raises R&D spending.

Idea space growth g_H is not directly observable, but the model predicts it declines during the transition as spillovers attenuate: $g_H = g_H^* + \frac{\delta\beta}{\gamma}(1 - d/\lambda)$ (Section 2.4). We approximate this transition with a linear trend $g_H(t) \approx g_H^{1991} + \delta_H \cdot t$ (centered at 1991), absorbed by the constant and time trend. Spacing growth g_d is proxied by changes in measured similarity. Substituting into equation (32):

$$g_{R\&D,t} = a_0 + a_1 \cdot t + a_2 \cdot \Delta(-1 \times \text{Similarity})_t + \epsilon_t \tag{33}$$

The model's testable prediction is $a_2 > 0$: years with greater spreading out should see faster R&D growth.

Table 4 reports the regression estimates. Column (1) uses annual differences but the signal-to-noise ratio is low ($R^2 = 0.032$). Columns (2)–(3) use multi-year differences (3-year and 5-year) to smooth through measurement error and timing issues in R&D accounting. Using 5-year differences (Column 3), spreading out significantly increases R&D spending ($a_2 = 0.438$, $p < 0.10$), confirming the model's sign prediction.

Structural interpretation Strictly, the TFP regression involves Δd (because log TFP is linear in d), while the R&D equation involves $g_d = \Delta d/d$ (because log R&D involves $\log d$). The same empirical proxy $-\Delta$Sim cannot serve as both simultaneously. In practice, the approximation is close: for small changes, $\Delta d \approx d \cdot g_d$, so the two are proportional up to a slowly-moving scale factor. If $-\Delta$Sim were exactly proportional to g_d, the regression coefficients in equation (32) would identify structural parameters: $\theta = (a_2 + 1)/(1 + \eta)$ and $g_H = a_0/[1 + \alpha(1 - \theta)]$. Under the baseline ($\alpha = 1$, $\eta = 1$), the 5-year estimates imply $\theta = 0.719$ and $g_H = 2.70\%$/year (Table 4, bottom panel), which we compare with externally calibrated values below. The ratio of changes in similarity to changes in log R&D spending is 0.2–0.4 depending on the time horizon; the model predicts $g_d/g_{R\&D} = 1/3$ under the

Table 4: Time Series Evidence: Technological Proximity and Aggregate R&D Growth

		Multi-Year	
	Annual	3-Year	5-Year
	(1)	(2)	(3)
$a_2 : -1 \times \Delta_t$Sim	0.165	0.448**	0.438*
	(0.177)	(0.219)	(0.244)
a_1 : Year (1991=0)	-2.88e-4	-7.79e-4	-1.65e-3*
	(3.05e-4)	(6.58e-4)	(9.65e-4)
a_0: Constant	0.034***	0.102***	0.173***
	(0.006)	(0.013)	(0.018)
R-squared	0.032	0.107	0.147
Observations	67	65	63
Regression-implied parameters (for comparison):			
θ (Variable cost share)	0.583	0.724	0.719
g_H^{1991} (baseline, %/year)	2.40	2.66	2.70
δ_H (acceleration, pp/year)	-0.020	-0.020	-0.026

Notes: Each column reports estimates from a separate regression. Dependent variable is log aggregate R&D growth over the specified horizon. Columns 2–3 use three-year and five-year differences ($\Delta_3 \log$ and $\Delta_5 \log$) to smooth through high-frequency measurement error. Similarity is standardized (standard deviation = 1) and indexed to 0 in 1991. The model predicts $a_2 > 0$ (spreading out raises R&D). Regression-implied parameters assume $a_2 = \theta(1+\eta) - 1$ and $a_0 = g_H^{1991}[1 + \alpha(1-\theta)]$; these are shown for comparison with the primary calibration ($\theta = 0.69$ from NSF survey data; $g_H = 2.67\%$/year from the model identity, equation (37)). Standard errors in parentheses. R&D from Bloom et al. (2020), 1948–2015. Average change in standardized similarity is -0.005, -0.015, and -0.029 over annual, 3-year, and 5-year horizons, respectively. *** — $p < 0.01$, ** — $p < 0.05$, * — $p < 0.1$.

baseline, and the empirical ratio brackets this prediction.[26] The regression is thus consistent with the model's quantitative predictions, not just its sign prediction.

6.3 Contributions to Research Productivity Decline

We decompose research productivity using externally calibrated parameters. This is a calibrated decomposition conditioned on the structure of the model — it quantifies how much of the observed productivity decline is attributable to spatial forces under the model's maintained assumptions.

Following Bloom et al. (2020), we define *research productivity* as the ratio of TFP growth to the level of research effort: $\Pi_t \equiv g_{TFP,t}/R\&D_t$. Taking logs and time derivatives, the

[26] From equation (32) with $g_d = \frac{\alpha}{2} g_H$, the model predicts $g_d/g_{R\&D} = \frac{\alpha/2}{1+\alpha/2+\alpha\theta(\eta-1)/2}$. Under $\alpha = 1$, $\eta = 1$, θ cancels and this simplifies to 1/3.

growth rate of research productivity is:

$$g_{\Pi} = g_{g_{TFP}} - g_{R\&D} \tag{34}$$

where $g_{g_{TFP}} = \frac{d\ln(g_{TFP})}{dt}$ measures how fast TFP growth itself is changing, and $g_{R\&D}$ is the growth rate of research effort. Over 1948–2015, TFP growth declined by a factor of 3 (Figure 11), implying $g_{g_{TFP}} = -\ln(3)/67 \approx -1.6\%$ per year. Combined with research effort growth of $g_R = 4.0\%$ per year over 1948–2015, research productivity declined at:

$$g_{\Pi} = -1.6\% - 4.0\% = -5.6\% \text{ per year} \tag{35}$$

This estimate is close to Bloom et al. (2020)'s estimate of -5.1% per year.[27]

Calibration In symmetric equilibrium, the zero-profit condition and $n = H/d$ yield a level identity:

$$R\&D = \tau H d \tag{36}$$

Taking growth rates: $g_{R\&D} = g_H + g_d$. This is equivalent to the cost-channel decomposition in equation (32): the zero-profit condition ensures the two decompositions coincide.[28] Under the baseline ($\alpha = 1$), $g_d = \frac{1}{2} g_H$ (Section 2.4), so $g_{R\&D} = \frac{3}{2} g_H$. With observed $g_{R\&D} = 4.0\%$/year:

$$g_H = \tfrac{2}{3} \times 4.0\% = 2.67\%/\text{year} \tag{37}$$

This estimate is parameter-free under the baseline, requiring no regression coefficient or similarity–distance mapping. The convex hull of patent embeddings projected to 7 principal components expands at approximately 0.5%/year. This is a conservative lower bound: a 7D hull captures expansion along $7/k_{eff}$ effective dimensions, so consistency with $g_H = 2.67\%$ requires $k_{eff} \approx 38$ independently expanding dimensions — plausible in 1024-dimensional GTE embeddings, where PCA concentrates variance in top components and the remaining dimensions contribute sub-linearly. The convex hull corroborates the direction and order of magnitude of idea-space expansion, though precise dimensional accounting requires knowledge of the full variance spectrum.

We calibrate θ from external data. National Center for Science and Engineering Statis-

[27]The small discrepancy reflects differences in aggregation — Bloom et al. (2020) aggregate to decadal differences first — and in time period — they combine the 1930s and 1940s using the coarser Gordon (2016) data with BLS/BEA data starting in 1948.

[28]Substituting $g_d = \frac{1}{2} g_H$ into equation (32) gives $(2-\theta)g_H + (2\theta - 1)\frac{1}{2} g_H = \frac{3}{2} g_H$ regardless of θ.

tics (2025) data show labor and direct research costs comprise 69% of business R&D, giving $\theta = 0.69$.[29] This implies $\gamma\tau = 1/(2\theta) = 0.72 > 0.5$, empirically confirming the spreading-out condition (Proposition 2).[30] That the regression-implied and externally-calibrated parameters are close (θ: 0.719 vs. 0.69; g_H: 2.70 vs. 2.67) provides a consistency check, but the calibrated values do not depend on the similarity–distance mapping.

Variety growth and α The model predicts that the number of inventions grows at $g_n = (1 - \alpha/2)g_H$. Over 1976–2023, the number of U.S. utility patents granted grew at 3.9%/year, while the number of unique inventor entities (using PatentsView disambiguation, Section S9.2) grew at 2.0%/year. Patent counts probably overstate invention growth: the model requires $g_n < g_H = 2.67\%$, so 3.9%/year is inconsistent with any $\alpha > 0$. The wedge could reflect continuation patents, defensive portfolios, and strategic fragmentation of single ideas across multiple filings — implying that ideas per patent have been falling. Unique entity growth provides a closer proxy. Inverting $g_n = (1 - \alpha/2)g_H$ with $g_n = 2.0\%$ gives $\alpha = 0.50$, at the low end of the sensitivity range (Table 6). The mapping from entity growth to g_n depends on whether inventions per entity–year are rising or falling: if rising, $g_n > 2.0\%$ and $\alpha < 0.50$; if falling, $g_n < 2.0\%$ and $\alpha > 0.50$. Either way, entity growth likely rules out the upper half of the sensitivity range ($\alpha > 1$), suggesting a spatial share of at least 42% — the baseline is conservative.

The baseline sets $\alpha = 1$ and $\eta = 1$ alongside the calibrated $\theta = 0.69$ and $g_H = 2.67\%$/year. From the balanced growth path relationships (Section 2.4), $g_d = \frac{\alpha}{2} g_H = 1.33\%$/year, $g_n = (1 - \frac{\alpha}{2})g_H = 1.33\%$/year, and $g_q = g_d = 1.33\%$/year.

We explore sensitivity to the burden of knowledge elasticity α (how fast fixed entry costs scale with frontier size) and the R&D cost curvature η, neither of which is directly observed. For each (α, η) pair, we hold $\theta = 0.69$ fixed and solve for g_H by constraining the four R&D components in equation (31) to sum to 4.0%/year:

$$g_H = \frac{4.0\%}{1 + \frac{\alpha}{2}\Big[1 + \theta(\eta - 1)\Big]} \tag{38}$$

When $\eta = 1$, this simplifies to $g_H = 4.0\%/(1 + \alpha/2)$, independent of θ.[31]

[29] This mapping is not direct because some labor costs (e.g., team formation) represent fixed costs and some non-labor costs (e.g., materials) represent variable costs.

[30] From the equilibrium conditions, $\theta = \frac{1}{2\gamma\tau}$, so $\gamma\tau = \frac{1}{2\theta}$.

[31] When $\eta \neq 1$, a strict balanced growth path with constant cost shares does not exist; the decomposition is a local approximation at current parameter values.

Table 5: Decomposition of Research Productivity Decline (Baseline)

Component	**Type**	**Baseline** α=1, η=1, θ=0.69
TFP deceleration		−1.60%/yr
Spatial drag acceleration	Spatial	−0.11
Other factors	—	−1.49
R&D spending growth		+4.00%/yr
Entry expansion	Spatial	+1.33
Quality scaling	Spatial	+0.92
Fishing out	Non-spatial	+0.92
Burden of knowledge	Non-spatial	+0.83
Total productivity decline		−5.60%/yr
Spatial contribution		−2.36 **(42%)**
Non-spatial contribution		−3.24 (58%)

Notes: Decomposition of $g_{\Pi} = g_{g_{TFP}} - g_R = -5.6\%$/year (1948–2015) into spatial and non-spatial components. Spatial forces: TFP drag acceleration, entry expansion, quality scaling. Non-spatial forces: fishing out, burden of knowledge. $\theta = 0.69$ from National Center for Science and Engineering Statistics (2025) survey data; $g_H = 2.67\%$/yr from model identity (37). TFP drag from Table 3 Column 4. R&D components computed from equation (31) using balanced growth path rates $g_d = \frac{\alpha}{2} g_H$, $g_n = (1 - \frac{\alpha}{2}) g_H$.

On the TFP side, the regression (Table 3) estimates that the spatial drag from spreading out worsened from −0.041%/year (1948) to −0.136%/year (2015). The change in drag was −0.095 percentage points over 67 years, or 0.095/1.4 = 6.8% of the TFP growth deceleration. The spatial contribution to $g_{g_{TFP}}$ is 6.8% × −1.6% = −0.11%/year, common to all calibrations.

On the R&D side, the model decomposes 4.0%/year spending growth into four forces using equation (31): entry expansion ($(1 - \frac{\alpha}{2}) g_H$), quality scaling ($\theta \frac{\alpha}{2} g_H$), fishing out ($\theta \eta \frac{\alpha}{2} g_H$), and burden of knowledge ($(1-\theta)\alpha g_H$). Quality scaling is fundamentally spatial: the equilibrium condition $q = d/\gamma$ implies quality must keep pace with spacing. Fishing out reflects the non-spatial curvature of R&D costs (η), a technological constraint that applies even without spreading out. Table 5 reports the baseline decomposition and Table 6 reports sensitivity of the spatial share to α and η.

The baseline attributes 42% of the total research productivity decline to spatial forces. The burden of knowledge elasticity α is the dominant source of calibration uncertainty: varying α from 0.5 to 1.5 shifts the spatial share from 33% to 55%. The model requires $\alpha \in (0, 2)$ for consistency with the other evidence it explains — spreading out, rising quality, and rising variety all hold, but $\alpha \geq 2$ would imply a declining number of inventions (Corollary 4), contradicting the data. The variety-growth moment suggests $\alpha \leq 1$, corresponding to a spatial share of at least 42%. Lower α means fixed entry costs grow slowly with the frontier,

Table 6: Sensitivity of Spatial Share to α and η

	Spatial share (%)	
α	$\eta = 1$ (quadratic costs)	$\eta = 0.625$ (Guceri and Liu 2019)
0.50	55	58
0.75	48	51
1.00	**42**	**46**
1.25	37	41
1.50	33	37

Notes: Each cell reports the share of the -5.6%/year research productivity decline attributed to spatial forces (TFP drag acceleration + entry expansion + quality scaling). $\theta = 0.69$ throughout (National Center for Science and Engineering Statistics (2025)). $g_H = 4.0\%/[1 + \alpha(1 + \theta(\eta - 1))/2]$; when $\eta = 1$, this simplifies to $g_H = 4.0\%/(1 + \alpha/2)$ independent of θ. α governs the elasticity of fixed entry costs to frontier size ($f = \phi H^{\alpha}$); η governs R&D cost curvature ($c(q) \propto q^{1+\eta}$). TFP drag (-0.11%/yr) is common to all calibrations.

so more R&D growth is channeled through entry expansion (spatial) rather than burden of knowledge (non-spatial). R&D cost curvature η has a secondary effect: Guceri and Liu (2019) estimate a user cost elasticity of -1.6 from UK R&D tax credit shocks, implying $\eta = 0.625$, which raises the spatial share by 3–5 percentage points by shifting R&D growth from fishing out (non-spatial) toward quality scaling (spatial). The variable cost share θ matters only when $\eta \neq 1$; at $\eta = 1$, θ cancels from the decomposition entirely.

Across the range of calibrations, spatial forces account for roughly 33–58% of the research productivity decline, with the baseline at 42%.

7 Conclusion

Inventors are spreading out across an expanding idea space. We develop a spatial competition model in which inventors choose locations in idea space, trading off territorial pricing power against rising entry costs, generating equilibrium spreading out as the knowledge frontier expands. Using validated NLP representations applied to nearly two centuries of US patent claims (1836–2023), we document a substantial and consistent decline in contemporaneous invention similarity. This pattern is robust across spatial scales, within and between technology classes, after correcting for multi-patent entities, and is corroborated by over 150 years of declining patent interference rates.

A calibrated decomposition conditioned on the model's structure attributes about 40% of the Bloom et al. (2020) research productivity decline to spatial forces — spillover attenuation, adaptation drag, entry expansion, and quality scaling — with the spatial share ranging from

33% to 58% depending primarily on how fast the burden of knowledge scales with frontier size. The TFP regression corroborates Bloom et al. (2013) quasi-experimental elasticities, yielding nearly identical magnitudes from independent identification strategies. These results establish that the geography of idea space is a first-order determinant of aggregate innovation productivity.

Methodologically, we demonstrate that representation choice determines conclusions about invention similarity. Widely-used TF-IDF produces the opposite trend from validated neural embeddings. Our validation-based model selection framework — evaluating multiple NLP approaches against interference cases, human judgments, and patent classifications spanning 1850–2023 — provides a template for text-as-data measurement in economics.

