Mapping Inventions in Idea Space: Spreading Out Over an Expanding Frontier, 1836–2023*

Ina Ganguli Jeffrey Lin Vitaly Meursault Nicholas Reynolds

January 17, 2026

Abstract How do inventors position themselves across the landscape of technological possibilities? Using US patent text from 1836 to 2023, we document that inventors are spreading out: contemporary inventions have become substantially less similar over nearly two centuries. We corroborate this trend using patent interference data spanning 150 years, which shows declining rates of simultaneous invention. To explain this pattern, we develop a spatial competition model where inventors spread across idea space to escape crowding as technological opportunities expand. Spreading reduces knowledge spillovers and raises adaptation costs, creating a drag on TFP growth, while simultaneously boosting aggregate R&D spending through entry into new technological territories—spending that increases research inputs without raising productivity. By relating TFP and R&D growth to changes in measured similarity, we validate the model's predictions and estimate key parameters. Spatial forces—inventors spreading across idea space—account for more than two-fifths of the long-run decline in US research productivity documented by Bloom et al. (2020), with traditional forces (fishing out, burden of knowledge) explaining the remainder. Methodologically, we demonstrate that measuring technological similarity requires systematic validation of text analysis methods. Standard approaches can yield misleading conclusions; our validation framework using expert assessments and human judgments identifies superior methods and provides validated similarity measures for innovation research.

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. This paper subsumes a prior paper titled "Patent Text and Long-Run Innovation Dynamics: The Critical Role of Model Selection."

1 Introduction

How do inventors navigate the expanding landscape of technological possibilities? In choosing what to invent, they face a central trade-off: working on projects closer to others enables knowledge spillovers, but intensifies competition, while pursuing more distant ideas reduces competition but foregoes valuable spillovers. How inventors position themselves in "idea space"—the multidimensional landscape of all potential inventions—shapes the nature of innovation and research productivity.

This paper provides the first empirical evidence that inventors are spreading out across idea space: over nearly two centuries, American inventors have produced inventions that are increasingly dissimilar to each other. We document this striking secular decline in the similarity of US patents from 1836 to 2023, using state-of-the-art, validated language models applied to the full text of patent claims. We corroborate this trend using patent interference data spanning 150 years, which shows declining rates of simultaneous invention.

To interpret this pattern, we develop a spatial competition model where inventors choose locations in expanding idea space, weighing knowledge spillovers against competitive pressure. The model features a circular "idea space" with distance-dependent spillovers and adaptation costs: nearby inventors share knowledge but compete for profits, while distant inventors operate independently but forgo spillovers. As the space of technological opportunities expands—driven by accumulating knowledge that raises the fixed costs of frontier research (Jones 2009)—inventors spread out to maintain profitability. The equilibrium generates testable predictions for how spreading affects both innovation outcomes (TFP growth) and innovation inputs (R&D spending), and yields predictions consistent with established empirical regularities including rising R&D intensity (Hirschey et al. 2012), increasing patent rents (Bessen et al. 2018), and the cross-sectional relationship between technological proximity and knowledge spillovers (Bloom et al. 2013).

The model predicts that spreading out reduces aggregate TFP growth through two mechanisms: knowledge spillovers weaken as inventors become more distant, and downstream firms incur greater costs adapting technologies from distant inventors. Simultaneously, spreading out increases aggregate R&D spending: new inventors enter to serve newly accessible territories, and existing inventors intensify research quality in response to reduced competition. This creates a "spatial drag" on research productivity—TFP growth per unit of R&D input—through both reduced output and increased input.

We test these predictions using annual time-series data on US TFP growth, aggregate R&D spending, and our validated similarity measures over 1948–2015. The evidence strongly supports the model. Declining similarity correlates negatively with TFP growth and posi-

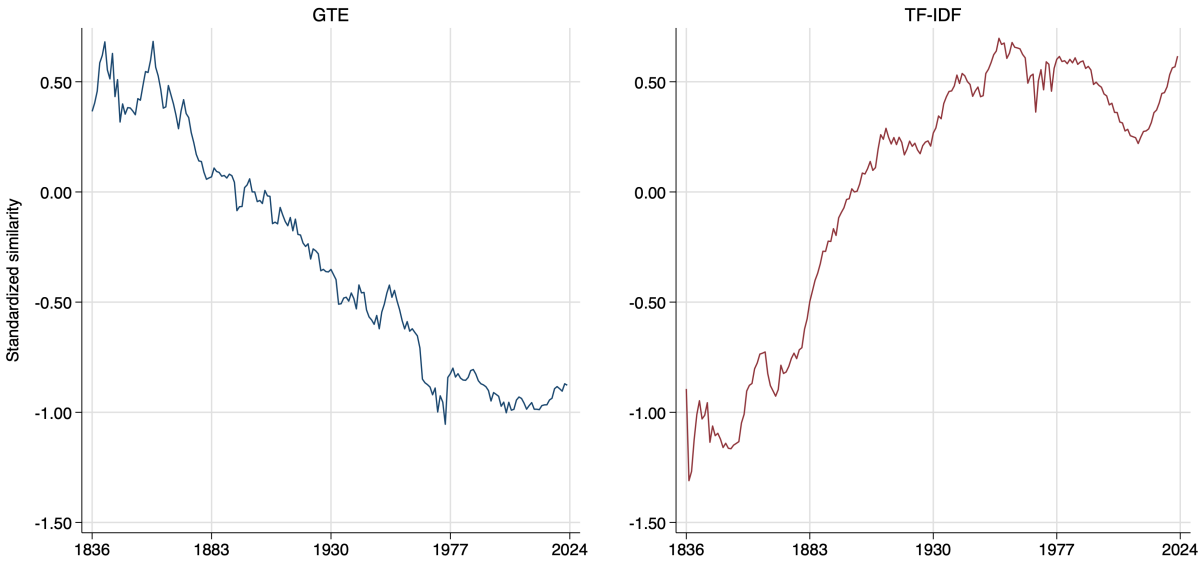


Figure 1: Similarity Trends Depend on Representation Choice

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 3. The 1.5σ-decline in similarity using the validated GTE representations contrasts sharply with the spurious 1.5σ-increase suggested by TF-IDF.

tively with R&D growth, as predicted. Structural time-series regressions validate the model's quantitative predictions and allow us to estimate key parameters, including the spillover attenuation rate, the share of variable versus fixed R&D costs (72% variable, consistent with NSF survey data), and the rate of idea space expansion (approximately 2.7% per year). These estimates enable a growth accounting decomposition: spatial forces—the spreading of inventors across idea space—account for more than two-fifths of the long-run decline in US research productivity documented by Bloom et al. (2020), with traditional mechanisms (fishing out, burden of knowledge) explaining the remainder.

However, measuring proximity in idea space is challenging: different text-based representations can yield opposite conclusions. As Figure 1 shows, the choice of Natural Language Processing (NLP) model fundamentally shapes inferences. With modern, neural net based GTE text embeddings (left panel), we observe the predicted historical decline in similarity. But with a widely-used traditional word-counting method (TF-IDF, right panel), the very same texts instead show increasing similarity: a result at odds with both intuition and theory.

To address this methodological challenge, we contribute a comprehensive validation and

model selection framework tailored to measuring invention similarity. Instead of presuming any single model is best, we systematically compare leading NLP approaches—ranging from TF-IDF to neural embeddings—using three independent, domain-relevant validation tasks. These tasks span nearly two centuries, test similarity at both detailed and coarse levels, and involve both general human knowledge and institutional expert ground truth. Perhaps our sharpest test asks which NLP model best predicts patents which are involved in an interference case— meaning they were deemed by patent examiners to reflect identical inventions (Ganguli et al. 2020). This approach reflects growing consensus in NLP that no method is superior for all tasks, but must be evaluated with data and benchmarks specific to its applied use (Ash and Hansen 2023; Grimmer et al. 2022).

Our results show that GTE and PaECTER embeddings substantially outperform both legacy (TF-IDF) and other modern representations (Doc2vec, USE, S-BERT); GTE, in particular, is robust across all validation regimes and time periods. This comprehensive evaluation directly addresses the concerns raised by Ash and Hansen (2023) of divergent economic conclusions from unvalidated text models.

In Appendix A, we illustrate the importance of validated measurement by re-examining trends in "breakthrough" inventions (Kelly et al. 2021). In this case the qualitative conclusions align across models, but downstream economic results are more robust and consistent for the NLP model which performed best at the validation tasks .

This paper contributes to both theory and empirical methodology. Theoretically, we advance the study of endogenous research allocation in idea space (Bloom et al. 2013; Dasgupta and Maskin 1987; Lamantia and Pezzino 2016); our use of comparative statics complements recent dynamic models (Bryan and Lemus 2017; Carnehl and Schneider 2025; Hopenhayn and Squintani 2021; Jones 2009), providing direct predictions for cross-sectional similarity. Our spatial framework provides a unified explanation for multiple empirical patterns previously analyzed in isolation. We connect our spatial equilibrium to the research productivity puzzle (Bloom et al. 2020). The model reveals three distinct spatial forces that contribute to research productivity decline, complementing "fishing out" (Kortum 1997) and the rising burden of knowledge (Jones 2009), traditional mechanisms that are also described in the model. One, spillovers weaken as inventors spread apart, reducing the effective output per unit of R&D input. Second, downstream firms incur larger costs adapting distant technologies, reducing productivity gains. Third, new inventors enter to serve new territories, absorbing resources without raising average productivity.

Empirically, we build on a rich literature measuring invention similarity through classifications (Akcigit et al. 2017; Bloom et al. 2013; Fleming 2001; Youn et al. 2015), keywords (Arts et al. 2025; Azoulay et al. 2019), citations (Berkes and Gaetani 2020; Verhoeven et al.

2016), and text (Feng 2020; Kelly et al. 2021; Lee and Hsiang 2019). While some compare methods (Arts et al. 2018, 2021; Cheng et al. 2022), we provide the first task-specific, comprehensive validation for economic applications to patent text. Our validated similarity is contemporaneous—comparing inventions to each other at the same time—and can serve as a building block for work on novelty, disruptiveness, and breakthroughness (Akcigit et al. 2017; Kelly et al. 2021; Park et al. 2023). Bloom et al. (2013) demonstrate that classification-based proximity meaningfully predicts knowledge spillovers and product market rivalry. We build on this work by developing text-based measures that provide finer-grained, continuous similarity metrics validated against multiple ground-truth sources.