Several directions remain open. Investigating causes of spreading out beyond the burden of knowledge — changes in research organization, funding structures, or technological opportunities — would deepen understanding of innovation dynamics. Firm-level heterogeneity in spreading-out behavior could reveal how the burden of knowledge differentially shapes innovation strategies across sectors.

Acknowledgments

We gratefully acknowledge support from an NBER Innovation Policy Grant. We also received excellent RA support from Josh Chapman, Cameron Fen, Annette Gailliot, Joseph Huang, Jake Moore, Isaac Rand, and Aaron Rosenbaum. We especially thank Matt Clancy for helpful conversations and conference discussion and Enrico Berkes for sharing data from CUSP. Finally, we received useful feedback from Darya Davydova, Gaétan de Rassenfosse, Luise Eisfeld, Deanna James, Semyon Malamud, Kyle Mangum, Roxana Mihet, Bryan Stuart, Yanos Zylberberg, workshop participants at Bristol, EPFL, LSE, the Philadelphia Fed, and Pittsburgh, and conference participants at the NBER Innovation Information Initiative Technical Working Group Meeting, TADA 2023, and NBER SI Innovation.

References

Aghion, Philippe and Peter Howitt (1992). "A Model of Growth Through Creative Destruction." *Econometrica*, pp. 323–351.

Akcigit, Ufuk, William R. Kerr, and Tom Nicholas (2017). "The Mechanics of Endogenous Innovation and Growth: Evidence from Historical US Patents." Working Paper. Harvard University.

Arora, Ashish, Sharon Belenzon, and Lia Sheer (2021). "Knowledge Spillovers and Corporate Investment in Scientific Research." *American Economic Review* 111.3, pp. 871–898. DOI: 10.1257/aer.20171742.

Arts, Sam, Nicola Melluso, and Reinhilde Veugelers (2025). "Beyond Citations: Measuring Novel Scientific Ideas and their Impact in Publication Text." *Review of Economics and Statistics*, pp. 1–33. DOI: 10.1162/rest_a_01561.

Ash, Elliott and Stephen Hansen (2023). "Text Algorithms in Economics." *Annual Review of Economics* 15.1, pp. 659–688. DOI: 10.1146/annurev-economics-082222-074352.

Atkin, David, Azam Chaudhry, Shamyla Chaudry, Amit K. Khandelwal, and Eric Verhoogen (2017). "Organizational Barriers to Technology Adoption: Evidence from Soccer-Ball Producers in Pakistan." *Quarterly Journal of Economics* 132.3, pp. 1101–1164. DOI: 10.1093/qje/qjx010.

Azoulay, Pierre, Christian Fons-Rosen, and Joshua S. Graff Zivin (2019). "Does Science Advance One Funeral at a Time?" *American Economic Review* 109.8, pp. 2889–2920. DOI: 10.1257/aer.20161574.

Bergeaud, Antonin, Adam B. Jaffe, and Dimitris Papanikolaou (2025). "Natural Language Processing and Innovation Research." Working Paper 33821. National Bureau of Economic Research. DOI: 10.3386/w33821.

Berkes, Enrico (2016). "Comprehensive Universe of U.S. Patents (CUSP): Data and Facts." Working paper. URL: https://www.dropbox.com/s/mwzwekr4f98l9sm/cusp_wp.pdf.

Berkes, Enrico and Ruben Gaetani (2020). "The Geography of Unconventional Innovation." *Economic Journal* 131.636, pp. 1466–1514. DOI: 10.1093/ej/ueaa111.

Bessen, James, Peter Neuhäusler, John L. Turner, and Jonathan Williams (2018). "Trends in Private Patent Costs and Rents for Publicly-Traded United States Firms." *International Review of Law and Economics* 56, pp. 53–69. DOI: 10.1016/j.irle.2018.07.001.

Bisbee, James, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson (2024). "Synthetic Replacements for Human Survey Data? The Perils of Large Language Models." *Political Analysis*, pp. 1–16. DOI: 10.1017/pan.2024.5.

Bloom, Nicholas, Charles I. Jones, John Van Reenen, and Michael Webb (2020). "Are Ideas Getting Harder to Find?" *American Economic Review* 110.4, pp. 1104–1144. DOI: 10.1257/aer.20180338.

Bloom, Nicholas, Mark Schankerman, and John Van Reenen (2013). "Identifying Technology Spillovers and Product Market Rivalry." *Econometrica* 81.4, pp. 1347–1393. DOI: 10.3982/ECTA9466.

Bochkay, Khrystyna, Stephen V. Brown, Andrew J. Leone, and Jennifer Wu Tucker (2023). "Textual Analysis in Accounting: What's Next?" *Contemporary Accounting Research* 40.2, pp. 765–805. DOI: 10.1111/1911-3846.12825.

Bryan, Kevin A. and Jorge Lemus (2017). "The Direction of Innovation." *Journal of Economic Theory* 172, pp. 247–272. DOI: 10.1016/j.jet.2017.09.005.

Butler, John P. (1993). "Patent Interference Case Files: 1838–1900." Special List 59. National Archives and Records Administration. URL: https://www.archives.gov/research/guide-fed-records/groups/241.html.

Calvert, Ian A. and Michael Sofocleous (1982). "Three Years of Interference Statistics." *Journal of the Patent Office Society* 64, p. 699.

— (1986). "Interference Statistics for Fiscal Years 1983 to 1985." *Journal of the Patent & Trademark Office Society* 68, p. 385.

— (1989). "Interference Statistics for Fiscal Years 1986 to 1988." *Journal of the Patent & Trademark Office Society* 71, p. 399.

— (1992). "Interference Statistics for Fiscal Years 1989 to 1991." *Journal of the Patent & Trademark Office Society* 74, p. 822.

— (1995). "Interference Statistics for Fiscal Years 1992 to 1994." *Journal of the Patent & Trademark Office Society* 77, p. 417.

Carmody, Sean (2023). *ngramr: Retrieve and Plot Google n-Gram Data*. Manual.

Carnehl, Christoph and Johannes Schneider (2025). "A Quest for Knowledge." *Econometrica* 93.2, pp. 623–659. DOI: 10.3982/ECTA22144.

Cer, Daniel, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil (2018). "Universal Sentence Encoder for English." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, pp. 169–174. DOI: 10.18653/v1/D18-2029.

Champsaur, Paul and Jean-Charles Rochet (1989). "Multiproduct Duopolists." *Econometrica* 57.3, pp. 533–557.

Chiopris, Caterina (2024). "The Diffusion of Ideas." Working Paper. URL: https://www.caterinachiopris.com/_files/ugd/b45409_ba6a9e005f5c428ba55811d3dc219580.pdf.

Clancy, Matthew S. (2018). "Inventing by Combining Pre-Existing Technologies: Patent Evidence on Learning and Fishing Out." *Research Policy* 47.1, pp. 252–265. DOI: 10.1016/j.respol.2017.10.015.

Cohen, Lauren, Umit G. Gurun, and Scott Duke Kominers (2019). "Patent Trolls: Evidence from Targeted Firms." *Management Science* 65.12, pp. 5461–5486. DOI: 10.1287/mnsc.2018.3147.

Dasgupta, Partha and Eric Maskin (1987). "The Simple Economics of Research Portfolios." *Economic Journal* 97.387, pp. 581–595. DOI: 10.2307/2232925.

Dell, Melissa (2024). *Deep Learning for Economists*. arXiv: 2407.15339 [econ.GN].

Di Simone, Daniel V., James B. Gambell, and Charles F. Gareau (1963). "Characteristics of Interference Practice." *Journal of the Patent Office Society* 45, pp. 503–591.

Dixit, Avinash K. and Joseph E. Stiglitz (1977). "Monopolistic Competition and Optimum Product Diversity." *American Economic Review* 67.3, pp. 297–308.

Dominguez-Olmedo, Ricardo, Moritz Hardt, and Celestine Mendler-Dunner (2024). *Questioning the Survey Responses of Large Language Models*. arXiv: 2306.07951 [cs.CL].

Feng, Sijie (2020). "The Proximity of Ideas: An Analysis of Patent Text Using Machine Learning." *PLOS ONE* 15.7, pp. 1–19. DOI: 10.1371/journal.pone.0234880.

Fleming, Lee (2001). "Recombinant Uncertainty in Technological Search." *Management Science* 47.1, pp. 117–132. DOI: 10.1287/mnsc.47.1.117.10671.

Fudenberg, Drew and Jean Tirole (1991). *Game Theory*. Cambridge, MA: MIT Press.

Ganguli, Ina, Jeffrey Lin, and Nicholas Reynolds (2020). "The Paper Trail of Knowledge Spillovers: Evidence from Patent Interferences." *American Economic Journal: Applied Economics* 12.2, pp. 278–302. DOI: 10.1257/app.20180017.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy (2019). "Text as Data." *Journal of Economic Literature* 57.3, pp. 535–574. DOI: 10.1257/jel.20181020.

Ghosh, Mainak, Sebastian Erhardt, Michael E. Rose, Erik Buunk, and Dietmar Harhoff (2024). *PaECTER: Patent-Level Representation Learning using Citation-Informed Transformers*. arXiv: 2402.19411 [cs.IR].

Goli, Ali and Amandeep Singh (2024). "Frontiers: Can Large Language Models Capture Human Preferences?" *Marketing Science* 43.4, pp. 709–722. DOI: 10.1287/mksc.2023.0306.

Gordon, Robert J. (2016). *The Rise and Fall of American Growth: The US Standard of Living since the Civil War*. Princeton, NJ: Princeton University Press.

Grimmer, J., M.E. Roberts, and B.M. Stewart (2022). *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton University Press.

Grossman, Gene M. and Elhanan Helpman (1993). *Innovation and Growth in the Global Economy*. MIT Press.

Guceri, Irem and Li Liu (2019). "Effectiveness of Fiscal Incentives for R&D: Quasi-Experimental Evidence." *American Economic Journal: Economic Policy* 11.1, pp. 266–291. DOI: 10.1257/pol.20170403.

Hall, Bronwyn H. (2009). "Business and Financial Method Patents, Innovation, and Policy." *Scottish Journal of Political Economy* 56.4, pp. 443–473. DOI: 10.1111/j.1467-9485.2009.00493.x.

Hall, Bronwyn H., Adam Jaffe, and Manuel Trajtenberg (2005). "Market Value and Patent Citations." *RAND Journal of Economics* 36.1, pp. 16–38.

Hall, Bronwyn H. and Rosemarie Ham Ziedonis (2001). "The Patent Paradox Revisited: An Empirical Study of Patenting in the U.S. Semiconductor Industry, 1979–1995." *RAND Journal of Economics* 32.1, pp. 101–128.

Hippel, Eric von (1994). ""Sticky Information" and the Locus of Problem Solving: Implications for Innovation." *Management Science* 40.4, pp. 429–439. DOI: 10.1287/mnsc.40.4.429.

Hirschey, Mark, Hilla Skiba, and M. Babajide Wintoki (2012). "The Size, Concentration and Evolution of Corporate R&D Spending in US Firms from 1976 to 2010: Evidence and Implications." *Journal of Corporate Finance* 18.3, pp. 496–518. DOI: 10.1016/j.jcorpfin.2012.02.002.

Hopenhayn, Hugo and Francesco Squintani (2021). "On the Direction of Innovation." *Journal of Political Economy* 129.7, pp. 1991–2022. DOI: 10.1086/714093.

Howitt, Peter (1999). "Steady Endogenous Growth with Population and R&D Inputs Growing." *Journal of Political Economy* 107.4, pp. 715–730.

Hsieh, Cheng-Yu, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister (2023). *Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes*. arXiv: 2305.02301 [cs.CL].

Jaffe, Adam B. (1986). "Technological Opportunity and Spillovers of R&D: Evidence from Firms' Patents, Profits, and Market Value." *American Economic Review* 76.5, pp. 984–1001.

Jaffe, Adam B., Manuel Trajtenberg, and Rebecca Henderson (1993). "Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations." *Quarterly Journal of Economics* 108.3, pp. 577–598. DOI: 10.2307/2118401.

Jones, Benjamin F. (2009). "The Burden of Knowledge and the "Death of the Renaissance Man": Is Innovation Getting Harder?" *Review of Economic Studies* 76.1, pp. 283–317. DOI: 10.1111/j.1467-937X.2008.00531.x.

Jones, Charles I. (1995). "R&D-Based Models of Economic Growth." *Journal of Political Economy* 103.4, pp. 759–784.

Kantor, Shawn and Alexander Whalley (2014). "Knowledge Spillovers from Research Universities: Evidence from Endowment Value Shocks." *Review of Economics and Statistics* 96.1, pp. 171–188. DOI: 10.1162/REST_a_00357.

Kelly, Bryan, Dimitris Papanikolaou, Amit Seru, and Matt Taddy (2021). "Measuring Technological Innovation over the Long Run." *American Economic Review: Insights* 3.3, pp. 303–320. DOI: 10.1257/aeri.20190499.

Klemperer, Paul and A. Jorge Padilla (1997). "Do Firms' Product Lines Include Too Many Varieties?" *RAND Journal of Economics* 28.3, pp. 472–488.

Kogan, Leonid, Dimitris Papanikolaou, Amit Seru, and Noah Stoffman (2017). "Technological Innovation, Resource Allocation, and Growth." *Quarterly Journal of Economics* 132.2, pp. 665–712. DOI: 10.1093/qje/qjw040.

Kortum, Samuel S. (1997). "Research, Patenting, and Technological Change." *Econometrica* 65.6, pp. 1389–1419.

Kusupati, Aditya, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi (2024). *Matryoshka Representation Learning*. arXiv: 2205.13147 [cs.LG].

Lamantia, Fabio and Mario Pezzino (2016). "R&D Spillovers on a Salop Circle." *Managerial and Decision Economics* 37.7, pp. 485–494. DOI: 10.1002/mde.2734.

Le, Quoc and Tomas Mikolov (2014). "Distributed Representations of Sentences and Documents." In: *Proceedings of the 31st International Conference on Machine Learning*. Ed. by Eric P. Xing and Tony Jebara. Vol. 32. Proceedings of Machine Learning Research 2. PMLR, pp. 1188–1196. URL: https://proceedings.mlr.press/v32/le14.html.

Lee, Jieh-Sheng and Jieh Hsiang (2019). *PatentBERT: Patent Classification with Fine-Tuning a Pre-Trained BERT Model*. arXiv: 1906.02124 [cs.CL].

Li, Zehan, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang (2023). *Towards General Text Embeddings with Multi-Stage Contrastive Learning*. arXiv: 2308.03281 [cs.CL].

Lucking, Brian, Nicholas Bloom, and John Van Reenen (2019). "Have R&D Spillovers Declined in the 21st Century?" *Fiscal Studies* 40.4, pp. 561–590. DOI: 10.1111/1475-5890.12195.

McInnes, L., J. Healy, and J. Melville (2018). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. arXiv: 1802.03426 [stat.ML].

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (2013). *Efficient Estimation of Word Representations in Vector Space*. arXiv: 1301.3781 [cs.CL].

Miller, George A. (1992). "WordNet: A Lexical Database for English." In: *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23–26, 1992*.

Monath, Nicholas, Christina Jones, and Sarvo Madhavan (2021). "PatentsView: Disambiguating Inventors, Assignees, and Locations." Tech. rep. American Institutes for Research. URL: https://s3.amazonaws.com/data.patentsview.org/documents/PatentsView_Disambiguation_Methods_Documentation.pdf.

Monroe, Burt L., Michael P. Colaresi, and Kevin M. Quinn (2017). "Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict." *Political Analysis* 16.4, pp. 372–403. DOI: 10.1093/pan/mpn018.

Murata, Yasusada, Ryo Nakajima, Ryosuke Okamoto, and Ryuichi Tamura (2014). "Localized Knowledge Spillovers and Patent Citations: A Distance-Based Approach." *Review of Economics and Statistics* 96.5, pp. 967–985. DOI: 10.1162/REST_a_00422.

Nard, Craig Allen (2010). "Legal Forms and the Common Law of Patents." *Boston University Law Review* 90.1, pp. 51–108. URL: https://www.bu.edu/law/journals-archive/bulr/documents/nard.pdf.

National Center for Science and Engineering Statistics (2025). *Business Enterprise Research and Development (BERD) Survey*. URL: https://ncses.nsf.gov/surveys/business-enterprise-research-development/2023.

Park, Michael, Erin Leahey, and Russell J. Funk (2023). "Papers and Patents Are Becoming Less Disruptive Over Time." *Nature* 613.7942, pp. 138–144. DOI: 10.1038/s41586-022-05543-x.

Peretto, Pietro F. (1998). "Technological Change and Population Growth." *Journal of Economic Growth* 3.4, pp. 283–311. DOI: 10.1023/A:1009799405456.

— (2018). "Robust Endogenous Growth." *European Economic Review* 108, pp. 49–77. DOI: 10.1016/j.euroecorev.2018.02.006.

Reimers, Nils and Iryna Gurevych (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. arXiv: 1908.10084 [cs.CL].

Romer, Paul M. (1990). "Endogenous Technological Change." *Journal of Political Economy* 98.5, S71–S102.

Salop, Steven C. (1979). "Monopolistic Competition with Outside Goods." *Bell Journal of Economics*, pp. 141–156.

Schnoebelen, Tyler, Julia Silge, and Alex Hayes (2022). *tidylo: Weighted Tidy Log Odds Ratio*. Manual.

Scotchmer, Suzanne (1991). "Standing on the Shoulders of Giants: Cumulative Research and the Patent Law." *Journal of Economic Perspectives* 5.1, pp. 29–41.

Smith, Noah A. (2020). "Contextual Word Representations: Putting Words into Computers." *Communications of the ACM* 63.6, pp. 66–74. DOI: 10.1145/3347145.

Sparck Jones, K. (1972). "A Statistical Interpretation of Term Specificity and its Application in Retrieval." *Journal of Documentation* 28.1, pp. 11–21. DOI: 10.1108/eb026526.

Teece, David J. (1977). "Technology Transfer by Multinational Firms: The Resource Cost of Transferring Technological Know-How." *Economic Journal* 87.346, pp. 242–261. DOI: 10.2307/2232084.

Thompson, Peter and Melanie Fox-Kean (2005). "Patent Citations and the Geography of Knowledge Spillovers: A Reassessment." *American Economic Review* 95.1, pp. 450–460. DOI: 10.1257/0002828053828509.

U.S. Patent and Trademark Office (2023). *Data Download Tables*. PatentsView. URL: https://patentsview.org/download/data-download-tables.

Verhoeven, Dennis, Jurriën Bakker, and Reinhilde Veugelers (2016). "Measuring Technological Novelty with Patent-Based Indicators." *Research Policy* 45.3, pp. 707–723. DOI: 10.1016/j.respol.2015.11.010.

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou (2024). "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models." In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*. NIPS '22. Curran Associates Inc.

Weitzman, Martin L. (1998). "Recombinant Growth." *Quarterly Journal of Economics* 113.2, pp. 331–360.

A Methodological Applications: Consequences of Representation Choice

Having established GTE's superiority through systematic validation (Section 3) and demonstrated its robustness across multiple dimensions of our main finding (Section 4), we now illustrate the practical consequences of using unvalidated representations for innovation research.

We revisit Kelly et al. (2021)'s influential analysis of breakthrough inventions — patents dissimilar from prior art but similar to subsequent innovations. This application demonstrates how representation choice can meaningfully affect both interpretation and robustness even when qualitative conclusions align.

A.1 Methodology

Kelly et al. (2021) employ TF-BIDF (backward-looking TF-IDF) to identify breakthrough patents, defined as those in the top 10% of a measure capturing dissimilarity from the past five years combined with similarity to the subsequent five years. They residualize this measure on year fixed effects and normalize breakthrough counts by US population to construct a time series of breakthrough invention rates from 1840 to 2010.

We replicate their analysis using both TF-BIDF and GTE representations, examining sensitivity along three dimensions: (i) representation choice (TF-BIDF versus GTE), (ii) residualization on year fixed effects, and (iii) normalization by population versus total patents. This comprehensive approach isolates the impact of representation choice while examining other methodological decisions.

A.2 Results

Our replication using TF-BIDF (Figure 13, Panel A) closely mirrors Kelly et al. (2021)'s key finding (in their Figure 4a) that breakthrough patent rates per capita fluctuated before 1980, then increased sharply through 2010 (despite minor methodological differences).[32]

However, this TF-BIDF-based pattern proves highly sensitive to methodological choices (Figure 13, Panels B-D). Normalizing by total patents rather than population reveals that the peak breakthrough rate occurred before 1870, not recently (Panel C). Omitting year fixed

[32]Our replication differs slightly from Kelly et al. (2021) in data source (ProQuest patent claims versus Google Patents full text) and IDF computation (five-year rolling window versus patent-specific backward lookups for computational efficiency). These differences do not affect qualitative patterns.

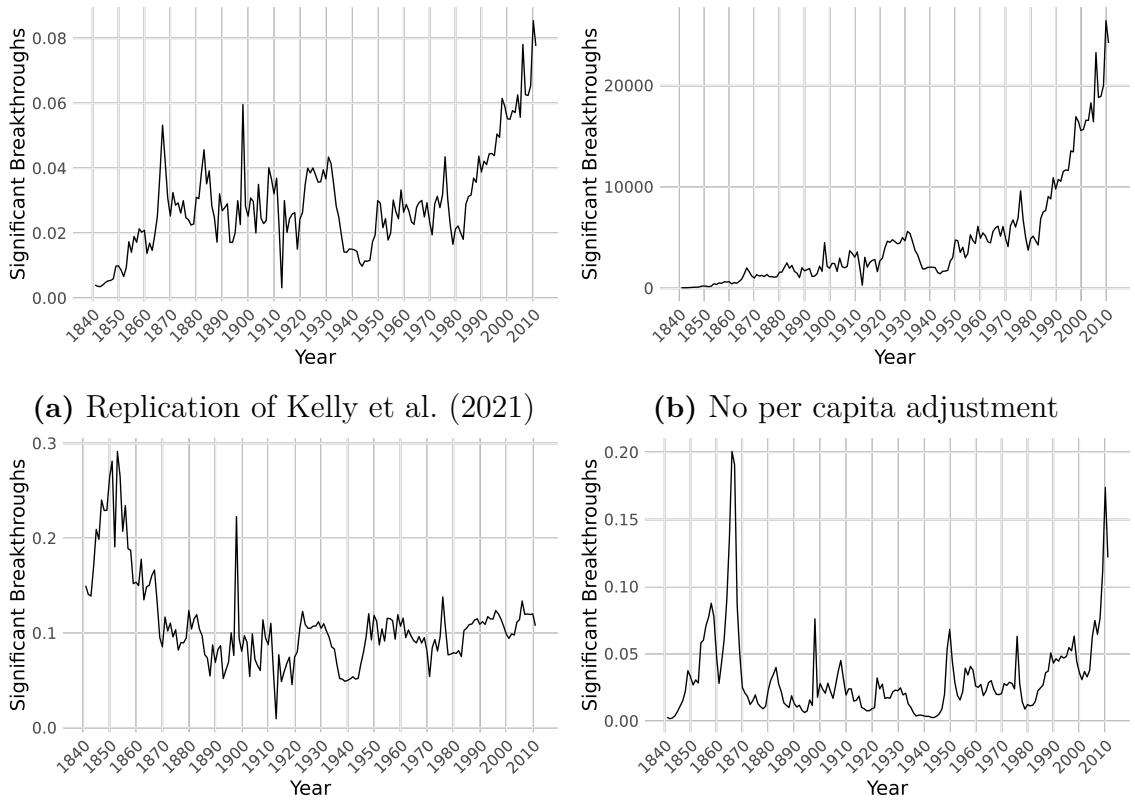
effects produces a qualitatively different historical pattern with two distinct peaks (Panel D). This sensitivity raises concerns about the robustness of conclusions drawn from unvalidated representations.

Figure 14 presents the same analysis using our validated GTE representations. Several important patterns emerge. One, GTE confirms the qualitative finding of elevated breakthrough rates in recent decades, lending support to Kelly et al. (2021)'s central conclusion. Two, GTE-based measures prove far more robust to methodological choices: omitting year fixed effects (Panel D) produces much less dramatic changes compared to TF-BIDF, and different normalization approaches (Panels B-C) yield more consistent historical patterns.

A.3 Implications

This analysis illustrates the practical value of validation-based model selection. While both TF-BIDF and GTE support Kelly et al. (2021)'s qualitative finding of elevated recent breakthrough rates, the representations differ meaningfully. GTE's robustness to methodological choices increases confidence in the findings, whereas TF-BIDF's sensitivity raises questions about which specification to trust.

More broadly, this exercise demonstrates why our validation framework matters for applied work. Researchers using TF-IDF might reasonably conclude it is validated — it correlates with patent classes and performs better than chance on our validation tasks. But comparative evaluation reveals it performs substantially worse than alternatives and produces results sensitive to seemingly innocuous methodological choices. Validation-based selection thus provides not only more accurate measures but also more reliable foundations for empirical analysis.



(a) Replication of Kelly et al. (2021)

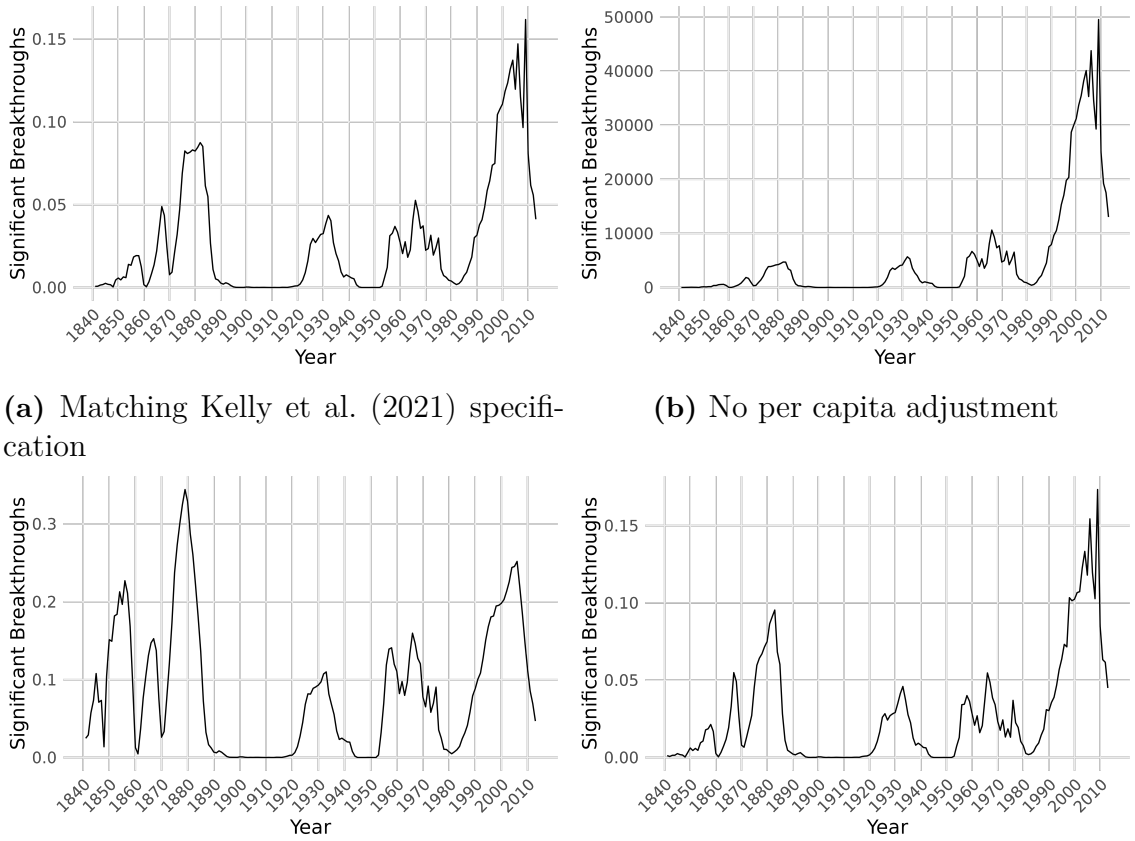
(b) No per capita adjustment

(c) Normalized by total patents

(d) No year fixed effects

Figure 13: Breakthrough Inventions Using TF-BIDF: Sensitivity Analysis

These panels show breakthrough invention rates using TF-BIDF representations under different specifications. The results are highly sensitive to normalization and residualization choices, raising concerns about robustness.



(a) Matching Kelly et al. (2021) specification

(b) No per capita adjustment

(c) Normalized by total patents

(d) No year fixed effects

Figure 14: Breakthrough Inventions Using GTE: Sensitivity Analysis

These panels show breakthrough invention rates using validated GTE representations. While confirming elevated recent rates, GTE reveals similar historical peaks and demonstrates greater robustness to specification choices compared to TF-BIDF.

Supplemental Appendix
For Online Publication Only

Appendix S1 Equilibrium Existence and Technical Conditions

S1.1 Second-order conditions and no spatial deviation

Pricing second-order condition From the revenue function $R_i(p_i) = 2p_i\tilde{h}(p_i)$ where $\frac{\partial \tilde{h}}{\partial p_i} = -\frac{1}{2\tau}$:

$$\frac{\partial^2 R_i}{\partial p_i^2} = 2\frac{\partial \tilde{h}}{\partial p_i} + 2\frac{\partial \tilde{h}}{\partial p_i} = 4\frac{\partial \tilde{h}}{\partial p_i} = -\frac{2}{\tau} < 0 \tag{S1.2}$$

The revenue function is strictly concave in price, so the first-order condition $p = \tau d$ is indeed a maximum.

Quality second-order condition From the profit function $\pi_i = R_i(q_i) - c(q_i) - f$, where $\frac{\partial R_i}{\partial q_i} = d$ (constant) and $\frac{\partial c}{\partial q_i} = \gamma q_i$:

$$\frac{\partial^2 \pi_i}{\partial q_i^2} = 0 - \gamma < 0 \tag{S1.3}$$

The profit function is strictly concave in quality, so the first-order condition $q = d/\gamma$ is a maximum.

No spatial deviation In symmetric equilibrium, no inventor gains by unilaterally relocating. With linear spillover decay, moving distance ϵ toward one neighbor (from distance d to $d - \epsilon$) while moving away from the other (from d to $d + \epsilon$) leaves total spillovers unchanged:

$$\frac{1}{2}\beta\left(1 - \frac{d-\epsilon}{\lambda}\right)q + \frac{1}{2}\beta\left(1 - \frac{d+\epsilon}{\lambda}\right)q = \beta q\left(1 - \frac{d}{\lambda}\right) \tag{S1.4}$$

The linear spillover function ensures gains from proximity to one neighbor exactly offset losses from distance to the other. Similarly, demand effects are symmetric: boundaries with each neighbor shift in opposite directions, leaving first-order profits unchanged. Linear spillovers thus guarantee that symmetric spacing constitutes a Nash equilibrium in locations.

S1.2 Spillover reach condition

With linear spillovers, the spillover function is active only when $d < \lambda$. In equilibrium, $d^* = \sqrt{\frac{\phi H}{\tau - \frac{1}{2\gamma}}}$, so spillovers remain active when H is not too large relative to spillover reach

λ. Specifically, we require:

$$H < \lambda^2 \left(\tau - \frac{1}{2\gamma}\right) / \phi \tag{S1.5}$$

This ensures $d^* < \lambda$, so that the spillover mechanism operates throughout. When H grows very large and this condition is violated, the model transitions to a no-spillover regime where $Q = q$. We focus on the spillover-active regime, which is most relevant for understanding how spreading out affects productivity when knowledge flows remain present but weakening.

S1.3 Full Coverage Constraint

Our equilibrium characterization assumes that all downstream firms adopt a technology from some inventor — i.e., *full coverage*. To verify this, we must check that even the most distant firm prefers adoption to its outside option.

Adoption decision. A firm at distance h from inventor i that adopts the technology obtains total surplus (net of licensing fee):

$$\text{Total surplus with adoption} = Q_i - p_i - \tau h \tag{S1.6}$$

Full coverage condition. For all downstream firms to adopt, even the boundary firm (at distance $d/2$ from its nearest inventor) must weakly prefer adoption to the baseline productivity of zero (log TFP = 0, or TFP = 1):

$$Q - p - \frac{\tau d}{2} \geq 0 \tag{S1.7}$$

Verification. Substituting equilibrium values $Q = \frac{d}{\gamma}\left(1 + \beta - \frac{\beta d}{\lambda}\right)$ and $p = \tau d$:

$$\frac{d}{\gamma}\left(1 + \beta - \frac{\beta d}{\lambda}\right) - \tau d - \frac{\tau d}{2} \geq 0 \tag{S1.8}$$

$$\frac{d}{\gamma}\left(1 + \beta - \frac{\beta d}{\lambda}\right) \geq \frac{3\tau d}{2} \tag{S1.9}$$

$$\frac{1}{\gamma}\left(1 + \beta - \frac{\beta d}{\lambda}\right) \geq \frac{3\tau}{2} \tag{S1.10}$$

This simplifies to a parameter restriction:

$$\boxed{\frac{1}{\gamma}\left(1 + \beta - \frac{\beta d}{\lambda}\right) \geq \frac{3\tau}{2}} \tag{S1.11}$$

Interpretation. Full coverage requires that the quality delivered (including spillover

benefits) is sufficiently high relative to the price and adaptation costs. The left side represents the effective quality-cost ratio, accounting for spillovers. The right side is the adaptation burden faced by boundary firms.