For innovation research, our release of patent GTE and PaECTER representations for 1836—2023 can serve as a new standard, facilitating analysis of knowledge spillovers (Ganguli et al. 2020; Jaffe et al. 1993; Murata et al. 2014; Thompson and Fox-Kean 2005) and formal tests of idea-space theory (Akcigit et al. 2017; Clancy 2018; Dasgupta and Maskin 1987; Olsson 2000).

Methodologically, our work illustrates best practices for economic text-as-data analysis. Echoing Grimmer et al. (2022), we emphasize structured validation; in line with Ash and Hansen (2023), we show the risks of over-reliance on a single model. Our approach demonstrates (1) the value of evaluating multiple models, including non-proprietary options (open-source GTE competes with costly OpenAI APIs); (2) the need for bespoke, domain-specific validation tasks; and (3) the importance of sharing validated benchmarks as community resources. Our methodology offers a template for extracting meaningful economic measures from text, avoiding pitfalls of "black box" representations.

The rest of the paper proceeds as follows. Section 2 develops our stylized model of invention in idea space. Section 3 tests the spreading-out prediction and demonstrates its robustness. We show that the decline in similarity (i) remains after accounting for multi-patent entities, confirming it reflects how independent inventors space themselves; (ii) appears at both global and local spatial scales; (iii) persists within and across technology fields, ruling out industrial composition shifts; (iv) holds within cohorts defined by technology "age"; and (v) is corroborated by an independent data source: purpose-digitized historical patent interference rates over more than 150 years, indicating a decline in simultaneous inventions.

Section 5 details our validation and model selection methodology. Section 6 presents validation results, including evidence from patent interference cases (USPTO determinations of identical discoveries), human non-expert annotations (for historical patents), and official patent classifications (expert-labeled technological groupings). Section 7 concludes.

2 A Theory of Invention in Idea Space

This section develops a stylized model to illustrate how inventors position themselves across the landscape of technological possibilities. The central insight is simple: as knowledge accumulates, the space of potential inventions expands, faster than the number of active inventors. This creates a "spreading-out" force that drives inventors apart in idea space.

The key mechanism operates through rising entry costs. As the technological frontier expands, reaching it requires ever-greater sunk investments—more education, larger teams, more sophisticated equipment (Jones 2009). Inventors respond to rising entry costs by spreading out. When adaptation costs create sufficient product differentiation, inventors can restore profitability by capturing larger territories and charging higher patent prices—provided the revenue gain from serving more idea customers exceeds the cost of producing higher quality ideas. This generates the model's most distinctive prediction: as idea space expands, equilibrium spacing increases, making inventions systematically less similar over time. Other predictions align with established empirical patterns.

This spreading-out equilibrium also helps explain the secular decline in research productivity. As inventors move farther apart, knowledge spillovers weaken with distance, downstream firms incur larger costs adapting distant technologies, and resources flow to entering and serving new idea "territories" without raising average productivity. Combined with traditional fishing-out (Kortum 1997) and burden of knowledge (Jones 2009) mechanisms, these spatial forces contribute to the sharp decline in research productivity documented by Bloom et al. (2020).

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).[1] This *opportunity space* captures the breadth of feasible technological possibilities at the knowledge frontier. 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.

[1]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. These could represent individuals, teams, or firms; we abstract from the team-formation margin in Jones (2009), which is instead reflected in the increasing fixed cost of entry.[2] We 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:

$$f(H) = \phi H, \quad \phi > 0 \tag{1}$$

This specification captures the *burden of knowledge* documented by Jones (2009): as the frontier of knowledge expands, reaching it requires greater sunk investments.

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:

$$A_i(h) = Q_i - \tau h \tag{2}$$

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 mod-

[2] "Entry" refers to undertaking a research project at a location in idea space, not necessarily market entry by a new firm.

erately 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.[3]

The firm's net surplus after paying licensing fee p_i is:[4]

$$\text{Net Surplus}(h) = A_i(h) - p_i = Q_i - \tau h - p_i \tag{3}$$

Downstream firms use licensed technologies to produce differentiated consumption goods for final consumers. 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, justifying our reduced-form specification. The distinction 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.

[3]The literature on technology transfer and adoption provides additional evidence for substantial adaptation frictions. Teece (1977) finds technology transfer costs of 15-59% of project value, increasing with technological distance. Hippel (1994) documents that information is fundamentally "sticky," requiring substantial resources to transfer across contexts. Atkin et al. (2017) provide quasi-experimental evidence of large costs adapting production technology even within the same industry. These studies suggest adaptation costs are a first-order friction in technology deployment, consistent with the model's τh term being quantitatively important.

[4]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 captures 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.

Firms choose which inventor to license from to maximize net surplus (equation 3). This creates spatial competition among inventors, analyzed in Section 2.2.

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

$$c(q_i) = \frac{1}{2}\gamma q_i^2 \tag{4}$$

where $\gamma > 0$ is a cost scaling parameter capturing the convexity of R&D production. The variable q_i represents quality investment. This accommodates the "fishing out" mechanism (Kortum 1997): as the technological frontier advances, producing further increments requires progressively more resources.

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

$$Q_i = q_i + \frac{1}{2}\beta\left(1 - \frac{d_c}{\lambda}\right)q_c + \frac{1}{2}\beta\left(1 - \frac{d_r}{\lambda}\right)q_r \tag{5}$$

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).

Economic interpretation. An inventor's R&D investment q_i generates private knowledge that is costly to produce. But downstream firms benefit from the *total* knowledge available—both the inventor's own ideas and freely accessible spillovers from nearby inventors. These spillovers could arise from published research and patents, tacit knowledge diffusion, complementary ideas, or other sources of non-excludable knowledge. Distance-dependent spillovers capture the intuition that similar problems might have similar solutions.

Spillover decay. 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: inventors gain no advantage from asymmetric positioning.

Spillovers as pure externalities. In symmetric equilibrium where all inventors choose the same quality q, each inventor receives spillovers $\beta s(d)q$ regardless of their own investment. Crucially, this makes spillovers a *pure positive externality* that does not enter the private zero-profit condition: inventors' location and quality choices respond to private costs and revenues, not spillover benefits they receive (which cancel out in the symmetric equilibrium). This structure yields clean analytical solutions while preserving the result that spillovers matter for social welfare but not for equilibrium spacing.

No strategic complementarity. In this specification, spillovers do not create strategic complementarity: $\frac{\partial^2 Q_i}{\partial q_i \partial q_{-i}} = 0$. An inventor's marginal benefit from investing in quality q_i is independent of neighbors' investments q_{-i}. This additive structure yields clean analytical solutions and preserves the spreading-out result.[5]

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.[6]

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.

Optimal Pricing and Quality Inventor i chooses price p_i and quality q_i to maximize profit $\pi_i = R_i - c(q_i) - f(H)$, 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:

$$Q - p_i - \tau\tilde{h} = Q - p - \tau(d - \tilde{h}) \quad \Rightarrow \quad \tilde{h} = \frac{d}{2} + \frac{p - p_i}{2\tau} - \frac{q - q_i}{2\tau} \tag{6}$$

[5]Bloom et al. (2013) provide quasi-experimental evidence of strategic complementarity in R&D: when neighbors' R&D increases (due to exogenous tax credit shocks), firms increase their own R&D investment. Multiplicative spillovers (e.g., $Q_i = q_i(1 + \beta s(d) q_{-i})$) would capture this but require numerical solutions. Importantly, strategic complementarity would *reinforce* our spreading-out result: as inventors spread apart and spillovers weaken, each inventor faces lower marginal returns to R&D (since $\frac{\partial Q_i}{\partial q_i} = 1 + \beta s'(d) q_{-i} < 1$), making spreading more attractive relative to quality investment. Our additive baseline thus provides a conservative estimate of spatial effects.

[6]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.

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:

$$d + \frac{p}{\tau} - \frac{2p_i}{\tau} = 0 \quad \Rightarrow \quad \boxed{p = \tau d} \tag{7}$$

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:

$$2p_i \cdot \frac{1}{2\tau} = \frac{p_i}{\tau} = \gamma q_i \quad \Rightarrow \quad \boxed{q = \frac{d}{\gamma}} \tag{8}$$

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:

$$R - c(q) - f(H) = 0 \tag{9}$$

Substituting revenue $R = pd = \tau d^2$, cost (4), quality (8), and fixed cost (1):

$$\tau d^2 - \frac{1}{2}\gamma\left(\frac{d}{\gamma}\right)^2 - \phi H = 0 \quad \Rightarrow \quad \boxed{d^*(H) = \sqrt{\frac{\phi H}{\tau - \frac{1}{2\gamma}}}} \tag{10}$$

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 H are characterized by equations (5), (7), (10). These equilibrium values determine the number of inventors $n^* = H/d^*$, realized quality Q^* (8), 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 (Appendix 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 Appendix S1, a unique symmetric equilibrium exists. All downstream firms adopt a technology, and all inventors earn zero profits.*

Proof. See Appendix Appendix S1. □

With existence and uniqueness established, we proceed to analyze how equilibrium quantities respond to changes in opportunity space H.

2.3 Comparative Statics

We now analyze how equilibrium variables respond to changes in the opportunity space H. As knowledge accumulates, the space of potential inventions expands. (For example, in Weitzman (1998)'s combinatoric framework, H represents all yet-untried combinations of existing ideas and grows with the stock of knowledge.) These comparative statics yield testable predictions about innovation dynamics which align with established empirical findings along with new insights.

2.3.1 Spreading Out

Spreading out $dd/dH > 0$ is the central comparative static of interest. It captures the intuition that as the knowledge frontier expands, inventors spread out to capture larger territories. This result follows immediately from equation (10), which also yields the parameter restriction $\tau\gamma > 1/2$ for a real solution.