When is this satisfied? The constraint is more easily satisfied when:

- R&D costs are low (γ small): Inventors can afford to produce high quality
- Spillovers are strong (β large, λ large): Realized quality Q is boosted by neighbors
- Spacing is small (d small): Spillovers are stronger and boundary firms are closer
- Adaptation costs are low (τ small): Boundary firms don't lose much productivity

Relationship to spreading-out condition. The spreading-out condition is $\tau\gamma > \frac{1}{2}$, or equivalently $\gamma > \frac{1}{2\tau}$. Rearranging the full coverage condition:

$$\gamma < \frac{2}{3\tau}\left(1 + \beta - \frac{\beta d}{\lambda}\right) \tag{S1.12}$$

For both conditions to hold simultaneously, we need:

$$\frac{1}{2\tau} < \gamma < \frac{2}{3\tau}\left(1 + \beta - \frac{\beta d}{\lambda}\right) \tag{S1.13}$$

This parameter region is non-empty when:

$$1 + \beta - \frac{\beta d}{\lambda} > \frac{3}{4} \tag{S1.14}$$

which always holds for any $\beta > 0$ and $\lambda > 0$, since the left side is strictly greater than 1. Therefore, the two conditions are compatible.

S1.4 Zero-profit condition.

The zero-profit condition $d^2(\tau - \frac{1}{2\gamma}) = \phi H$ has a unique positive solution:

$$d^* = \sqrt{\frac{\phi H}{\tau - \frac{1}{2\gamma}}} \tag{S1.15}$$

provided $\tau\gamma > \frac{1}{2}$. This is exactly the spreading-out condition from Proposition 2.

Uniqueness of symmetric equilibrium. Given spacing d, the pricing and quality first-order conditions uniquely determine (p^*, q^*) by strict concavity of profit functions. The

zero-profit condition then uniquely determines spacing d^* (and thus $n^* = H/d^*$) given H. The symmetric equilibrium is therefore the unique solution to the system of first-order and zero-profit conditions.

Asymmetric equilibria. We do not rule out existence of asymmetric equilibria where inventors choose heterogeneous qualities, prices, or irregular spacing. Characterizing these would require solving boundary value problems with heterogeneous agents, which is beyond our scope. We focus on the symmetric equilibrium as the natural focal point: it is stable under small perturbations, analytically tractable, and captures the key economic forces.

Appendix S2 Technical Details of NLP Models

This section provides technical details about the NLP models evaluated in Section 3.

S2.1 Model Architectures and Training

Embedding sizes vary considerably across models. A doc2vec model typically produces embeddings of 100-300 dimensions, USE generates 512-dimensional vectors, S-BERT and GTE yield larger embeddings of 768 or 1,024 dimensions, and PaECTER, designed specifically for patents, uses 1,024-dimensional embeddings. OpenAI embeddings have dimensions of 1,536 by default, but they use Matryoshka representation learning technology, allowing reductions in embedding size with limited loss in performance (Kusupati et al. 2024).

The objective functions and training processes differ significantly across models. Doc2vec models use two architectures: Distributed Memory (PV-DM), which predicts a target word from surrounding context and a document vector, and Distributed Bag of Words (PV-DBOW), which predicts context words from the document vector alone. In contrast, USE and subsequent models use multi-task learning, jointly training on multiple objectives including paraphrase identification and sentence similarity. The second stage typically includes paraphrase identification and sentence similarity tasks, with the explicit goal of producing embeddings that are broadly applicable and semantically meaningful.

GTE utilizes a contrastive learning objective (Li et al. 2023), which explicitly aims to both bring similar sentences closer and different ones further apart. PaECTER adapts this approach to patents, fine-tuning on citation data. Details of the OpenAI embedding models are proprietary, but the technology and training data are likely similar to that underlying the large language model GPT-4.

Appendix S3 Validation Framework Details

This section provides more discussion about the validation framework outlined in Section 3.

The central challenge in selecting among NLP representations is that we cannot directly observe the true similarity between inventions. This creates a fundamental evaluation problem: how can we determine which representation best captures technological similarity when similarity itself is unobservable?

Figure S3.1 illustrates our solution through a four-step pipeline. Steps 1 and 2 show how patent text gets mapped to numerical representations (Step 1) and then to similarity measures (Step 2). Our key contribution is Step 3: validation-based model selection using external ground truth — independent measures of technological similarity that do not rely on the text representations we seek to validate. For each validation task, we compare similarity measures derived from different NLP representations against these external benchmarks to identify which representations align best with independent assessments of technological proximity.

The pipeline's Steps 1 and 2 can produce different similarity measures from the same patent text — as Figure S3.1 illustrates with representations A, B, and C yielding different similarity patterns. Step 3 addresses this multiplicity by evaluating each representation using external validation:

$$V^j(m) = S^j\left(1 - d^m(\mathbf{p}), g^j(\mathbf{p})\right) \tag{S3.16}$$

where $1 - d^m(\mathbf{p})$ measures similarities using representation m, $g^j(\mathbf{p})$ provides ground truth from validation task j, and S^j quantifies the correspondence between the two measures.

For example, in our interference validation task, g^j creates binary indicators for whether patent pairs were in interference, while $1-d^m_{ik} \equiv \frac{C^m_i \cdot C^m_k}{||C^m_i||||C^m_k||}$ computes cosine similarity between patents i and k using representation m. The score function S^j measures how well similarity rankings predict interference status using Receiver Operating Characteristic Area Under Curve (ROC AUC) or Precision-Recall Area Under Curve (PR AUC). Only after validation (Step 3) do we proceed to Step 4: computing our final measure of invention similarity for testing the spreading-out prediction.

This framework addresses the core problem illustrated in Figure 1: that different representations can yield different conclusions about the same underlying similarity patterns. Rather than assuming any particular representation correctly captures technological similarity, we evaluate each method against multiple independent benchmarks. Representations that consistently align with external ground truth across different validation tasks are more likely to provide reliable measures for economic analysis.

Text of patent 1
Text of patent 2
Text of patent 3

Step 1: Numerical representations in three idea spaces $(C_i(m_k), k \in \{A, B, C\})$

Repr. A

$$\begin{bmatrix} 0.44 & 0.03 & 0.55 & 0.44 \\ 0.42 & 0.33 & 0.2 & 0.62 \\ 0.3 & 0.27 & 0.62 & 0.53 \end{bmatrix}$$

Repr. B

$$\begin{bmatrix} 0.13 & 0.51 & 0.18 \\ 0.85 & 0.49 & 0.85 \\ 0.51 & 0.07 & 0.43 \end{bmatrix}$$

Repr. C

$$\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Step 2: Measurements of pairwise similarities $(Sim^{m_k}(p_i, p_j))$

Repr. A

Pat 1	Pat 2	Cos. Sim.
1	2	0.82
1	3	0.94
2	3	0.87

Repr. B

Pat 1	Pat 2	Cos. Sim.
1	2	0.71
1	3	0.48
2	3	0.96

Repr. C

Pat 1	Pat 2	Cos. Sim.
1	2	1
1	3	0
2	3	0

Step 3: Validation-based selection $\left(V^l(m_k), l \in \{(i), (ii), (iii)\}\right)$

Task (i)

Repr.	Perf.	Rank
Repr. A	0.91	1
Repr. B	0.87	2
Repr. C	0.84	3
Baseline	0.51	4

Task (ii)

Repr.	Perf.	Rank
Repr. A	0.46	1
Repr. B	0.23	2
Repr. C	0.18	3
Baseline	0.03	4

Task (iii)

Repr.	Perf.	Rank
Repr. A	0.85	2
Repr. B	0.93	1
Repr. C	0.73	3
Baseline	0.05	4

Step 4: Compute downstream measure based on the best representation:

For example, "Breakthrough" patents $(q^m(p_i))$ (Kelly et al. 2021) or average patent pair similarity $(q^m(p_i, p_j))$ by year (this paper).

Figure S3.1: Overview of the NLP pipeline

Our approach to validation differs from some prior literature in that it is intrinsically linked with model selection. In this respect, it conforms with methods in forecasting and machine learning, where it is often acknowledged that we cannot select the best model *a priori*, necessitating a structured selection procedure.[33] Our work demonstrates that this principle extends to NLP applications, where model selection can have substantial effects on results and interpretations.

The validation approach also reveals what different representations actually capture. Some methods may excel at detecting broad technological categories while others better identify fine-grained technical relationships. Understanding these differences allows us to select methods most appropriate for testing our theory's prediction about declining invention similarity over time.

When validation tasks disagree about which representation performs best, we weight results based on task relevance for our specific application, the reliability of each task's ground truth, and the magnitude of performance differences. This structured approach moves beyond arbitrary model selection to evidence-based choices grounded in multiple independent assessments of representation quality.

[33]Model selection is a well-established practice in econometrics and forecasting, often using criteria such as the Akaike Information Criterion (AIC). In machine learning, out-of-sample testing is commonly used for model selection, where models are evaluated on data not used for training. The "winning" model is typically determined by a score function, such as root mean squared error. Ash and Hansen (2023) provide examples outside of innovation economics where different text representations lead to divergent conclusions.

Appendix S4 Detailed Validation Task Results

This section provides the detailed tables and figures underlying the validation results summarized in Section 3.4.

S4.1 Interference Task Details

Patent interferences were US patent office administrative proceedings that decided the priority of invention when two or more independent parties claimed to have invented the same thing at the same time. A specialized patent examiner initiated an interference upon encountering another pending US patent application containing the "same patentable invention" (37 CFR § 1.601). We use 215 interference cases decided between 2001 and 2014, obtained from the US patent office's e-FOIA Reading Room and encoded by Ganguli et al. (2020). The 215 cases correspond with 440 distinct patent applications, producing 96,580 application pairs of which 322 are interfering pairs.

We evaluate each representation's ability to classify interfering versus non-interfering application pairs using four complementary metrics: F1 score (harmonic mean of precision and recall), F10 score (weighting recall by $\beta^2 = 100$ relative to precision), ROC AUC, and PR AUC.

PaECTER achieves the highest F1 score (67%), followed closely by GTE (65%) and OpenAI embeddings (63%), significantly outperforming S-BERT (56%) and TF-IDF (49%). When prioritizing interference detection (F10 score), GTE, PaECTER, and OpenAI embeddings perform nearly identically, retrieving 90%, 90%, and 89% of true interferences respectively. The top three models generate 1.6–2.7 times fewer false positives than S-BERT and 2.8–4.7 times fewer than TF-IDF. The threshold-independent metrics confirm these patterns: PaECTER leads PR AUC at 0.65, followed by GTE (0.64) and OpenAI (0.62), with S-BERT (0.52) and TF-IDF (0.45) trailing substantially.

S4.2 Human Judgment Task Details

We evaluate the four remaining competitive models (PaECTER, GTE, S-BERT, TF-IDF) using relative similarity judgments from non-expert annotators on historical patents (1880–1920). Annotators saw two text fragments per patent: the "improvement in" statement and the first 500 characters of claims. Four annotators each completed 100 comparisons. See Online Appendix S5 for full instructions.

GTE demonstrates the strongest alignment ($\beta_1 = 0.62$), followed by S-BERT (0.54), PaECTER (0.51), and TF-IDF (0.35). The relative performance ordering differs from the

Table S4.1: Rankings: Threshold-based Metrics

(a) Separate F1-max. thresh.

Rank	Repr.	TP	FP	F1
1	PaECTER	168	58	0.668
2	GTE	186	114	0.645
3	OpenAI	182	123	0.625
4	S-BERT	143	90	0.561
5	TF-IDF	111	64	0.491
6	USE	85	58	0.405
7	doc2vec	47	57	0.247
8	Class	98	792	0.168

(b) Separate F10-max. thresh.

Rank	Repr.	TP	FP	F10
1	GTE	260	1,236	0.899
2	PaECTER	265	1,862	0.897
3	OpenAI	255	1,118	0.886
4	S-BERT	250	3,001	0.816
5	TF-IDF	242	3,997	0.765
6	USE	235	4,984	0.721
7	Class	209	6,255	0.618
8	doc2vec	173	13,516	0.422

These tables show rankings of model performance by F1/F10 scores and underlying true positives (TP) and false positives (FP). The total number of patent applications is 440; the total number of patent application pairs is 96,580; the total number of true interfering pairs is 322.

Table S4.2: Rankings: Non-threshold-based Metrics

(a) ROC AUC

Rank	Repr.	ROC AUC
1	PaECTER	0.991
2	GTE	0.991
3	OpenAI	0.988
4	S-BERT	0.984
5	TF-IDF	0.976
6	USE	0.964
7	Class	0.855
8	doc2vec	0.839

(b) PR AUC

Rank	Repr.	PR AUC
1	PaECTER	0.654
2	GTE	0.640
3	OpenAI	0.617
4	S-BERT	0.517
5	TF-IDF	0.448
6	USE	0.356
7	Class	0.211
8	doc2vec	0.172

These tables show rankings of model performance by ROC and PR AUC scores in the interference task.

interference task, where PaECTER led. PaECTER's weaker historical performance likely reflects its fine-tuning on patent data from 1985–2022.

Note on LLM-Based Validation We explored using Large Language Models (Claude 3.5 Sonnet and GPT-4) as a scalable alternative to human annotation. However, LLMs showed notable disagreement with human annotators and with each other: Claude selected GTE as best-performing, while GPT-4 chose S-BERT. See Appendix S6 for detailed results.

Table S4.3: Human Agreement with Similarity Rankings by Representation

	Dep. Var.: More similar pair = 1			
	PaECTER	GTE	BERT	TF-IDF
(Intercept)	0.28***	0.20**	0.24***	0.37***
	(0.07)	(0.06)	(0.07)	(0.07)
Human Choice = 1	0.51***	0.62***	0.54***	0.35***
	(0.09)	(0.08)	(0.09)	(0.10)
R^2	0.27	0.38	0.29	0.12
Adj. R^2	0.26	0.38	0.28	0.11
Num. obs.	83	90	91	89

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

This table shows regression results evaluating the agreement between human annotators and relative similarity rankings of patent pairs according to different representations. GTE far outperforms the other models.

S4.3 Classification Task Details

We use CPC assignments from the May 2023 vintage. We evaluate similarity at two levels: eight top-level technology sections and 123 three-character technology classes. For each classification level and quarter-century period from 1850 to 2023, we randomly sample 200 patents from each category. An important limitation is that patent classifications emphasize administrative utility rather than technological similarity per se.

S-BERT demonstrates notably strong classification performance, leading on top-level sections by both ROC AUC and PR AUC. Both S-BERT and PaECTER outperform GTE at the finer three-character class level. However, the class agreement task rewards surface-level classification consistency rather than semantic similarity per se — two patents in the same class need not be conceptually similar, and two similar patents may span class boundaries. This task-specific variation reinforces our multi-task validation strategy: different similarity concepts demand different validation approaches.