It is instructive for economic intuition to derive this result from the zero-profit condition. Differentiating totally with respect to H:

$$\frac{dR}{dd}\frac{dd}{dH} - \frac{dc}{dd}\frac{dd}{dH} - \frac{df}{dH} = 0 \quad \Rightarrow \quad \frac{dd}{dH} = \frac{df/dH}{dR/dd - dc/dd} = \frac{\phi}{\frac{dR}{dd} - \frac{dc}{dd}} \tag{11}$$

The comparative static is thus simply: spacing increases with H if and only if marginal revenue of increasing spacing exceeds its marginal cost. From $R = \tau d^2$, marginal revenue is $\frac{dR}{dd} = 2\tau d$, and marginal cost is $\frac{dc}{dd} = \gamma q \frac{dq}{dd} = \gamma \cdot \frac{d}{\gamma} \cdot \frac{1}{\gamma} = \frac{d}{\gamma}$. Thus, $\frac{dR}{dd} - \frac{dc}{dd} > 0$ yields the spreading-out condition $\tau\gamma > 1/2$.

Intuitively, when opportunity space expands, rising fixed costs squeeze profits. Inventors can restore profitability by spreading out to capture larger territories. This is profitable when

the revenue gain from serving more firms exceeds the cost increase from producing higher quality. The key mechanism is pricing power from differentiation. With adaptation costs, downstream firms face productivity losses when moving away from their nearest inventor. This creates differentiation that allows inventors to charge higher prices ($p = \tau d$) on larger territories.

Proposition 2 (Spreading Out)**.** *For parameters satisfying $\tau\gamma > \frac{1}{2}$, marginal revenue of expanding spacing exceeds marginal cost, and equilibrium spacing increases with opportunity 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 $dQ/dH = \frac{1}{\gamma}\left(1 + \beta - \frac{2\beta d}{\lambda}\right)\frac{dd}{dH} > 0$. This aligns with empirical evidence of increasing R&D spending per firm (Hirschey et al. 2012). Hall et al. (2005) find that citations per patent are increasing over time, consistent with rising realized idea quality. Kogan et al. (2017) document rising value per patent over time. Kelly et al. (2021) find a growing rate of "breakthrough" innovations.

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

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: $\frac{dR}{dH} = 2\tau d \cdot \frac{dd}{dH} = \frac{dR}{dd} \cdot \frac{dd}{dH} > 0$. This is consistent with the findings of Bessen et al. (2018) that patent rents have increased substantially over time.

Corollary 2. *Rising Prices and Revenue per Inventor: $dp/dH > 0$ and $dR/dH > 0$.*

A key comparative static is the number of inventions. It follows immediately from equation 10 and the definition $n = H/d$ that the number of inventions rises with idea space: $dn/dH > 0$. This is consistent with the growing number of issued patents 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 3. *Rising Number of Inventions $dn/dH > 0$.*

Our baseline model uses linear entry costs $f(H) = \phi H$. More generally, any entry cost function $f(H)$ that increases in H would yield the same qualitative result that the number of inventions grows more slowly than idea space H. Entry costs that increase sublinearly—e.g., $f(H) = \phi H^\alpha$ for $\alpha < 1$—would yield $dn/dH > 0$. Entry costs that increase too quickly—i.e., superlinearly $\alpha >> 1$—could lead to the counterfactual prediction that the number of

inventions *declines* with idea space. In what follows, we proceed with the baseline linear entry cost, as superlinear entry cost growth is not empirically well-supported.

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 remained roughly constant—implying a 95% decline in research productivity. Can our model account for this pattern?

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.

$$\rho = \frac{Q}{\frac{1}{2}\gamma q^2 + \phi H} \tag{12}$$

Second, define aggregate research productivity Π: average TFP delivered to downstream firms per aggregate R&D input. (This corresponds to measure in Bloom et al. (2020), who compute TFP growth relative to total effective research employment.) Define aggregate R&D spending and aggregate TFP[7]:

$$\text{Agg R\&D} = n \cdot [c(q) + f(H)] = n \cdot \left[\frac{1}{2}\gamma q^2 + \phi H\right] \tag{14}$$

$$\text{Agg TFP} = Q - \frac{\tau d}{4} \tag{15}$$

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

$$\text{Average TFP} = \frac{1}{d}\int_{-d/2}^{d/2}(Q - \tau|h|)\,dh = \frac{1}{d}\left(Qd - \frac{\tau d^2}{4}\right) = Q - \frac{\tau d}{4} \tag{13}$$

Then:

$$\Pi \equiv \frac{\text{Agg TFP}}{\text{Agg R\&D}} = \frac{Q - \frac{\tau d}{4}}{n \cdot [\frac{1}{2}\gamma q^2 + \phi H]} \tag{16}$$

This quantity measures the *social* return to R&D investment. Aggregate productivity Π is lower than per-inventor productivity ρ for two reasons.[8] 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 result in monopolistic competition models (Dixit and Stiglitz 1977) where entry increases variety but not average quality.[9] Specifically, it builds on the hybrid innovation frameworks of Howitt (1999) and Peretto (1998), which feature both horizontal expansion (new varieties/firms) and vertical innovation (quality improvement). A key but underappreciated feature of these models is that horizontal R&D expands the set of product lines without improving average productivity—new entrants match but don't exceed existing quality levels. While these papers addressed the 'scale effects' controversy of the 1990s, neither connected this structural feature to aggregate research productivity decline.

These are *average* returns. In equilibrium, marginal private returns equal marginal costs ($\partial R/\partial q = 0$), but marginal social returns exceed marginal costs due to positive spillovers: when inventor i raises quality q_i, this benefits neighbors through the spillover function $\beta(1 -$

[8]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.

[9]Unlike standard CES production models where variety creates direct gains through aggregation ($Y = [\int y(\omega)^\rho d\omega]^{1/\rho}$) (Grossman and Helpman 1993, ch. 3), technology adoption involves discrete choice: each downstream firm selects one production process from available alternatives. A firm cannot "blend" multiple patent designs—it must implement a single technology. Perfect substitution at the firm level is less a strong modeling assumption but more reflective of the binary nature of technology adoption. Variety benefits manifest through territorial coverage (downstream firms can access better-matched technologies) rather than firm-level aggregation. For modelling consumption, our reduced-form specification of downstream firm demand (equation (3)) is consistent with either horizontal product differentiation or CES preferences in final demand. While these microfoundations have different welfare implications—CES generates additional gains from variety expansion—they yield identical predictions for inventor equilibrium and productivity measurement.

$d/\lambda)q_i$. This classic positive externality implies equilibrium R&D investment is below the social optimum.[10] Importantly, as inventors spread out (d increases), the spillover externality $\beta(1 - d/\lambda)$ shrinks—the wedge between private and social returns narrows, but this reflects weakening knowledge flows rather than improved efficiency.

Both productivity measures decline as opportunity space expands. First, consider the three components of per-inventor productivity from equation 12. Three forces drive declining productivity:

(1) Weakening spillovers from spreading out. As inventors spread out (d increases with H), the spillover benefit $\beta(1 - d/\lambda)$ shrinks—knowledge flows decay with distance. While quality investment q rises to serve larger territories, realized quality $Q = q(1 + \beta - \beta d/\lambda)$ grows more slowly because the spillover multiplier declines. This limits output growth to order $\sqrt{H}$ despite increasing R&D effort.

(2) Fishing out—convex R&D costs. Quality investment exhibits diminishing returns due to convex costs $\frac{1}{2}\gamma q^2$ (Kortum 1997). Serving larger territories requires higher quality ($q = d/\gamma$ with $d \sim \sqrt{H}$), but costs grow quadratically in quality. Hence variable R&D costs $\frac{1}{2}\gamma q^2 = \frac{d^2}{2\gamma}$ grow linearly with H—faster than the sublinear output growth. This is the classic "fishing out" effect: additional R&D investment yields progressively smaller quality improvements.

(3) Growing burden of knowledge. Fixed entry costs ϕH rise directly with opportunity space (Jones 2009). As the knowledge frontier expands, mastering the existing stock to contribute new ideas becomes costlier. This burden grows linearly with H, squeezing the output-cost ratio independent of spillovers or R&D efficiency. Each generation of inventors faces higher fixed costs simply to reach the frontier.

Together, these forces create a stark imbalance: output grows as $\sqrt{H}$ (sublinearly, limited by weakening spillovers) while total costs—variable R&D plus knowledge burden—grow as H (linearly). Hence productivity $\rho \sim H^{-1/2}$ declines.[11]

Aggregate productivity decline inherits and extends these forces. To understand the

[10]The marginal social benefit of quality exceeds marginal private benefit by $2\beta(1 - d/\lambda)$ in symmetric equilibrium, where the factor of 2 reflects spillovers to both neighbors in the circular model.

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

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

Since $d^*(H) = \sqrt{\frac{\phi H}{\tau - 1/(2\gamma)}}$ increases with H (Proposition 2), and $\partial\rho/\partial d < 0$ from both the

forces driving aggregate productivity down, we decompose its derivative using economically meaningful primitives. Taking derivatives of equations (14) and (15) with respect to H:

$$\frac{d(\text{Agg R\&D})}{dH} = \underbrace{\frac{dn}{dH} \cdot [c(q) + f(H)]}_{\text{(4) Extensive margin: entry}} + \underbrace{n \cdot c'(q)\frac{dq}{dH}}_{\text{(2) Fishing out}} + \underbrace{n \cdot f'(H)}_{\text{(3) Burden of knowledge}} \tag{17}$$

$$\frac{d(\text{Agg TFP})}{dH} = \underbrace{\frac{dq}{dH}\left[1 + \beta\left(1 - \frac{d}{\lambda}\right)\right]}_{\text{Quality investment with spillovers}} - \underbrace{\frac{\beta q}{\lambda}\frac{dd}{dH}}_{\text{(1) Spillover attenuation}} - \underbrace{\frac{\tau}{4}\frac{dd}{dH}}_{\text{(5) Adaptation drag}} \tag{18}$$

By the quotient rule:

$$\frac{d\Pi}{dH} = \frac{1}{\text{Agg R\&D}}\left[\frac{d(\text{Agg TFP})}{dH} - \Pi \cdot \frac{d(\text{Agg R\&D})}{dH}\right] < 0 \tag{19}$$

This expression reveals why aggregate productivity declines. The term $\frac{d(\text{Agg TFP})}{dH}$ represents TFP output growth, while $\Pi \cdot \frac{d(\text{Agg R\&D})}{dH}$ represents the "productivity-adjusted" R&D cost growth. When R&D costs grow faster than TFP output—which they must if productivity is declining—the gap drives productivity further down.