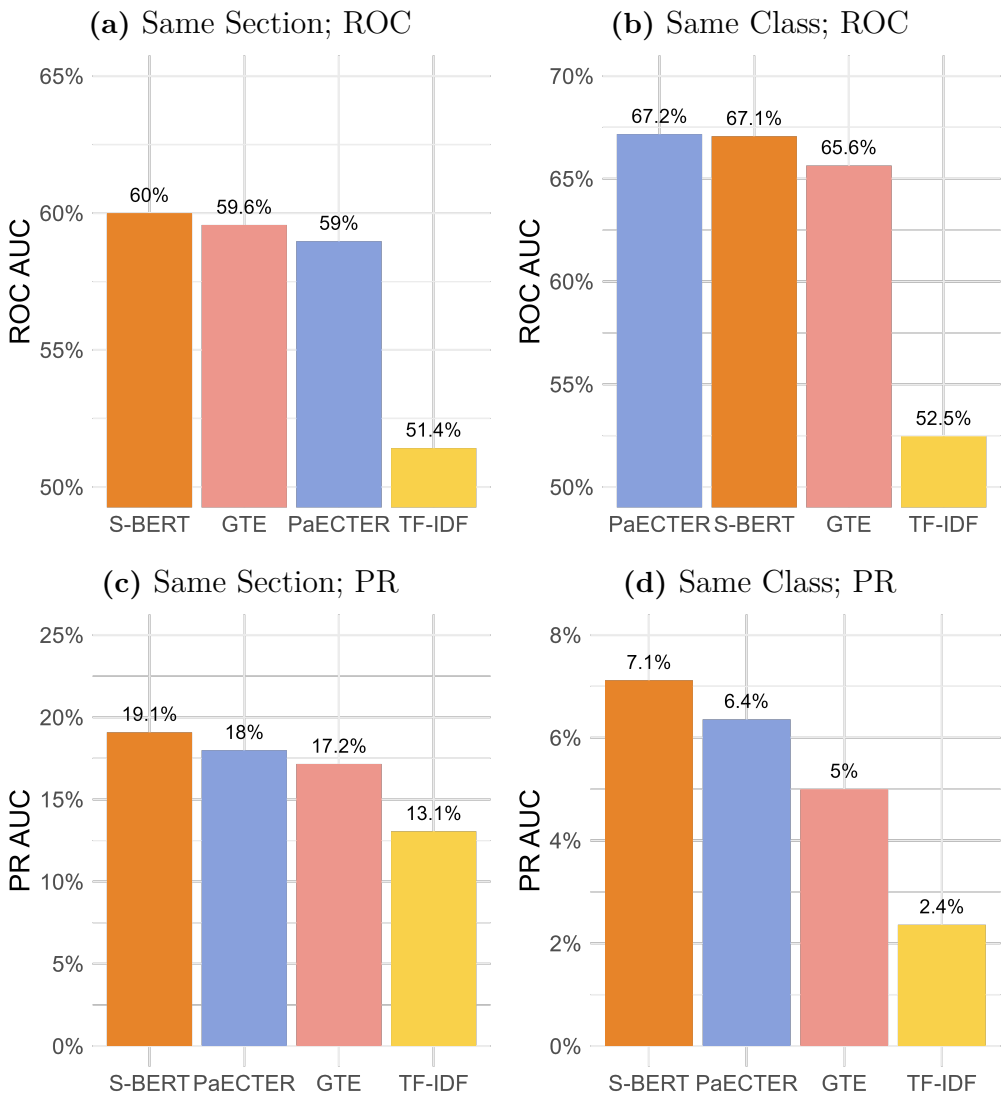


Figure S4.2: Representation Performance on Common Section and Class Tasks

These plots show performance for different representations in the common top-level technology section and 3-digit technology class tasks.

Appendix S5 Instructions for the Non-Expert Human Judgement Task

You will be comparing the similarity of two pairs of patents to determine which pair is more similar to each other. Read through each pair carefully. Then compare the key aspects of each pair of patents, including the following (feel free to use "scratchpad" column to take notes, but that's not necessary):

- The general field or domain the patents relate to
- The specific problem each patent is trying to solve
- The key components of the solution each patent proposes
- Any other major similarities or differences between the patents in each pair

Based on analyzing these factors, assess the overall similarity of the patents in each pair. Determine which pair of patents you think is more similar to each other.

If you don't understand the text enough to assess the above, feel free to google to understand meaning of unfamiliar words or concepts. But try to avoid reading parts of the patent that are outside the snippet (for example, using google patents).

In the "anno_more_similar_1_or_2_or_0" column, put only the number of the pair (1 or 2) that you judge to be more similar. If you are unsure about which is better, put 0 there.

To make it easier to annotate in Excel, adjust the width of the text_pair_1 and text_pair_2 columns and click the "wrap text" button.

Example

Pair 1

IMPROVEMENT: Improvements in Train-Binding Harvesters and Mowers

CLAIMS: The combination of the wedge-shaped platform 15, secondary platform 47, door 35, carriage 46, pivoted reciprocating extension-rake 41, chain 64, and the pulleys 60, these members constructed and operating substantially as and for the purposes herein specified. 2. In combination with the main frame B, the detachable arm 63, having the binder mounted thereon, substantially as and for the purposes herein specified. 3. The combination of the arm 63, eyebolt H

IMPROVEMENT: Improvement in Incandescent Electric Lamps

CLAIMS: 1. The combination, with the incandescing conductor of an electric lamp and the key for controlling the circuit thereof, of an adjustable resistance located within the base of the lamp and cut in or out of the circuit in any desired proportion by the key, so that the lamp may be used at any desired power less than its normal capacity, substantially as set forth. 2. A carbon resistance made substantially as described, and provided with a series of metallic contacts, in combination with a key havin

Pair 2

IMPROVEMENT: Improvements in Wire Fences

CLAIMS: 1. In a wire fence a vertical brace or tie having two legs, a horizontal wire having horizontal bends disposed between said two legs, a plate having at each end a pair of horizontally-extending prongs or fingers with spaces between the same, and a connecting-portion d, the back side of said connecting portion being disposed within said horizontal bend, the horizontal wire passing throughsaid spaces, and the front side of said prongs or fingers being clamped around said legs, substantially as and

IMPROVEMENT: Improvements in Hitches

CLAIMS: 1. A trailer hitch comprising a bar, means for rigidly securing said bar vertically on a vehicle bumper, a loop loosely mounted on the lower portion of the bar, said bar having an opening in its upper portion, a bracket removably mounted on the bar, said bracket including a second vertical bar engaged at its lower end in the loop, a forwardly projecting rigid pin on the upper end portion of the second-named bar engaged in the opening of the first-named bar, and a ball rigidly mounted on the seco

Possible Reasoning

Pair 1 The first patent relates to harvesting/mowing equipment, while the second is about incandescent electric lamps. Very different domains. The first patent aims to improve the binding mechanism on a harvester/mower. The second allows adjusting the power level of an electric lamp. The first uses components like platforms, doors, carriages, rakes and pulleys in its solution. The second uses an adjustable resistance, metallic contacts, and a key. The two patents are solving very different problems in unrelated fields using dissimilar components and mechanisms.

Pair 2 Both patents relate to connection/attachment mechanisms, the first for wire fences and the second for trailer hitches. More related domains than Pair 1. The first patent aims to provide an improved way to brace and tie together wires in a fence. The second provides an

improved trailer hitch mechanism. Both make use of bars, loops, brackets, and engagement of components to create their attachment solutions. While the specific applications differ, both patents essentially aim to solve connection/attachment problems using some similar components like bars, loops and brackets.

Conclusion The patents in Pair 2 seem to have more in common in terms of their general domain, the type of problem they are solving, and some of the key components used, compared to the very different patents in Pair 1. Pair 2 appears more similar overall.

More difficult pairs

Many patent pairs will be more tenuously connected than others; even when patent pairs seem dissimilar, try to think about how they might be trying to solve similar problems or using similar technology.

Here are some examples of dissimilar things that might still be the more similar patent pair in a row:

- Sewing Machines and Closet Hanging Rods are very different technologies, but are both related to clothing/home goods
- Flutes and Tube Sprinklers are very different technologies, but are both tubes with holes in them

Often the patents themselves are small but complicated improvements in technologies you are already familiar with. Even if it is hard to understand the improvement, try to think about how you can connect the technologies in each pair of patents (even tenuously), keeping in mind again:

- The general field or domain the patents relate to
- The specific problem each patent is trying to solve
- The key components of the solution each patent proposes
- Any other major similarities or differences between the patents in each pair

Appendix S6 LLMs for Patent Similarity Assessment

Human annotation, while valuable, can be costly and challenging, especially when comparing technical documents like patents. To address these limitations and provide a scalable approach to our validation setup, we explore the use of Large Language Models (LLMs) for annotation tasks. While this approach introduces its own set of limitations, it offers potential benefits in terms of scalability and cost-effectiveness.

We do not view this as an exercise in using LLMs as survey respondents. Recent research across various disciplines has shown that LLMs often do not reflect human judgments in statistically accurate ways (Bisbee et al. 2024; Dominguez-Olmedo et al. 2024; Goli and Singh 2024). In light of these findings, we cannot assume that LLMs have the same underlying concept of idea similarity as humans. Rather, we explore whether this is the case to a useful degree by comparing LLM results with human annotations, allowing us to assess the potential utility of LLMs in this context.

Our approach is conceptually similar to the distillation techniques used in LLM research, where outputs from larger models are used to improve or evaluate smaller models (Hsieh et al. 2023). In our case, we are not improving capabilities but testing them, using larger LLMs to evaluate the performance of smaller embedding models that share many elements with LLMs.

We employed two state-of-the-art (as of July 2024) language models, Claude 3.5 Sonnet (`claude-3-5-sonnet-20240620`) and GPT-4o (`gpt-4o-2024-05-13`), to perform the same similarity judgment task as human annotators. We provided the models with identical patent pair comparisons, using carefully designed prompts based on the human annotator instructions (see S6.1 for the full prompt).

Our prompts were structured to mirror the human annotation process closely, incorporating a "chain of thought" (CoT) approach (Wei et al. 2024). The LLMs were instructed to analyze key aspects of each patent pair in a "scratchpad" section before making a final judgment, mirroring the format of human annotations.

S6.1 LLM Prompt for Patent Similarity Assessment

You will be comparing the similarity of two pairs of patents to determine which pair is more similar to each other.

Here is the first pair of patents:

<pair1> {PAIR1} </pair1>

And here is the second pair of patents:

<pair2> {PAIR2} </pair2>

Read through each pair carefully. Then, in a <scratchpad>, compare the key aspects of each pair of patents, including:

- The general field or domain the patents relate to
- The specific problem each patent is trying to solve
- The key components of the solution each patent proposes
- Any other major similarities or differences between the patents in each pair

Based on analyzing these factors, assess the overall similarity of the patents in each pair. Determine which pair of patents you think is more similar to each other.

In an <answer> tag, output only the number of the pair (1 or 2) that you judge to be more similar. If you are unsure about which is better, output 0. Do not include any other text or explanation. Close the answer tag with </answer>. You shouldn't have a bias towards answering either 1 or 2; the answer should be only evidence-based. If you don't have a reasonable level of confidence, it's better to output a 0.

S6.2 LLM Results

To analyze the agreement between LLM judgments and embedding-based similarity rankings, we use the following regression setup:

$$I[Sim(2) > Sim(1)]^{Emb} = \beta_0^{LLM} + \beta_1^{LLM} I[Response = 2]^{LLM} + \epsilon \tag{S6.17}$$

where LLM $\in \{\text{Claude}, \text{GPT}\}$ and Emb $\in \{\text{PaECTER}, \text{GTE}, \text{S-BERT}, \text{TF-IDF}\}$. The coefficient β_1 represents the increase in the probability that the embedding indicates pair 2 is more similar when the LLM chooses pair 2. Higher β_1 suggests a stronger LLM-embedding agreement.

Each LLM produced outputs for 100 comparisons. However, the number of observations in our regressions is lower, reflecting the removal of cases where the LLM responded with 0 (indicating it couldn't decide). This ensures that our analysis focuses on clear judgments made by the LLMs.

We present the results of our LLM-based regressions in Table S6.1. The ranking of representations differs between the two LLMs and from our human annotation results. For Claude, the ranking is GTE >S-BERT >TF-IDF >PaECTER, while for GPT-4o, it's S-BERT >GTE >PaECTER >TF-IDF. Despite these differences, both LLMs consistently show that newer embedding models (GTE, S-BERT) outperform the traditional TF-IDF approach, aligning with our human annotation findings in this crucial aspect.

Table S6.1: LLM Agreement with Embedding-Based Similarity Rankings

	PaECTER		GTE		S-BERT		TF-IDF	
	Claude	GPT	Claude	GPT	Claude	GPT	Claude	GPT
(Intercept)	0.14	0.17	0.08	0.14	0.16	0.11	0.16	0.31*
	(0.10)	(0.10)	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)	(0.13)
Claude=1	0.52***		0.60***		0.58***		0.54***	
	(0.11)		(0.10)		(0.10)		(0.10)	
GPT4o=1		0.57***		0.58***		0.71***		0.35*
		(0.12)		(0.11)		(0.10)		(0.15)
R^2	0.19	0.26	0.28	0.28	0.28	0.43	0.23	0.08
Num. obs.	92	72	91	76	90	67	94	68

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Regression results showing the agreement between Claude 3.5 Sonnet (`claude-3-5-sonnet-20240620`), GPT-4o (`gpt-4o-2024-05-13`), and the relative similarity rankings of patent pairs according to different patent text representations.

The variability in results between human annotators and different LLMs underscores the potential limitations of using LLMs as proxies for human judgment in this context. However, the consistent underperformance of TF-IDF across all evaluation methods (human and LLM) provides strong evidence for the superiority of newer embedding techniques in capturing patent similarity. This suggests a potential use for LLMs as a cost-effective way to test the validation tasks before deploying them to human annotators, streamlining the overall validation process.

Appendix S7 Why Are Deep Learning Models Better? An In-Depth Look at Why S-BERT Is Better than TF-IDF.

In this section, we explore the performance differences between S-BERT and TF-IDF. First, we compare a 21st-century bicycle patent and a 19th-century velocipede patent to illustrate S-BERT's ability to identify semantic similarities. Second, we examine unigram frequencies in the Google Books Ngram database. Unigrams characteristic of patent pairs with high TF-IDF similarity overweight period-specific language similarities, rather than similarity of ideas represented by the patents. We then present details of the characteristic unigram methodology, an additional Google Books Ngram analysis, and a synonym-based analysis that further highlights S-BERT's ability to capture semantic similarity.

S7.1 Example: Bicycle versus Velocipede

Figure S7.1 shows a bicycle patent from the 21st century and a velocipede patent from the 19th century. Despite these patents originating from different time periods and employing distinct terminologies, S-BERT successfully identifies them as similar, positioning them in the 87th percentile of similarity. At the same time, the similarity according to TF-IDF is 0. This example illustrates the S-BERT's ability to capture semantic nuances and contextual similarities despite changes in language.

Both patents introduce improvements in the design or function of two-wheeled vehicles. A velocipede is an archaic term for a type of bicycle. Although Patent 1 focuses on the "front frame for a bicycle" while Patent 2 is more broadly about an "improved velocipede," they both involve common mechanical features such as tubes, frames, and axles. However, the patents do not share many common terms. Patent 1 talks about "front frame," "inner tubes," "upper tube," while Patent 2 mentions "friction-clutch," "spurs," "arms," etc.

S-BERT takes into account not just specific words, but also the context in which these words appear. Words with similar meaning that frequently appear in similar contexts will be assigned similar S-BERT vectors. Thus, S-BERT representations reflect that both patents are about two-wheeled vehicles, even if they use different terms. S-BERT is trained on a diverse dataset, which includes technical language. It can therefore encode terms like "frame," "tubes," and "axle" as related in general, even if they appear in different contexts.

TF-IDF is a simpler bag-of-words model that does not capture meaning in the same way (see Smith 2020). It considers only the frequency of individual words in each document and in the corpus as a whole. TF-IDF treats distinct terms such as "bicycle" and "velocipede" as

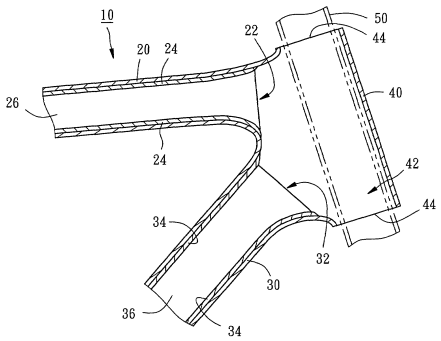
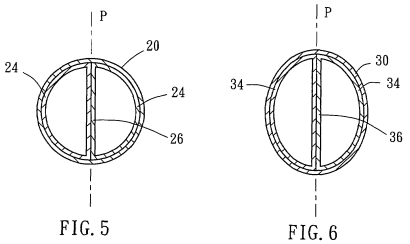
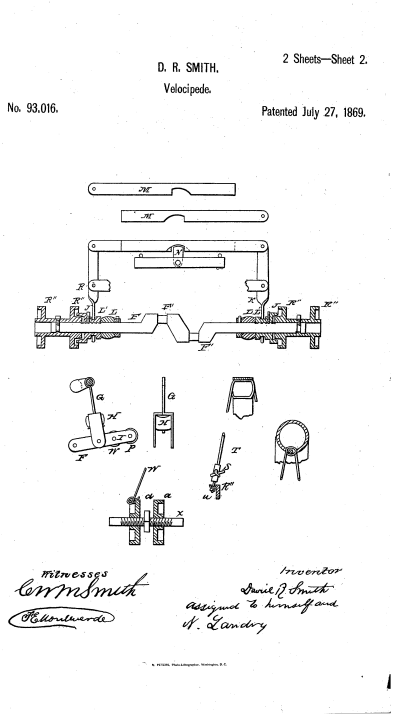
Patent 1: US7562890B2 (2009)	**Patent 2: US93016A (1869)**
Front frame for a bicycle. 1. A front frame for a bicycle, comprising: two first inner tubes abutted together; two second inner tubes abutted together; an upper tube of cured multiple layers of fiber reinforced rein material wound around the two first inner tubes so that there is no crack between the upper tube and ...	IMPROVED VELOCIPEDE. In the velocipede as constructed, and in combination therewith, the friction-clutch, spurs, arms, cross-bar, cam, guide-wheel, with hollow rim and axle, arranged and operated substantially as described. In witness whereof, I have hereunto set my hand and seal.
 FIG. 4  FIG. 5 FIG. 6	

Figure S7.1: A Conceptually Similar Pair of Patents

A velocipede is a type of bicycle. The text is truncated to the title and the beginning of the claims section of the patents. Optical Character Recognition (OCR) errors were fixed for this illustrative example. According to S-BERT, these patents are in the 87th percentile of similarity, whereas according to TF-IDF, the similarity is 0.

unrelated concepts. In sum, S-BERT is able to better capture the semantic and contextual similarities between these two patents that describe similar inventions but do not share a common vocabulary.