Equations (17) and (18) decompose the productivity decline into five distinct forces. Forces (1)-(3)—weakening spillovers, fishing out, and burden of knowledge—operate as described in the per-inventor analysis above. Forces (4)-(5) are aggregate-specific mechanisms. Three raise R&D costs; two reduce TFP output.

Forces raising R&D costs:

(2)-(3) Fishing out and burden of knowledge: As in the per-inventor analysis, convex costs ($n \cdot c'(q)\frac{dq}{dH}$) and rising fixed costs ($n \cdot \phi$) both contribute to R&D spending growth.

(4) Extensive margin 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. This extensive-intensive trade-off mirrors the classic monop-

---

$1/d$ term and the declining spillover factor $(1 + \beta - \beta d/\lambda)$:

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

olistic competition result (Dixit and Stiglitz 1977): entry increases variety (here, idea-space coverage) but not average quality (here, TFP growth per R&D dollar). Importantly, 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.

Forces reducing TFP output:

(1) Spillover attenuation: $-\frac{\beta q}{\lambda}\frac{dd}{dH}$ — As described above, knowledge flows weaken with distance.

(5) Adaptation costs. 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 delivered to downstream firms. While adaptation costs grow in absolute terms, they grow slower than total R&D costs ($\sim \sqrt{H}$ vs. $\sim H$), so as a *fraction* of R&D spending they decline—a small offsetting force that is overwhelmed by the other four forces driving productivity down.

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 $d\Pi/dH < 0$*:* From equation (19), aggregate productivity declines when $\frac{d(\text{Agg TFP})}{dH} < \Pi \cdot \frac{d(\text{Agg R\&D})}{dH}$—i.e., when TFP output growth is slower than productivity-adjusted R&D cost growth. Substituting the explicit expressions from (17) and (18) and simplifying using the equilibrium relationships yields the condition $\frac{1}{4} < \frac{1}{\gamma\tau}[1 + \beta - \beta d/(2\lambda)]$. Under the spreading-out condition $\tau\gamma > 1/2$, the right-hand side exceeds $1/2 > 1/4$, establishing the result. □

This decomposition contributes to understanding the puzzle documented by Bloom et al. (2020). Their finding is that "ideas are getting harder to find," requiring exponentially rising research to sustain constant growth. Our spatial model offers three complementary mechanisms to Kortum (1997)'s fishing out mechanism and Jones (2009)'s burden of knowledge: growing drags on TFP output from weakening spillovers and adaptation costs as inventors spread out, and extensive margin expansion in differentiated idea space.

2.4 Discussion

The model's central predictions are spreading out and declining productivity. Spreading out arises whenever rising entry costs pressure profits and marginal revenue of expanding spacing

exceeds marginal cost, a condition satisfied when adaptation costs create sufficient product differentiation ($\tau\gamma > 1/2$). The key insight—inventors restore profitability by serving larger territories—is robust to alternative cost and revenue structures. Productivity decline is primarily driven by the extensive margin force ($dn/dH > 0$), which obtains whenever entry costs don't rise too rapidly, ensuring spacing grows slower than H. The other margins—weakening spillovers, convex costs, burden of knowledge, adaptation costs—would likely operate under other spillover decay functions, convex cost structures, and rising knowledge requirements. Thus, the model's central predictions appear to emerge from general economic forces rather than specific functional forms.

While our novel contribution is the spreading-out prediction, the model's other implications—rising R&D per firm, rising patent values, rising entry—align with established empirical patterns. Natural experiments provide additional validation: railroad expansion in 19th century Germany led to intellectual divergence (Chiopris 2024), and university endowment shocks generate stronger spillovers to technologically-similar firms (Kantor and Whalley 2014).

We adopt specific functional forms as a deliberate choice to obtain closed-form solutions. Alternative specifications would likely preserve qualitative results but require numerical solutions, sacrificing analytical transparency. The tractability-realism trade-off favors analytically transparent mechanisms that generate testable comparative statics for empirical validation.

Several extensions merit future work but lie outside our current scope. Stochastic invention processes (Kortum 1997) and heterogeneous inventors would enrich the model's realism but complicate equilibrium characterization. The static framework abstracts from explicit knowledge accumulation dynamics (Jones 2009) and takes final demand as exogenous, focusing instead on cross-sectional equilibrium forces that generate transparent comparative statics. Alternative mechanisms—exhaustion of low-hanging fruit in specific fields, changing innovation organization (large R&D labs), evolving patent practices—could also generate declining similarity; our empirical strategy addresses these by examining trends within technology classes and using independent historical validation.

The model could apply at multiple scales of the technological landscape. While our exposition frames the analysis in terms of aggregate idea space, the interpretation could be extended to individual technology fields where opportunity space represents possible directions within a field. Field-level interpretation has an important virtue: spreading reflects a fundamental mechanism operating within fields, not merely compositional shifts across fields. Our empirical analysis documents spreading-out within established technology classes as well as between them, validating both interpretations. This field-level reading also reveals

a potential compositional margin: emergence of new fields with low-hanging fruit could attract dense entry, temporarily increasing aggregate average similarity even as mature fields continue spreading.

Mapping Comparative Statics to Growth Rates: The static model characterizes equilibrium outcomes as functions of opportunity space H. To connect with secular trends in innovation, we adopt a *treadmill interpretation*: as H grows over time through knowledge accumulation, the comparative statics become growth rates.

Formally, if $H = H(t)$ expands exogenously, equilibrium variables $d^*(H(t))$, $q^*(H(t))$, $n^*(H(t))$ trace out time paths. By the chain rule, the time derivative of any variable $X(t) = X(H(t))$ is:

$$\frac{dX(t)}{dt} = \frac{dX}{dH} \cdot \frac{dH}{dt}$$

Define the *growth rate* of variable X as $g_X \equiv \frac{d\ln X}{dt} = \frac{1}{X}\frac{dX}{dt}$. Substituting the chain rule:

$$g_X = \frac{1}{X} \cdot \frac{dX}{dH} \cdot \frac{dH}{dt} = \frac{dX/dH}{X/H} \cdot g_H$$

where $g_H \equiv \frac{d\ln H}{dt}$ is the growth rate of opportunity space. For example, spreading out ($dd/dH > 0$) translates to positive spacing growth: $g_d = \frac{dd/dH}{d/H} \cdot g_H > 0$.

This mapping extends to TFP and R&D growth. Average log TFP is $A \equiv \log \text{TFP} = Q - \frac{\tau d}{4}$, where realized quality $Q = q(1 + \beta - \beta d/\lambda)$ includes spillover benefits. Since A is already in logs, its growth rate is simply its time derivative:

$$g_{TFP} \equiv \frac{dA}{dt} = \frac{dQ}{dt} - \frac{\tau}{4}\frac{dd}{dt} \tag{21}$$

Expanding dQ/dt using the product rule on $Q = q(1+\beta-\beta d/\lambda)$, substitution and collecting terms:

$$g_{TFP} = \frac{dq}{dt}\left(1 + \beta - \frac{\beta d}{\lambda}\right) - \left(\frac{\beta q}{\lambda} + \frac{\tau}{4}\right)\frac{dd}{dt} \tag{22}$$

The first term captures quality growth dq/dt scaled by the spillover factor $(1 + \beta - \beta d/\lambda)$. The second term represents combined spatial drag: *spillover attenuation* $(\frac{\beta q}{\lambda}\frac{dd}{dt})$ as rising spacing weakens knowledge flows, and *adaptation costs* $(\frac{\tau}{4}\frac{dd}{dt})$ as rising spacing increases average mismatch. Both forces reduce TFP growth as inventors spread out ($dd/dt > 0$).

Similarly, aggregate R&D spending is $\text{R\&D}(t) = n(t) \cdot [c(q(t)) + f(H(t))]$, where $c(q) = \frac{1}{2}\gamma q^2$ and $f(H) = \phi H$. Its growth rate is:

$$g_{R\&D} \equiv \frac{d\ln(\text{R\&D})}{dt} = g_n + \theta \cdot (1 + \eta)g_q + (1 - \theta)g_f \tag{23}$$

where $g_n \equiv \frac{d\ln n}{dt}$, $g_q \equiv \frac{d\ln q}{dt}$, $g_f \equiv \frac{d\ln f}{dt}$, the parameter $\theta \equiv \frac{nc(q)}{nc(q)+nf(H)}$ is the variable cost share, and $\eta = 1$ captures quadratic cost curvature (so $(1+\eta)g_q = 2g_q$ reflects that R&D costs grow faster than quality due to convexity). This decomposes R&D spending growth into entry (g_n), quality scaling ($\theta \cdot 2g_q$, the "fishing out" effect), and rising fixed costs ($(1-\theta)g_f$, the burden of knowledge).

We treat model variables q and Q as flow rates—R&D effort and productivity increments per period—rather than cumulative stocks. Each period's problem takes the accumulated knowledge baseline as given; current R&D flow q produces increment Q above that baseline. The *treadmill* analogy is: as the knowledge frontier H expands over time (the treadmill speeds up), inventors must increase R&D effort q (run faster) just to maintain their position. Rising effort compensates for spreading out and rising entry costs, but productivity per R&D dollar declines—like a runner expending more energy per unit distance as the treadmill accelerates. This flow interpretation avoids explicit modeling of knowledge accumulation dynamics while preserving the model's predictive content about how equilibrium quantities respond to frontier expansion. The reduced-form mapping abstracts from depreciation and lifecycle dynamics that appear in fully dynamic models (Jones 2009), focusing instead on cross-sectional equilibrium forces that determine testable growth rate predictions.

Equations (22) and (23) map directly to growth accounting à la Bloom et al. (2020). Research productivity $\Pi \equiv g_{TFP}/R$ measures TFP growth (the flow output of research) relative to research effort (the stock input), so its growth rate is $g_\Pi = \frac{d\ln g_{TFP}}{dt} - g_R$. This combines all five forces from the comparative statics (equations (17)–(18)): spillover attenuation and adaptation costs affect TFP growth (the numerator), while entry expansion, fishing out, and burden of knowledge drive research effort growth (the denominator). Each force contributes to research productivity decline, enabling quantitative decomposition.

The model generates testable predictions that we evaluate in subsequent sections. Most distinctively, we validate the spreading-out prediction using nearly two centuries of US patent text data (Section 3). We then connect spreading out to research productivity decline (Section 4).

3 Inventors are Spreading Out in Idea Space

Our theory predicts that as idea space expands, inventors spread out, making inventions less similar over time. This section tests this prediction by measuring contemporaneous invention similarity across nearly two centuries of American patents.

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 Pro-

Quest. 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 (Li et al. 2023), a neural network model that uses contrastive learning to explicitly separate similar and dissimilar texts. GTE is selected based on comprehensive validation against multiple benchmarks in Section 5, where we demonstrate it outperforms alternative representations including TF-IDF, S-BERT, and PaECTER.

For each year, we compute average pairwise cosine similarity across all patents.[12] To 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.

3.1 Main Finding: Declining Similarity Using GTE

Figure 2 shows average annual pairwise patent similarity using GTE, our validated representation (the series is normalized to 0 in 1900).[13] GTE exhibits 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.

[12]We use an efficient computational method that reduces complexity from $O(N^2)$ to $O(N)$ for unit-normalized vectors, detailed in Appendix Appendix S2. To estimate cross-sectional standard deviations, we subsample up to 10,000 patents per year; in years with fewer patents, we use all available patents.

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

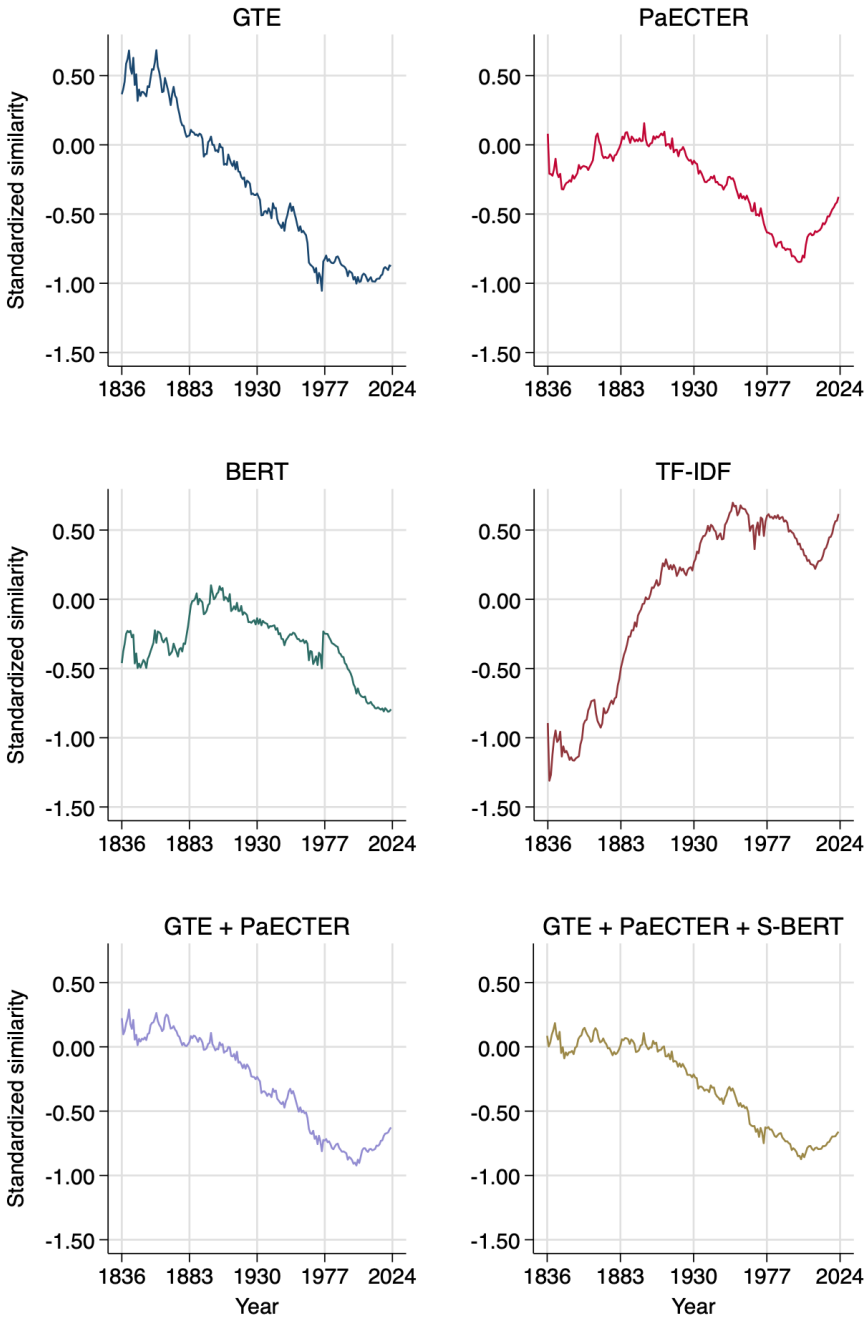


Figure 2: Similarity by Year and by Representation

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 S2 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 5.2. The bottom panels show ensemble estimates, discussed in Section 6.4.

Our main empirical finding is clear: **using GTE, 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, but remained essentially stable from 1980 to 2020. This timing tracks documented patterns in research productivity decline, which also moderated after 1980 (Bloom et al. 2020). Leveraging spillover elasticities from Bloom et al. (2013), we perform a back-of-the-envelope calculation: the 0.54σ decline in technological proximity from 1930 to 2015 could account for roughly 23% fewer citation-weighted patents, a 23% cumulative decline in R&D productivity, and approximately 6% of the overall decline in research productivity documented by Bloom et al. (2020) over this period (see Section 3.7 for detailed analysis).

Figure 2 also shows similarity trends from alternative representations for comparison. 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. (Interestingly, GTE, PaECTER, and BERT all shows similar declines from 1900 to 2000 of about 0.8–1.0σ.) Section 5 explains why these representations diverge and why our systematic validation identifies GTE as the most reliable measure for this analysis. The key insight: representation choice fundamentally affects conclusions, making validation-based model selection essential rather than optional.

3.2 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—portfolio strategies, product line

extensions, or claim differentiation—rather than the competitive-spillover trade-off central to our model. We therefore examine whether accounting for multi-patent entities affects our conclusions.

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 across the full patent corpus.[14] 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.

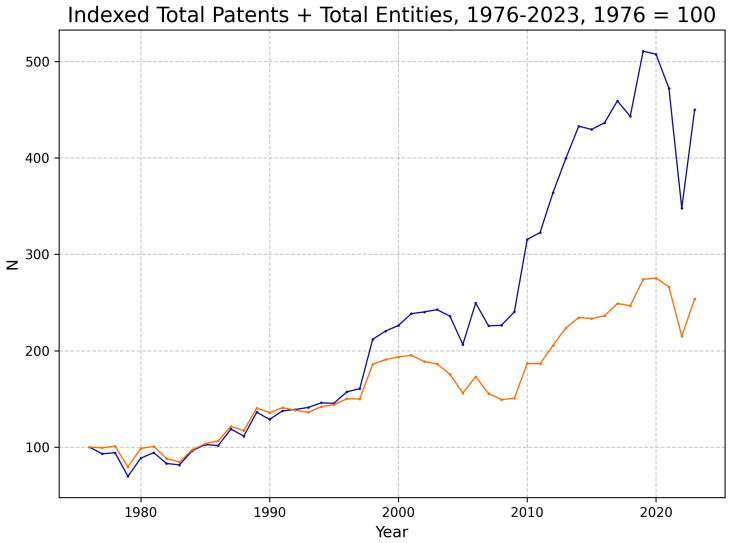


Figure 3: Growth in Patents and Patenting Entities

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.

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.

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

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.

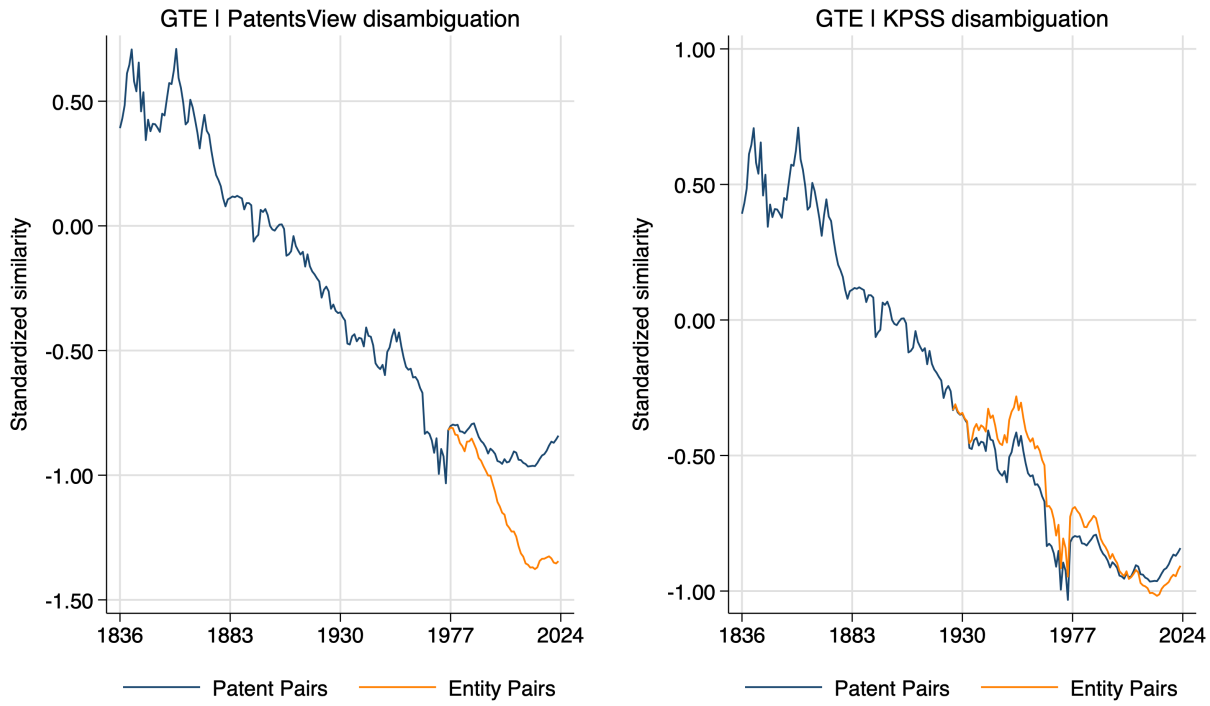


Figure 4: Similarity Correcting for Multi-Patent Entities

These plots show GTE similarity trends when sampling one patent per entity per year. The PatentsView disambiguation (left panel, 1976–2013) 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.