S7.2 TF-IDF Overweights Period-Specific Words versus Universal Synonyms

The bicycle/velocipede example suggests that TF-IDF overweights period-specific terms like velocipede, leading it to assign low similarity to pairs that might describe the same idea with different terms. Here we extend that analysis. We hypothesize that terms used in patent pairs assigned high similarity by TF-IDF should have a higher variance of usage over time. These period-specific terms might be archaic or modern, or they may have irregular fluctuations in usage.

Figure S7.2 presents some illustrative examples of unigram frequencies over time. Among the top-five most characteristic unigrams, TF-IDF unigrams are more volatile, which indicates more time-specific word usage.

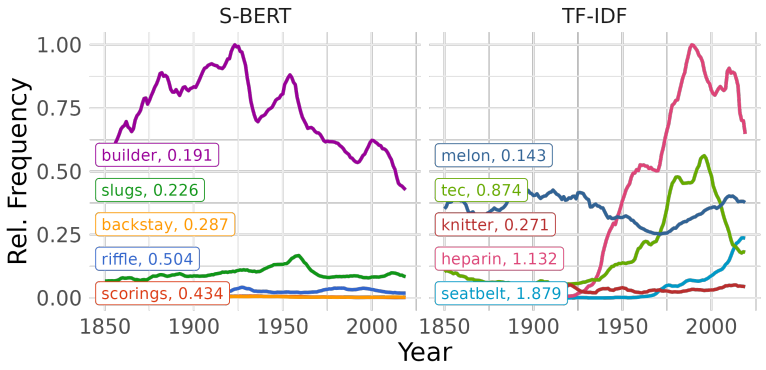
We further hand-picked examples of conceptually-similar words in panel (b). "Dresser," characteristic of S-BERT similar pairs, exhibits moderate use with little variation until the 2000s. In contrast, "vanity," characteristic of TF-IDF similar pairs, exhibits more volatility, steadily dropping in usage throughout the period between 1850 and 1970, followed by a small rise. Another example is shown in panel (c). "Verbal" and "cognitive" both increase after 1950. But the increase is more dramatic for "cognitive," and therefore this term characteristic of TF-IDF similar pairs has a larger coefficient of variation.

S7.3 Google Ngrams Analysis

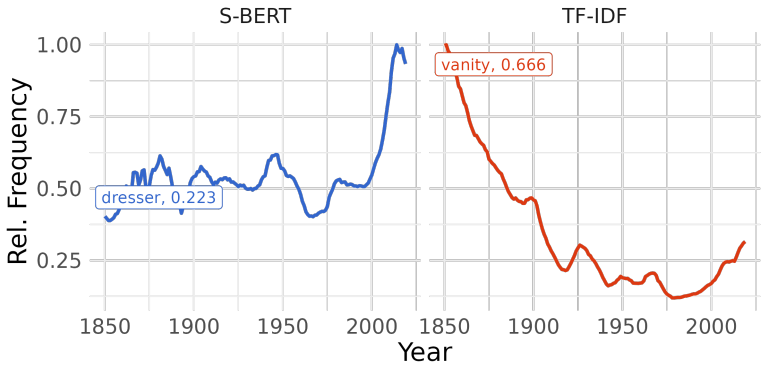
To gain insights into the time-specific nature of the words that TF-IDF focuses on, we turn to examining the tokens characteristic of patent pairs located closely in the TF-IDF space through the lens of Google Ngrams data. We identify characteristic tokens that differentiate patent pairs based on their similarity scores. Our analysis categorizes patent pairs into three groups: (i) those identified as similar by both S-BERT and TF-IDF, (ii) those recognized as similar only by S-BERT, and (iii) those recognized as similar only by TF-IDF. We exclude pairs with mutual agreement between models and determine characteristic unigrams for the latter two categories.

This analysis demonstrates that the unigrams characteristic of patent pairs with high TF-IDF similarity tend be more heavily used in specific time periods compared to the S-BERT unigrams, which can explain the outperformance of TF-IDF in the period classification task.

(a) Top-5 characteristic unigrams for each representation



(b) Hand-picked example 1



(c) Hand-picked example 2

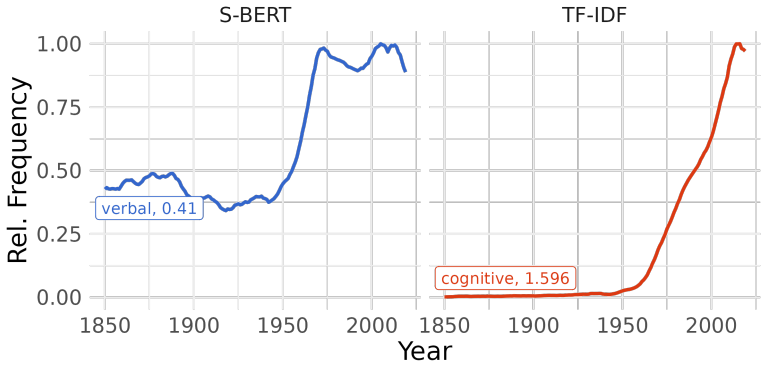


Figure S7.2: Frequency of characteristic unigrams of the pairs of patents classified as similar by S-BERT and TF-IDF

The plot is based on the Google Ngram Corpus (1850–2019). Frequency is normalized to the largest frequency on each plot. The number after the unigram label is the coefficient of variation, defined as the standard deviation divided by the mean. The characteristic unigrams are computed using the Monroe et al. 2017 algorithm.

The Google Books Ngrams dataset is a collection of word frequencies derived from the Google Books corpus,[34] which contains a vast array of books published over several centuries. This dataset enables the analysis of the usage patterns of words and phrases over time, providing a valuable resource for studying the evolution of language.

In NLP, characteristic tokens or words are specific lexical features that are highly indicative of a particular category, topic, or sentiment. These tokens serve as markers that can help in classifying or differentiating texts based on the target concept of interest, such as the party alignment of a political speech, or, in our case, whether a patent pair is deemed similar by S-BERT or TF-IDF. We use the Monroe et al. (2017) method implemented in the Schnoebelen et al. (2022) R library to systematically identify characteristic words. The method employs Bayesian shrinkage and regularization techniques to select and evaluate the relative importance of words that capture the target semantic concept.

Finding characteristic words requires a corpus of text split according to a categorical variable, which we obtain the following way. From the corpus of 11,200 patents used in the class and period validation task, we selected pairs that were in the top quartile of similarity scores according to S-BERT, TF-IDF, or both. We then categorized these pairs into three classes:

1. The representations agree
2. S-BERT identifies as similar, but TF-IDF does not *S-BERT Yes* category
3. TF-IDF identifies as similar, but S-BERT does not *TF-IDF Yes* category

We discard the pairs where both representations agreed and use the rest of the pairs as the input to Monroe et al. (2017) algorithm to find unigrams most characteristic of S-BERT and TF-IDF similarity. The output of the algorithm is the list of characteristic words for the categories *S-BERT Yes* and *TF-IDF Yes* along with the weighted log-odds that quantify the extent to which a unigram is more likely to appear in one category of patent pairs compared to the other.

Once the characteristic unigrams are obtained, we analyze their frequency from 1850 to the present using the Google Books Ngram corpus. For each unigram, we calculate the mean and standard deviation of its frequency over time. To obtain a measure of variation that is comparable between different unigrams we compute the coefficient of variation, defined as the standard deviation divided by the mean.

[34]Specifically, we use the "English 2019" corpus accessed using *ngramr* library in R programming language (Carmody 2023).

Figure S7.3 demonstrates the average coefficient of variation for *S-BERT Yes* and *TF-IDF Yes* characteristic unigrams. The difference is large, especially for the unigrams with the highest weighted log-odds. For the top 100 unigrams, the S-BERT coefficient of variation is 0.7 compared to 1.2 for TF-IDF (which means that the average standard deviation is 70% and 120% of the mean, respectively). As we increase the number of unigrams we include in the computation, the difference becomes smaller, but is always large: for all unigrams, the S-BERT coefficient of variation is 0.74 compared to 0.95 for TF-IDF.

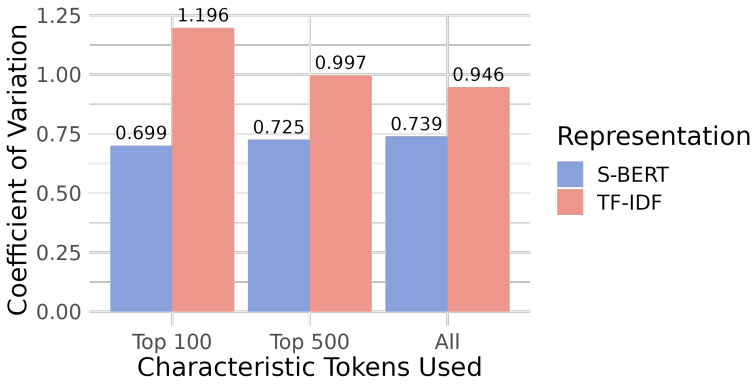


Figure S7.3: Average over-time coefficient of variation of the frequency of characteristic unigrams of the pairs of patents classified as similar by S-BERT and TF-IDF

The unigram frequency information is from the Google Ngram Corpus (1850–2019). The coefficient of variation is defined as the standard deviation divided by the mean. The characteristic unigrams are computed using the Monroe et al. 2017 algorithm.

The higher coefficient of variation of unigrams in the *TF-IDF Yes* category suggests that TF-IDF is sensitive to the linguistic peculiarities of specific time periods. This provides strong evidence for why TF-IDF is more effective at categorizing patents based on their temporal context.

S7.4 Synonyms Analysis

The objective of this analysis is to further explore the contrasting types of similarity captured by S-BERT and TF-IDF, particularly focusing on why S-BERT excels in class validation while TF-IDF shines in the period task. Our hypothesis posits that S-BERT, unlike TF-IDF, assigns a relatively lower weight to exactly overlapping words when determining similarity between patent pairs, and leans more towards semantic similarity and other forms of word "interchangeability." This distinction becomes apparent when analyzing patents within the same period that tend to exhibit period-specific overlapping language, even if they belong to different classes. Conversely, patents from the same class but different periods are more

likely to exhibit similarity at a conceptual or idea level, which is the main type of similarity we aim to capture.

In preparing the data for analysis, we further stratified patent pairs from the Class/Period validation sample into three strata: `tfidf_yes`, `S-BERT_yes`, and `agree` (using the 75th percentile similarity cutoff for yes). For instance, `S-BERT_yes` implies that according to S-BERT this pair is similar, but according to TF-IDF, it is not. We further categorized them as `same_class`, `same_period`, `both_same`, and `neither_same`. To focus on informative cases, pairs in `agree`, `both_same`, and `neither_same` categories were excluded. A sample of 200 pairs from each of the 4 strata (800 pairs in total) was selected.

To enrich our analysis, we employed WordNet, a lexical database of English (Miller 1992). In WordNet, nouns, verbs, adjectives, and adverbs are grouped into sets of synonyms (synsets), each expressing a distinct word sense. These synsets are interlinked by means of semantic relations. The relations include hypernyms (more abstract terms), hyponyms (more specific terms). For each word in each patent, we listed all word senses. For each word sense, we found the set of synonyms, hypernyms, and hyponyms. These, along with the original word, were concatenated. For instance, for the word "air," we obtained a set of related terms encompassing synonyms like "breeze," hypernyms like "gas," and hyponyms like "zephyr."

Each patent was then represented as the set of unique tokens in it (each counted once) and separately as the set of unique tokens plus their synonyms, hypernyms, and hyponyms. For each document pair, we calculated the exact word overlap and the word plus synonym plus hypernym plus hyponym overlap (Word+ overlap).

We then conducted a pair of analyzes with the aim of investigating whether the same text characteristics drive both S-BERT similarity and belonging to the `same_class` category, as well as TF-IDF similarity and belonging to the `same_period` category. In the first analysis of the pair, we ran regressions with S-BERT and TF-IDF on the LHS and the text characteristics (exact word overlap and Word+ overlap) on the RHS. This analysis aimed to explore the relationship between the similarity scores generated by S-BERT and TF-IDF and the text characteristics.

In the second analysis of the pair, we conducted a PR AUC analysis with `same_class` and `same_period` categories as the dependent variables and the text characteristics as predictors. This analysis aimed to explore how well the text characteristics predict the categorization of patents into `same_class` and `same_period` categories.

The findings from both analyses exhibited similar patterns: S-BERT similarity and `same_class` categorization were both driven by Word+ overlap, while TF-IDF similarity and `same_period` categorization were both driven by direct word overlap. These patterns

Table S7.1: Regression results for similarity scores and Wordnet-based measures on the `S-BERT_yes` and `tfidf_yes` patent sample

	TF-IDF	S-BERT
(Intercept)	0.31***	0.58***
	(0.02)	(0.02)
Word Overlap	0.39***	−0.29***
	(0.04)	(0.04)
Word+ Overlap	−0.01	0.13**
	(0.04)	(0.04)
R^2	0.15	0.06
Adj. R^2	0.15	0.06
Num. obs.	800	800

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

This table shows estimates from a regression where the dependent variables are the similarity scores generated by TF-IDF and S-BERT. The explanatory variables are Word Overlap, representing the exact word overlap between patent pairs, and Word+ Overlap, representing the overlap including synonyms, hypernyms, and hyponyms. The negative coefficients for S-BERT on Word Overlap and for TF-IDF on Word+ Overlap are observed due to the sampling strategy focusing on patents where the two models disagree.

led us to conclude that S-BERT's superior performance in `same_class` categorization can be attributed to its ability to capture the semantic similarity of words present in the patents, whereas TF-IDF's superior performance in `same_period` categorization can be attributed to its ability to capture direct word overlap.

The findings are shown in Table S7.1 and Figure S7.4, exhibiting expected patterns. Table S7.1 quantitatively shows how WordNet-derived measures relate to S-BERT and TF-IDF similarity scores. The regression coefficients indicate that S-BERT's similarity scores are negatively associated with direct word overlap but positively associated with Word+ overlap, suggesting a stronger emphasis on semantic similarity. The negative coefficient on direct word overlap reflects the axis of disagreement between models: conditioning on pairs where S-BERT and TF-IDF disagree, high word overlap mechanically predicts that TF-IDF scored the pair highly while S-BERT did not. Conversely, TF-IDF's similarity scores are positively associated with direct word overlap, indicating a preference for exact lexical matching.

Following the tabular analysis, Figure S7.4 visually represents the Precision-Recall Area Under Curve (PR AUC) values for Word and Word+ overlap measures across `same_class` and `same_period` categorizations. In the `same_class` categorization, it is discernible from the

figure that Word+ overlap (`sim_combined`) yields a higher PR AUC value of 0.49 compared to the Word overlap (`sim_1_2`) value of 0.43, underscoring the importance of capturing semantic relationships in addition to exact word overlap for classifying patents within the same class. Conversely, in the `same_period` categorization, Word overlap outperforms Word+ overlap with a PR AUC value of 0.588 against 0.512, indicating that direct word overlap is more pertinent for capturing period-specific similarities. The Figure also shows that S-BERT performs best on `same_class` task and TF-IDF performs best on the `same_period` task on the sub-sample used in this analysis, conforming with the full sample results discussed in Section S4.3.

In conclusion, one of the mechanisms through which S-BERT better captures idea similarity is through its ability to assign similar vectors to words located closely in the semantic graph (synonyms, hypernyms, hyponyms). This is consistent with the properties theoretically expected from S-BERT based on its architecture and training procedure. Our results show that these properties are useful in innovation economics by allowing S-BERT to capture the similarity of ideas in a way that transcends period-specific language.

S7.5 Why Is S-BERT Better? Conclusion

The Google Ngrams analysis and the patent pair example collectively offer robust evidence to support our initial observations. TF-IDF's strength lies in identifying patents from the same time period, primarily due to its sensitivity to words that are popular within specific temporal contexts. Conversely, S-BERT proves superior at classifying patents into the same technical class, given its ability to understand and capture the semantic essence of the text, highlighted by its association with synonym, hypernym, and hyponym overlap as opposed to the exact word overlap. These insights are important for choosing the more appropriate model for specific downstream tasks.

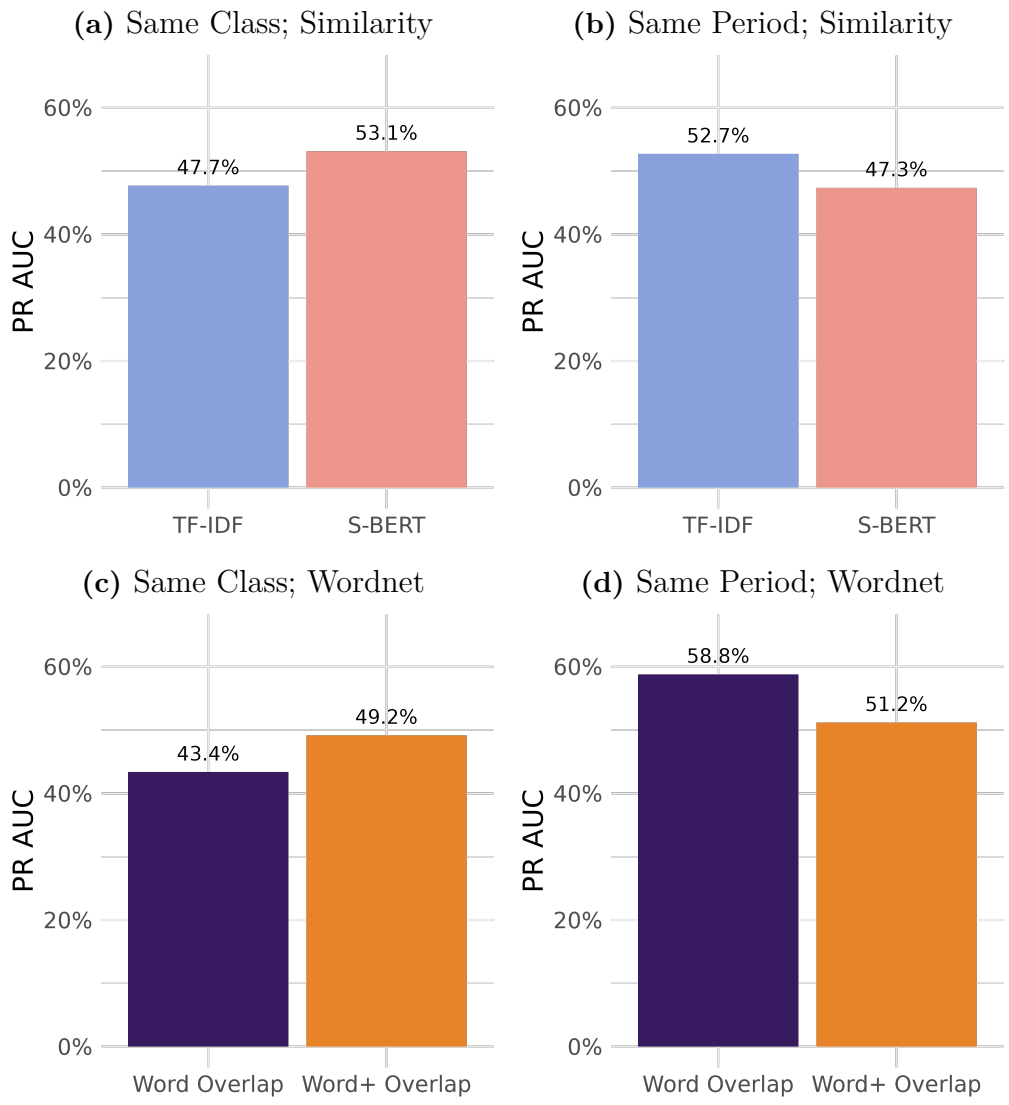


Figure S7.4: Similarity scores based on the S-BERT and TF-IDF representations and Wordnet-based measures for categorizing patent pairs as belonging to the same class and period

The sample includes patent pairs in the `S-BERT_yes` and `tfidf_yes` categories. We evaluate how well patent pairs can be classified as belonging to the same class or the same quarter-century period using two sets of similarity scores, based on S-BERT and TF-IDF representations, and two sets of Wordnet-based measures, Word Overlap and Word+ Overlap. "Word" represents exact word overlap and "Word+" encompasses word overlap along with their synonyms, hypernyms, and hyponyms as derived from Wordbet, a lexical database grouping English words into sets of synonyms and recording their semantic relationships.

Appendix S8 Visualizing Representation Differences

This section visualizes how different representations create different similarity spaces using two-dimensional projections of high-dimensional embeddings.

S8.1 Methodology

The raw data are obtained using the same sampling strategy outlined in the classification validation task (S4.3). We sampled patents from top-level technology sections and 25-year periods, 1850–2023.

We then plot 2-dimensional projections of the embedding spaces, where individual patents are marked with color according to their respective class or period. This visualization technique provides a geometrically intuitive perspective of the innovation space. It also lays a visual foundation for comparing the efficacy of different embedding techniques like S-BERT and TF-IDF.

The primary method we employ for visualization is dimensionality reduction through Uniform Manifold Approximation and Projection (UMAP) (McInnes et al. 2018). UMAP is noted for its ability to preserve both global and local structures during reduction, making it, roughly speaking, a non-linear variant of Principal Component Analysis (PCA).

To speed up the computation, we conduct the initial dimension reduction using PCA, which reduces the dimensionality of the S-BERT and TF-IDF representations to 50. Subsequently, UMAP is applied to these reduced representations. This two-step process harnesses the computational efficiency of PCA while benefiting from the geometric qualities of UMAP.

We manually tuned UMAP hyperparameters to achieve a more clustered representation that looked more like an "archipelago" than a singular "continent." This tuning aids in better visual separation among clusters within the innovation space.

S8.2 Plotting

One of the challenges we encountered during visualization was the overlapping of data points, especially in dense clusters. To mitigate this, we used a jittering technique which disperses each point slightly within its local neighborhood to reduce overlap, hence enhancing the visibility of individual clusters. Winsorization of extreme values produces the boxy appearance of the scatter plots.

The plots (refer to Figure S8.5) primarily serve as illustrative tools, providing a more tangible notion of the idea space. We use color coding to denote different top-level technology sections. Despite the inherent distortions, some observations could hint at underlying

structural differences between the representations.

S-BERT representations show clearer class boundaries compared to TF-IDF representations, suggesting that patent clustering is closer to the class structure. These visual patterns are consistent with the results in Section S4.3.

It is harder to draw conclusions from the general layout because of the distortions inherent in the projection. However, some observations stand out. For example, TF-IDF has more "dust" compared to S-BERT, which has more empty space. Also, the extended tails of the TF-IDF representations, hidden due to winsorizing, hint at increased variability due to the expression of similar ideas with different words, which may push these representations farther from the core.

S8.3 Results and Interpretation

Figure S8.5 illustrates how different representations create different similarity spaces using two-dimensional projections of high-dimensional embeddings. Patents are colored by top-level technology classifications to show clustering patterns. S-BERT shows tighter groupings by technology class compared to TF-IDF, suggesting it better captures technological relationships. S-BERT also reveals nuanced positioning — for example, a cluster of dark blue semiconductor patents near (-5, 0) is positioned between materials science and electrical engineering clusters, accurately reflecting their hybrid nature.

These visualizations demonstrate that representation choice fundamentally affects the similarity space rather than just adding noise to a consistent underlying structure. Different methods produce qualitatively different maps of technological relationships, making validation essential for selecting appropriate representations for economic analysis.

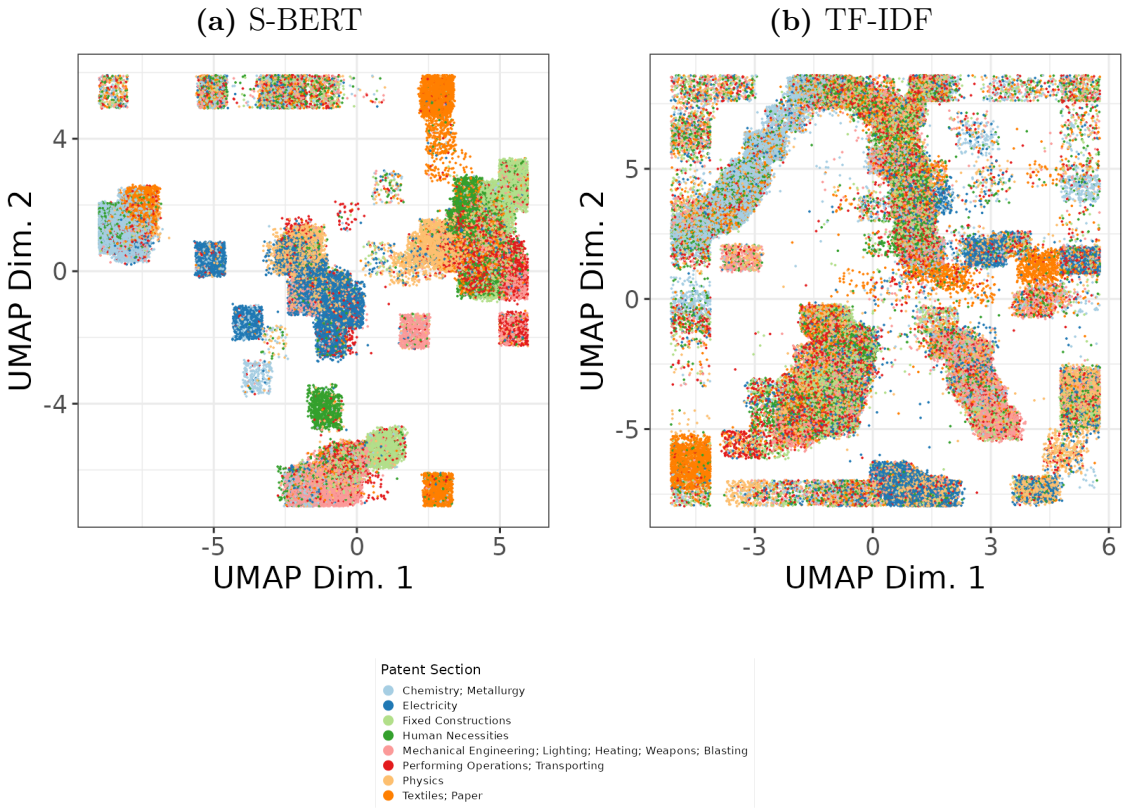


Figure S8.5: Visualizations of S-BERT and TF-IDF Representations

These plots show Uniform Manifold Approximation and Projections (UMAP) for S-BERT and TF-IDF representations using a sample of 111,251 patents stratified by top-level CPC Section and 25-year period. To constrain extreme values, the data were winsorized at the 5% and 95% levels along both axes.

Appendix S9 Similarity Methods and Results

S9.1 Computing Similarity

For the baseline similarity results, we use a simplification for computing pairwise similarity that reduces the complexity from $O(N^2)$ to $O(N)$ for unit-normalized vectors.

For unit-normalized vectors, the average pairwise cosine similarity can be computed as:

$$\text{Average Pairwise Cosine Similarity} = \frac{\left\|\sum_{i=1}^{N} \mathbf{v}_i\right\|^2 - N}{N(N-1)} \tag{S9.18}$$

Or equivalently:

$$\text{Average Pairwise Cosine Similarity} = \frac{\|\text{sum}(\mathbf{V})\|^2 - N}{N(N-1)} \tag{S9.19}$$

Where:

- $\mathbf{V}$ is a matrix of N unit-normalized vectors
- $\text{sum}(\mathbf{V})$ is the sum of all vectors
- $\|\cdot\|$ denotes the L_2 norm

Starting with the average pairwise dot product formula:

$$\text{Avg} = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (\mathbf{v}_i \cdot \mathbf{v}_j) \tag{S9.20}$$

Step 1: Consider the squared norm of the sum of all vectors:

$$\left\|\sum_{i=1}^{N} \mathbf{v}_i\right\|^2 = \left(\sum_{i=1}^{N} \mathbf{v}_i\right) \cdot \left(\sum_{j=1}^{N} \mathbf{v}_j\right) \tag{S9.21}$$

Step 2: Expand the dot product:

$$\left\|\sum_{i=1}^{N} \mathbf{v}_i\right\|^2 = \sum_{i=1}^{N} \sum_{j=1}^{N} (\mathbf{v}_i \cdot \mathbf{v}_j) \tag{S9.22}$$

Step 3: Separate diagonal and off-diagonal terms:

$$\left\|\sum_{i=1}^{N} \mathbf{v}_i\right\|^2 = \sum_{i=1}^{N} (\mathbf{v}_i \cdot \mathbf{v}_i) + \sum_{i \neq j} (\mathbf{v}_i \cdot \mathbf{v}_j) \tag{S9.23}$$

Step 4: Since vectors are unit-normalized, $\mathbf{v}_i \cdot \mathbf{v}_i = 1$:

$$\left\| \sum_{i=1}^{N} \mathbf{v}_i \right\|^2 = N + \sum_{i \neq j} (\mathbf{v}_i \cdot \mathbf{v}_j) \tag{S9.24}$$

Step 5: The sum over $i \neq j$ counts each unique pair twice:

$$\sum_{i \neq j} (\mathbf{v}_i \cdot \mathbf{v}_j) = 2 \times \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (\mathbf{v}_i \cdot \mathbf{v}_j) \tag{S9.25}$$

Step 6: Solve for the sum of unique pairs:

$$\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (\mathbf{v}_i \cdot \mathbf{v}_j) = \frac{\left\| \sum_{i=1}^{N} \mathbf{v}_i \right\|^2 - N}{2} \tag{S9.26}$$

Step 7: Apply the averaging factor:

$$\text{Avg} = \frac{2}{N(N-1)} \times \frac{\left\| \sum_{i=1}^{N} \mathbf{v}_i \right\|^2 - N}{2} \tag{S9.27}$$

$$= \frac{\left\| \sum_{i=1}^{N} \mathbf{v}_i \right\|^2 - N}{N(N-1)} \tag{S9.28}$$

S9.2 Multi-Patent Entitites

To address the rise in multi-patent entities in Section 4.2, we link patents to databases (Kogan et al. 2017; Monath et al. 2021) that disambiguate and assign unique identifiers to inventor and assignee names. First, we link patents that have assignees to assignee identifiers from the PatentsView disambiguation file (Monath et al. 2021). Second, unassigned patents are linked to unique individual inventors from the PatentsView disambiguation file. Some patents have multiple inventors. If there is a set of co-inventors that uniquely identifies a set of patents, then we concatenate these individual inventors into a single entity identifier. The result is a database with every patent linked to an entity identifier that could be a firm, an individual inventor, or a group of individual inventors. For the final step, one random patent was sampled for each entity to compute pairwise similarity for Figure 4. Varying the random seed multiple times yielded nearly identical quantitative results.

S9.3 Sampling for Other Estimates

Other statistics require calculating pairwise cosine distances, which are computationally expensive. Our approach was to sample. If the total number of patents in a year were under 10,000 (until the 1870s), then the matrix was calculated with all patents issued that year. If the number of annual patents was above 10,000, then we sampled 10,000 patents from each year. Then we formed patent pairs to compute a variety of statistics:

- The standard deviation of pairwise similarity, used to standardize similarity changes in Figures 1 and 2.
- Weighted similarity for Figure 5.
- Quantiles of pairwise similarity, described in Online Appendix S9.5.

S9.4 Alternative Normalizations

Figure S9.6 shows similar trends from alternative representations using a different normalization compared with our baseline results. Because different NLP representations have embedding spaces with unknown scaling, we divide each series by its maximum value, in order to better compare percentage changes. Each yields distinct patterns.

GTE exhibits clear secular decline in patent similarity from 1841 through the late 20th century. The trend is consistent and gradual, with minimum similarity reaching approximately 80% of the historical maximum — our main finding of spreading-out.

PaECTER suggests steadily declining patent similarity from 1898 through 1999, followed by partial retracing through 2023. However, PaECTER exhibits minimal overall variability — its minimum value is 98% of its maximum — suggesting either remarkable stability or limited sensitivity to temporal changes.