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.

The KPSS public firms disambiguation shows smaller effects, but this likely reflects its limited coverage: it excludes private firms, individual inventors, and non-practicing entities—perhaps actors most likely to file multiple similar patents. The PatentsView approach, which covers all assignees and individual inventors, provides more comprehensive entity identification for our research question.

Interpreting Recent Similarity Dynamics The stabilization of similarity after 1980 (or, after correcting for multi-patent entities, after 2015) presents an intriguing puzzle for our theoretical framework. Our model predicts continued spreading-out as long as idea space continues expanding and entry costs continue rising. Yet similarity remained roughly constant from 1980–2020 despite continued technological progress.

Several mechanisms could explain this stabilization.

- Advances in information technology, software, collaboration tools, improved access to knowledge repositories or methods of knowledge dissemination could have reduced entry costs and served as a countervailing force against dispersion.

- The emergence of entirely new technological domains (personal computing, internet, mobile, biotech) may have created localized clusters of related innovations: temporary "low-hanging fruit" that increased local density in specific regions of idea space.

- The rise of innovation platforms may have created network effects that pull inventors toward common standards and interfaces.

- The nature of patenting changed significantly after 1980 with the expansion of patentable subject matter to include software, business methods, and genetic sequences. These domains may have different natural similarity structures than mechanical and chemical inventions.

- Additionally, the rise of multi-patent entities may affect measured similarity even after our entity corrections, particularly if disambiguation algorithms are imperfect for the post-1999 period.

Our analyses in Section 3.3 and Online S2.5 provide some suggestive evidence of increased clustering at extremely local scales, especially after 1999. However, a full accounting of these various factors requires further research.

We view the post-1980 stabilization not as contradicting our theory but as highlighting that multiple forces shape the evolution of technological space. Our spreading-out mechanism operates alongside countervailing factors, and understanding their relative importance over time represents a valuable direction for future research. The strong co-movement of similarity and productivity decline through 1980, followed by joint stabilization, suggests our mechanism plays a meaningful role even if other forces matter in specific periods.

3.3 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 the introduction, local measures might be confounded by clustering around "low-hanging fruit" or other project-specific factors our model abstracts from.

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:

$$\text{Weighted Similarity} = \frac{1}{n}\sum_{i=1}^{n}\frac{\sum_{j\neq i}(1-d_{ij})e^{-\gamma d_{ij}}}{\sum_{j\neq i}e^{-\gamma d_{ij}}} \tag{24}$$

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 distance, while low γ emphasizes global average distance.

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.

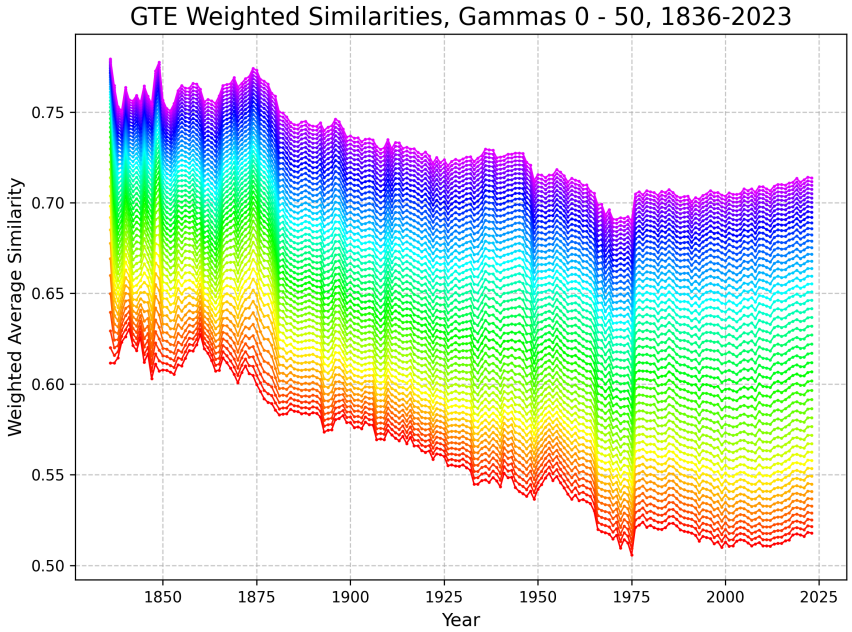


Figure 5: Similarity at Multiple Spatial Scales

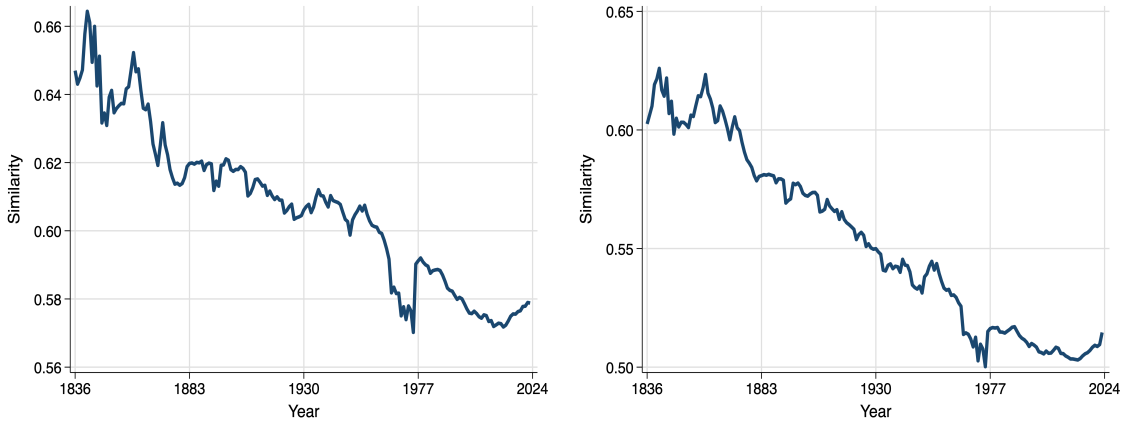
This figure shows GTE similarity trends using different spatial scales, with γ ranging from 0 (red line, lower envelope, global average) to 50 (purple line, upper envelope, emphasizing nearest neighbors). The secular decline in similarity is robust across all spatial scales, appearing at both local and global levels of idea space. The slight increase after 1999 is slightly faster for at local scales, indicating clustering.

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).

3.4 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. Em-



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

Figure 6: Similarity Within and Between Technology Classes

These plots show average pairwise patent similarity using GTE representations, decomposed into within-class (Panel A) and between-class (Panel B) components using three-digit CPC technology classifications. Both components exhibit secular decline, closely mirroring the overall trend from Figure 2.

pirically, 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-digit Cooperative Patent Classification (CPC) technology classes. (The CPC has eight top-level sections subdivided into over 120 three-digit 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. This reveals that inventors are spreading out at multiple levels of the technological landscape—not just moving across broad field boundaries, but also within established fields. The parallel decline in both components strengthens our interpretation that spreading-out reflects the fundamental mechanism emphasized by our theory rather than wholesale shifts in the industrial composition of innovation. It also 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 declining between-class similarity but stable or increasing within-class similarity. Instead, both decline in parallel.

3.5 Robustness: Similarity Trends by Technology Class Age

Another potential explanation for declining similarity is a shift in patent office standards or examination policies over time. If, for example, patent examiners have gradually raised the standard for novelty—requiring that each new patent be more distinct from prior inventions—this could produce a decline in similarity even in the absence of underlying changes in inventor behavior. Distinguishing between genuine changes in the inventive process versus shifting policy is therefore important for interpreting our findings.

To address this concern, Figure 7 examines similarity trends within technology classes as they age. Here, we define the "age" of a technology class as the number of years since it became substantively active, operationalized as the first calendar year in which at least 50 patents were issued in that class—serving as a proxy for the "birth" of the technology. For instance, class A01 ("Agriculture; forestry; animal husbandry; hunting; trapping; fishing") reached this threshold in 1843, while class C40 ("Combinatorial chemistry") did so in 2001.

Plotting average within-class similarity against class age (i.e., years since birth) reveals a clear pattern: as classes mature, within-class similarity steadily declines. Importantly, technology classes start at different historical moments, with "birth years" ranging from 1843 to 2001 (mean 1881, standard deviation 33 years). This staggering in class birth dates allows us to control for overall time trends and isolate the effect of maturity within fields.

The observed pattern—that mature classes display lower within-class similarity than younger classes, even within the same calendar year—suggests that declining similarity is not simply a product of changing patent office practices. Instead, this within-field, age-based comparison implies that the spreading out of inventions is closely tied to the endogenous evolution of the technological landscape, consistent with equilibrium forces emphasized by our theory.

3.6 Corroboration from Declining Interference Rates

We corroborate the declining similarity pattern using an entirely independent data source: patent interference rates from 1836 to 2014. This analysis provides particularly compelling evidence because pre-2001 interferences were not used in our validation process, making this a true out-of-sample test.

Patent interferences occurred when the USPTO 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

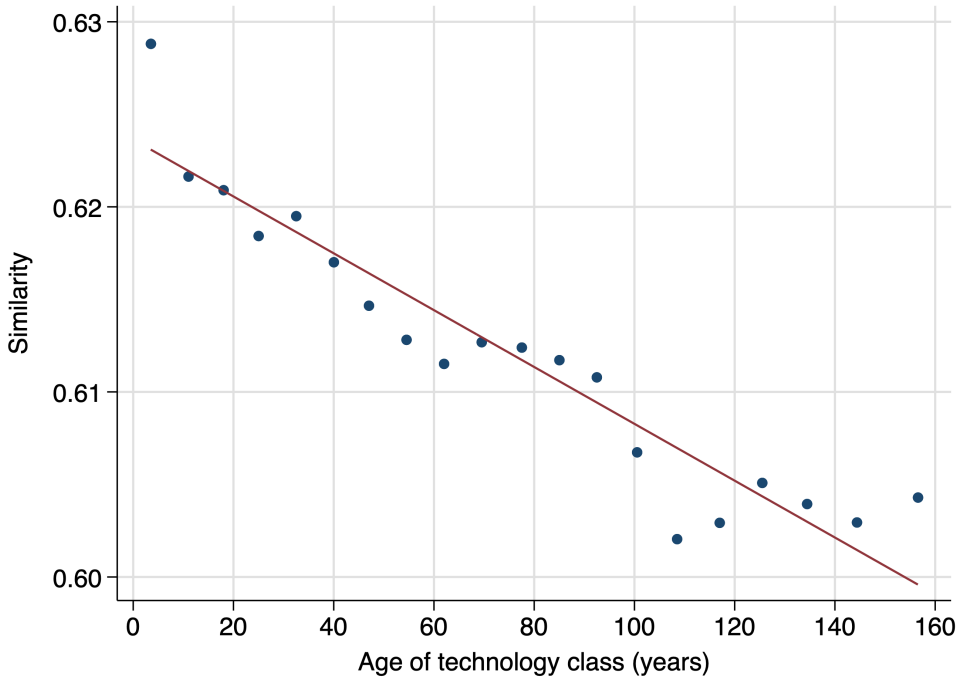


Figure 7: Similarity Within Class by Class Age

This binscatter plot shows 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. The steady decline in within-class similarity with increasing class age suggests that spreading out is a dynamic process tied to endogenous field evolution, rather than merely reflecting changing patent office standards.

spreading out in idea space.

Data Sources We construct a time series of interference rates from five distinct sources spanning different periods over 174 years:

- **1838–1900**: Patent Interference Case Files (Butler 1993): We digitized the finding guide to case files at the National Archives. Part I (1838–1869) allows tabulation of case files by year whereas Part II (1870–1900) does not. For 1838–1869, there are 29 case files per year. For 1870–1900, the finding guide identifies 2,682 surviving case files out of 27,271 sequentially-numbered cases. Surviving case files underestimate the true rate of interference. Not all cases assigned numbers are true interferences, so numbered cases likely overestimate the true rate of interference. The true number is bounded between 87 and 880 cases per year.

- **1864–1900**: We purpose-digitized the USPTO's *Registers of Interferences* from National Archives records, documenting 19,388 interference cases with an average of 504

annual terminations.[15] The Register data overlaps with the 31-year average upper bound estimated from Part II of the finding guide data.

- **1950–1962**: Summary statistics from Di Simone et al. (1963) report an average of 640 annual interferences.
- **1980–1994**: Data from Calvert and Sofocleous (1982, 1986, 1989, 1992, 1995) show an average of 237 annual interferences.
- **1998–2014**: Ganguli et al. (2020) document an average of 76 annual interferences.[16]

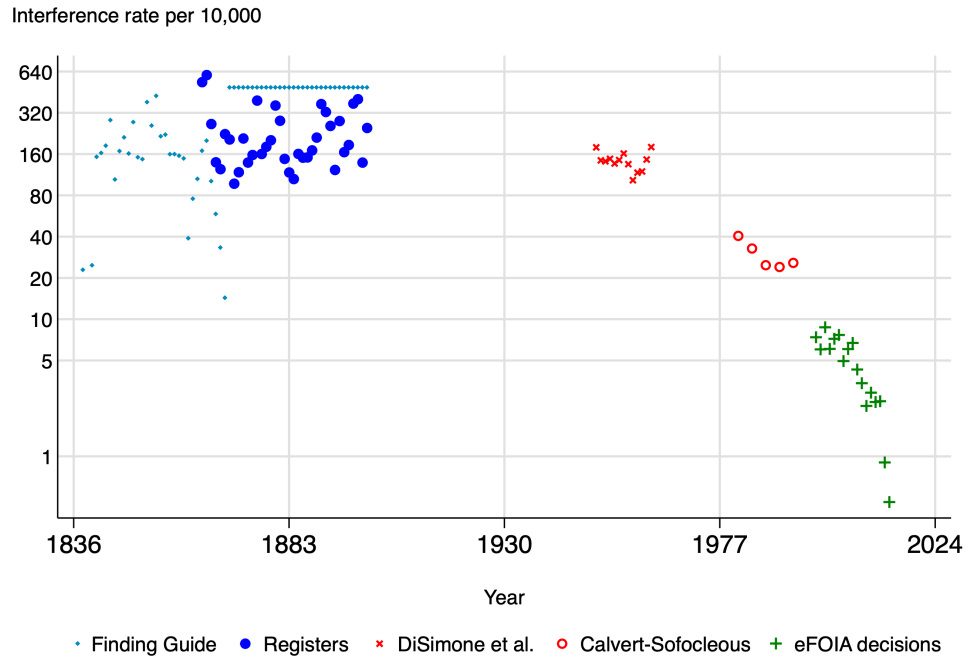


Figure 8: Interference Rate, 1838–2014

This plot shows the estimated interference rate per 10,000 issued utility patents across 150 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.

[15]See Online Appendix S3 for an example Register page.

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

Results Figure 8 reveals a striking and consistent decline in interference rates over 150 years. 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.[17]

This dramatic and steady reduction might be explained by changes in patent examination procedures alone. However, the USPTO's capacity to identify potential interferences likely improved over time, making early rates 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 completely 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 USPTO 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 3. 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.

3.7 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).

[17]The interference rate was at least 1.04% during 1838–1869 based on the finding guide data. The Copyright 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.

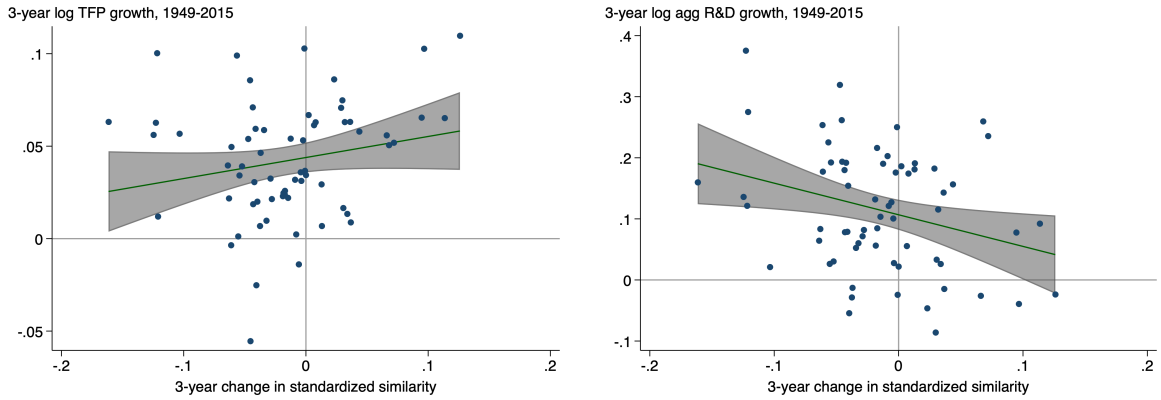


Figure 9: TFP Growth and Aggregate R&D Growth versus Changes in Idea Similarity
This figure plots 3-year log TFP growth and log aggregate R&D input growth from Bloom et al. (2020), 1948–2015, against 3-year changes in our GTE-PaECTER ensemble patent similarity measure. Declining similarity (spreading out) associates with lower TFP growth and faster R&D growth.

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.

4 Spreading Out and Research Productivity

The model predicts that as inventors spread out in idea space, aggregate downstream TFP growth slows due to weakening knowledge spillovers and rising adaptation costs. Moreover, aggregate R&D spending rises due to entry expansion, quality scaling, and rising entry costs. 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 9 illustrates the time series correlation between changes in similarity and changes in log TFP (and log aggregate R&D, discussed later), smoothed at 3-year intervals. Declining similarity (spreading out) associates with lower TFP growth and faster R&D growth, consistent with the equilibrium predictions of the model. We show that these estimated relationships validate the predictions of the model and identify key parameters. Using these identified parameters, we decompose the aggregate research productivity decline into its constituent forces.

4.1 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:

$$\Delta \log(\text{TFP})_t = b_0 + b_1 \cdot \Delta(-1 \times \text{Similarity})_t + b_3 \cdot t + \epsilon_t. \tag{25}$$

This equation relates TFP growth to (standardized) changes in technological distance $\Delta(-1 \times \text{Similarity})$ between inventors. If spreading out leads to declines in TFP growth (e.g., from spillover attenuation and adaptation costs) then $b_1 < 0$. The coefficient b_3 captures trend growth in TFP factors not captured by the theory.

This yields estimates $b_1 = -0.169$ (s.e. 0.057, $p < 0.01$) (Table 1, Column 1). The mean annual change in standardized similarity over 1948–2015 is -0.005, implying an average annual TFP drag from spreading-out of -0.084%/year.

Validation with Cross-Sectional Elasticity Estimates Our estimate validates with external quasi-experimental evidence. Bloom et al. (2013) and Lucking et al. (2019) estimate how firm-level TFP responds to its spillover pool, defined as the sum of its neighboring firms' R&D weighted by (inverse) idea-space distance. The spillover pool is instrumented using state R&D tax credit shocks. 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.[18]

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

[18]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 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.