S-BERT indicates steadily declining patent similarity from the early 20th century through 2023, with greater variability (minimum at 75% of maximum) than GTE or PaECTER. The pattern is qualitatively consistent with spreading-out but noisier.

TF-IDF shows a strikingly different pattern: sharp *increases* in similarity through 1960, followed by high but volatile similarity thereafter. TF-IDF exhibits extreme variability, with minimum similarity at just 20% of its maximum. This pattern directly contradicts our theoretical predictions and suggests inventors are clustering rather than spreading out.

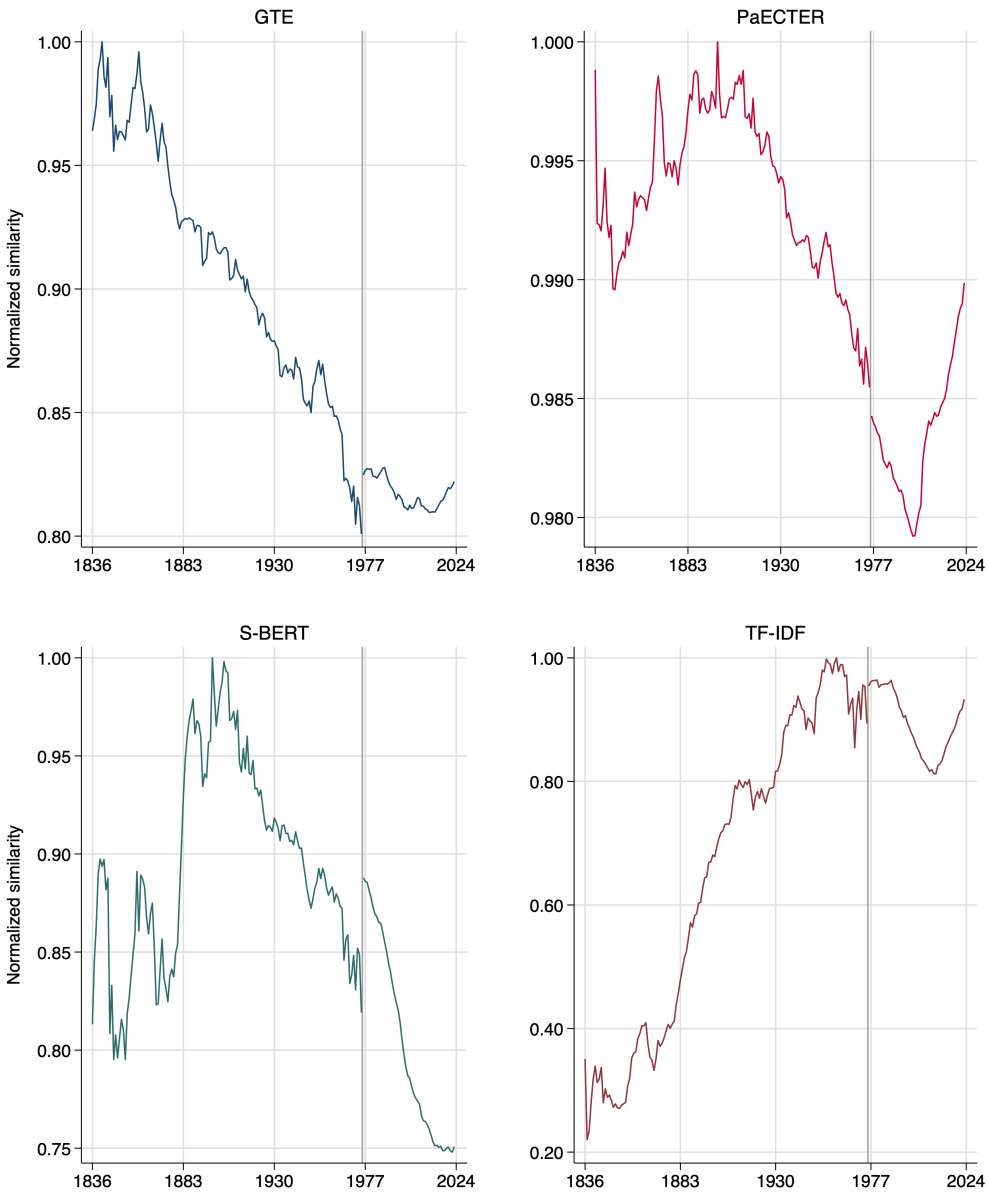


Figure S9.6: Similarity by Year and by Representation

These plots show normalized average pairwise US patent claim similarity by issue year and by representation. Each series is normalized to 1 at its maximum value.

The close correspondence of these results with Figure 2 confirms that the baseline standardization is not consequential for the qualitative dynamics of similarity. Instead, our baseline standardization allows for better comparability across representation by scaling changes in similarity to the size of each embedding space.

S9.5 Quantiles of pairwise distances

Several factors may contribute to changing average pairwise similarity over time beyond the mechanisms described in our model. Improved tools for knowledge dissemination and team management could serve as a countervailing force against dispersion. The emergence of new technological domains or innovation platforms might "pull" inventors towards "low-hanging fruit" or common standards and interfaces, increasing similarity.

Section 4.3 provides some evidence for increased local clustering, especially since 2000, as indicated by the high-γ weighted similarity dynamics. Figure S9.7 provides an alternative window into these dynamics by plotting 50 quantiles of pairwise similarity in each year. The secular decline in similarity across most of our sample period is robust across quantiles of patent similarity. The post-1999 increase in similarity is slightly faster for higher quantiles, providing some initial suggestive evidence of increased local clustering.

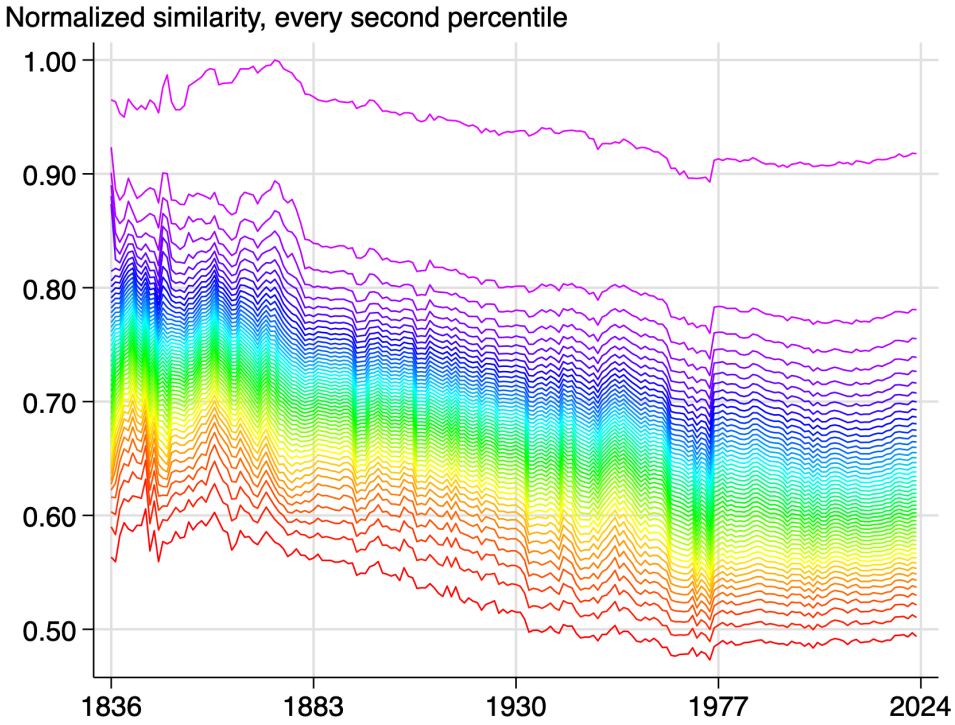


Figure S9.7: Similarity at Different Quantiles

This figure shows normalized GTE similarity trends for 50 quantiles of pairwise similarity. The secular decline in similarity is robust across all quantiles. Each percentile has a different natural scale, so dividing all of them by a common cross-sectional standard deviation distorts comparisons across series. Instead, normalizing by the global max preserves both the shape and the relative levels.

Appendix S10 Register of Interferences

Figure S10.1 shows an example page from one of the Register volumes. It displays two cases. Both cases record hearing dates of January 7, 1890. The subject of the first case was roll paper cutters and the competing inventors were named Ehrlich and Lawton. The case was decided in favor of Lawton on January 11. The subject of the second case, Blaine v. Hadley, was corn harvesters; the case was decided in favor of Hadley on April 29.

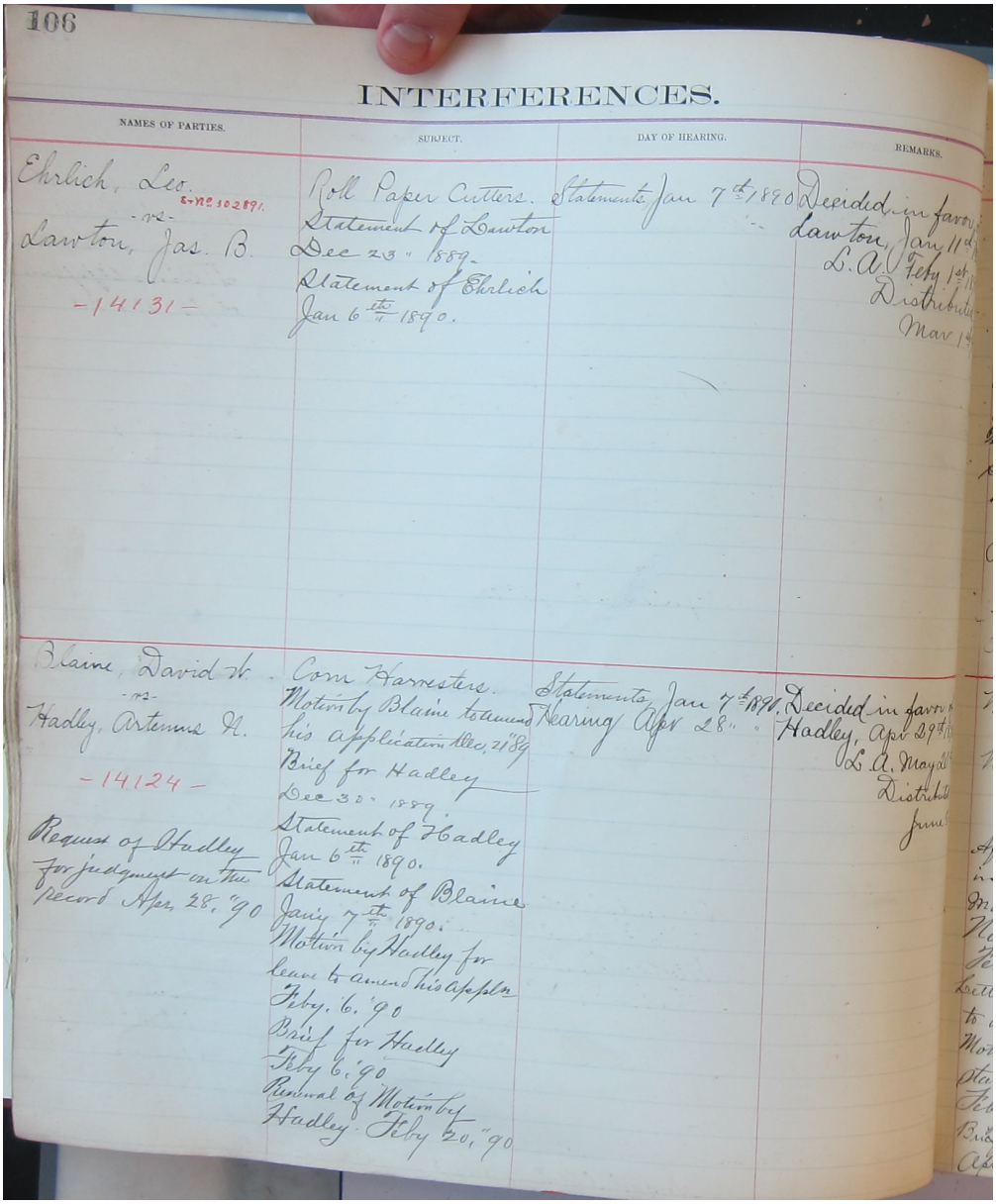


Figure S10.1: Example page from Register of Interferences

Appendix S11 Time-Series Evidence on Spacing and Quality

This appendix complements the cross-sectional analysis in Section 5 with time-series evidence on the spacing-quality relationship. We aggregate patent-level data to the CPC-class-year level and regress annual changes in quality measures on annual changes in similarity measures. CPC class and year fixed effects partial out level differences across technology classes and aggregate time-varying shocks common across classes, so identification comes from within-class deviations from common trends. Standard errors are clustered at the CPC class level.

Table S11.1 reports the results. Point estimates are predominantly negative: ten of twelve coefficients in Panels A and B have the predicted sign, consistent with the comovement prediction that spreading out within a technology class accompanies rising R&D investment. The strongest results appear for GTE 98th-percentile similarity, where co-inventor counts ($\beta = -3.844$, $p = 0.017$) and firm assignment ($\beta = -0.725$, $p = 0.096$) are both statistically detectable.

The estimates in Panels A and B are generally imprecise, reflecting the limited power in class-level annual changes ($N = 4{,}837$). Year-to-year variation in class-level similarity is noisy, attenuating coefficient estimates toward zero. Panel C addresses this power concern by extending the co-inventor analysis to the full 1836–2023 period using the CUSP dataset (Berkes 2016). The longer time series yields $N = 15{,}813$ CPC-class-year observations. Both GTE ($\beta = -0.011$, $p < 0.10$) and PaECTER ($\beta = -0.004$, $p < 0.05$) show statistically significant negative associations between similarity and co-inventor counts, consistent with the prediction. The cross-sectional analysis in Section 5, which exploits patent-level variation within class-years ($N = 219{,}772$), provides the sharper test.

Table S11.1: Time-Series Evidence: Changes in Similarity and Changes in Quality

	Δ Co-Inventors	Δ Firm Assignment	Δ Citations (5-yr)
Panel A: Δ Mean Similarity			
GTE	−3.364	−0.294	0.556
	(2.258)	(0.519)	(5.018)
PaECTER	−6.260	−0.411	−18.065
	(7.302)	(1.607)	(15.314)
Panel B: Δ 98th Percentile Similarity			
GTE	−3.844**	−0.725*	1.919
	(1.583)	(0.432)	(3.413)
PaECTER	−4.357	−1.558	−11.646
	(7.032)	(1.372)	(12.243)
Observations		4,837	
Fixed Effects		CPC Class, Year	
Panel C: Δ Mean Similarity, CUSP (1836–2023)			
GTE	−0.011*	—	—
	(0.005)		
PaECTER	−0.004**	—	—
	(0.001)		
Observations		15,813	
Fixed Effects		CPC Class, Year	

Notes: Each cell reports the coefficient from a separate bivariate regression of annual changes in the column variable on annual changes in the row similarity measure, at the CPC-class-year level. All specifications include CPC class and year fixed effects. Standard errors in parentheses, clustered at the CPC class and year levels. Panels A and B use modern patent data (1976–2023). Panel C uses CUSP data (Berkes 2016) spanning 1836–2023; firm assignment and citations are unavailable in CUSP. *** — $p < 0.01$, ** — $p < 0.05$, * — $p < 0.1$.

Appendix S12 Changelog

The analysis in this version of the paper differs slightly from a prior version circulated under the title "Patent Text and Long-Run Innovation Dynamics: The Critical Role of Model Selection" (NBER working paper 32934). This section documents those changes.

- We standardized corpora processing across representations. This led to revisions to PaECTER-based similarity measures in some years, after correcting prior data handling errors. Crucially, the prior errors did not affect PaECTER embeddings used in our validation tasks. There were also some slight revisions to TF-IDF similarity measures. As a result, we re-did the technology classification validation task. The ranking results remained the same, although the quantitative performance of TF-IDF representations worsened. Other representations were affected minimally across our results and validation tasks.

- We added sampling methods to obtain estimates of the standard deviation of pairwise similarity for each year and for each representation. This allowed us to compute standardized similarity as in Figure 2. Based on the sampled patent pair matrix, we were also able to estimate weighted average similarity (Section 4.3) and quantiles of average similarity (Section S9.5).

- We corrected a coding error in the between- and within-class similarity estimates (Figure 6). Intuitively, there are more between-class comparisons than within-class comparisons. Therefore, the between-class dynamics should resemble the overall similarity dynamics. In the updated version, they do so.