$\log(\text{SPILLTECH}) = 1.04$:

$$\Delta \log(\text{SPILLTECH}) = -0.144 \times 1.04 = -0.150 \tag{26}$$

$$\Delta \log(\text{TFP}) = 0.206 \times (-0.150) = -0.031 \tag{27}$$

This predicts a 3.1% cumulative decline in TFP due to spreading over 1981–2001, or approximately -0.16%/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 $\log(\text{SPILLTECH}) = 1.17$, this yields an annualized TFP drag of $(-0.287 \times 1.17 \times 0.287)/(2015 - 1948) = -0.14\%$/year.

This cross-sectional elasticity measures the effect of technological distance on downstream TFP. It incorporates both the direct spillover mechanism (weaker knowledge flows from distant neighbors) and any frictions in absorbing or adapting distant knowledge. The estimate does not separately decompose these channels—it captures their combined effect on measured TFP. It does not include other equilibrium adjustments that affect productivity such as increases in spacing, entry, or quality.

The smaller size of the time-series estimates is consistent with equilibrium adjustments in the model that are not accounted for in the cross-sectional elasticity. To see this, substitute the equilibrium relationships $q = d/\gamma$ and $dq/dt = (1/\gamma)(dd/dt)$ for unobserved terms in (22). This yields:

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

This formula reveals that the reduced-form coefficient $\hat{b}_1$ in equation (25) reflects a combination of negative and positive contributions to TFP from spreading out. As inventors spread out, spillover attenuation and adaptation costs contribute negatively to TFP. But at the same time, gross quality improvements contribute positively to growth $(1 + \beta)$, creating an offsetting force.

Reduced form specification The reduced-form equation (28) also contains an interaction term $d \cdot \Delta d$ with a negative coefficient $-\frac{2\beta}{\gamma\lambda}$. As inventors spread out, they invest more in R&D. But this marginal R&D spending has reduced impact on TFP when inventors are farther apart because of spillover attenuation. Thus, the positive contribution of spreading out to TFP coming from increased R&D diminishes.

We include this interaction term (with coefficient b_2) in the regression reported in Col-

Table 1: Time Series Evidence: Technological Proximity and TFP Growth

			Multi-Year	
	Annual		3-Year	5-Year
	(1)	(2)	(3)	(4)
$-1 \times \Delta_t \text{Sim}$	-0.169***	-0.171**	-0.278***	-0.269***
	(0.057)	(0.083)	(0.095)	(0.098)
$(-1 \times \Delta_t \text{Sim}) \times (-1 \times \text{Sim}_{t-1})$		-0.015	-0.408	-0.571*
		(0.342)	(0.320)	(0.312)
R-squared	0.174	0.174	0.280	0.293
Observations	67	67	65	63
Implied annualized TFP drag from spreading out (%/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 (Δ_3 log and Δ_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 and a time trend (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. ***—$p < 0.01$, **—$p < 0.05$, *—$p < 0.1$.

umn 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.) 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 precisely estimated: increasing distance reduces TFP growth (-0.269, $p < 0.01$), and crucially, the interaction term is large and statistically significant (-0.571, $p < 0.10$). The implied TFP drag in 1991 from spreading-out remains consistent with, and slightly smaller, than the cross-sectional elasticity estimate ($0.156 < 0.16$).

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 net drag of -0.041%/year. At 2015 similarity levels, spreading out generates a net negative contribution to TFP growth of -0.136%/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).

4.2 R&D Spending Growth

We turn next to the supply side, examining how spreading out affects aggregate R&D spending. Equation (23) provides a structural relationship between R&D growth and its components. With quadratic costs ($\eta = 1$), variable cost growth is $g_{c(q)} = 2g_q$. Substituting the equilibrium relationship $q = d/\gamma$ yields $g_q = g_d$, and noting that $g_n = g_H - g_d$ (since $n = H/d$) and $g_{f(H)} = g_H$ (as $f \propto H$), equation (23) becomes:

$$\begin{aligned} g_{R\&D} &= (g_H - g_d) + \theta(2g_d) + (1-\theta)g_H \\ &= g_H(2-\theta) + (2\theta - 1)g_d \end{aligned} \tag{29}$$

This structural equation reveals that R&D growth depends on two forces: (1) expansion of idea space g_H with coefficient $(2-\theta)$, reflecting both entry growth $(g_H - g_d)$ and rising fixed costs $(1-\theta)g_H$; and (2) spreading out g_d (which has ambiguous effects—increasing variable costs via quality but slowing entry). When $\theta > 0.5$ (variable costs dominate), spreading out raises R&D spending.

The baseline model assumes constant g_H, but idea space growth may itself vary over time. To allow for this, we permit time-varying idea space growth: $g_H(t) = g_H^{1991} + \delta_H \cdot t$, where t is centered at 1991. Substituting into equation (29):

$$g_{R\&D,t} = [g_H^{1991} + \delta_H \cdot t](2-\theta) + (2\theta - 1)g_{d,t} \tag{30}$$

This motivates a reduced-form regression with both a time trend and changes in distance.[19]

$$g_{R\&D,t} = a_0 + a_1 \cdot t + a_2 \cdot \Delta d_t + \epsilon_t \tag{31}$$

[19]Including the time trend also parallels the specification in the TFP regression (equation (25)), which allows for trend growth in factors not explicitly modeled, except that in this equation the time trend has an explicit structural interpretation.

Table 2: Time Series Evidence: Technological Proximity and Aggregate R&D Growth

		Multi-Year	
	Annual	3-Year	5-Year
	(1)	(2)	(3)
$-1 \times \Delta_t$Sim	0.165	0.448**	0.438*
	(0.177)	(0.219)	(0.244)
Year (1991=0)	-0.000	-0.001	-0.002*
	(0.000)	(0.001)	(0.001)
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
Implied annualized R&D growth from spreading out (%/year):			
	0.082	0.223	0.252
Implied parameters:			
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. Following equation (31), the coefficient on $-1 \times \Delta_t$Sim is $a_2 = (2\theta - 1)$ Year is indexed to 0 in 1991, so the constant $a_0 = g_H^1 991(2-\theta)$, where $g_H^1 992$ is the baseline growth rate of idea space in 1991. 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. ***—$p < 0.01$, **—$p < 0.05$, *—$p < 0.1$.

where the time variable is centered at 1991 ($t = 0$ in 1991). The constant $a_0 = g_H^{1991}(2-\theta)$ captures baseline R&D growth in 1991, the time trend $a_1 = \delta_H(2-\theta)$ captures acceleration or deceleration in idea space growth, and $a_2 = (2\theta - 1)$ measures the effect of spreading out on R&D spending.

Table 2 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), we find that spreading out significantly increases R&D spending ($a_2 = 0.438$, $p < 0.10$), consistent with the theory's predictions.

Variable costs represent 72% of total R&D spending. This is consistent with National Center for Science and Engineering Statistics (n.d.) data showing labor and direct research

costs comprise 69% of business R&D.[20] This implies $\gamma\tau = 0.695$.[21]

The baseline growth rate of idea space in 1991 is $g_H^{1991} = 0.173/(2-\theta) = 13.5\%/5\text{years} = 2.7\%/\text{year}$.[22] The deceleration in idea space growth is a miniscule -0.026 percentage points per year (with the 95% confidence interval including zero), suggesting idea space is expanding at a roughly constant rate.

4.3 Contributions to Research Productivity Decline

The time-series regressions in Sections 4.1–4.2 quantify how spreading out affects both TFP growth (the output of R&D) and aggregate R&D spending (the input). We now use these estimates to decompose the decline in research productivity growth—defined as TFP growth minus R&D input growth—into its constituent forces.

Figure 10 reveals the research productivity decline documented by Bloom et al. (2020). Aggregate R&D spending grew at a roughly constant rate of 4.0%/year over 1948–2015, representing a 23-fold increase in research effort. Yet despite this sustained growth in inputs, TFP growth fell from 2.1%/year in 1948 to 0.7%/year in 2015—a decline by a factor of 3. Following Bloom et al. (2020), we define *research productivity* as the ratio of TFP growth

[20]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.

[21]From the equilibrium conditions, the variable cost share is $\theta = \frac{c(q)}{c(q)+f(H)}$. Substituting $c(q) = \frac{d^2}{2\gamma}$ (from $c = \frac{1}{2}\gamma q^2$ and $q = d/\gamma$) and the zero-profit condition $\phi H = d^2(\tau - \frac{1}{2\gamma})$, we obtain:

$$\theta = \frac{\frac{d^2}{2\gamma}}{\frac{d^2}{2\gamma} + d^2(\tau - \frac{1}{2\gamma})} = \frac{\frac{1}{2\gamma}}{\tau} = \frac{1}{2\gamma\tau}$$

Thus $\gamma\tau = \frac{1}{2\theta}$. With $\theta = 0.719$, we get $\gamma\tau = 0.695 > 0.5$, empirically confirming the spreading-out condition (Proposition 2).

[22]As an independent validation, we measure idea space expansion directly from patent embeddings. Applying principal component analysis to reduce dimensionality (7 dimensions) and computing the approximate convex hull volume, we estimate idea space expanded at 1.6-1.7%/year. This serves as a lower bound on true expansion: PCA captures the dominant modes of variation, but innovation spreading into higher-order dimensions (beyond the top 7 PCs) would be compressed by dimensionality reduction. The regression-implied $g_H \approx 2.7\%/\text{year}$ likely reflects expansion across all dimensions. The consistency between these independent approaches—one geometric, one structural—validates our interpretation of H as idea space expansion.

(the output of research) to the level of research effort: $\Pi_t \equiv g_{TFP,t}/R_t$. Taking logs and time derivatives, the growth rate of research productivity is:

$$g_\Pi = g_{g_{TFP}} - g_R \tag{32}$$

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

$$g_\Pi = -1.6\% - 4.0\% = -5.6\% \text{ per year} \tag{33}$$

This estimate is very close to Bloom et al. (2020)'s estimate of -5.1% per year (the difference reflects our use of 5-year frequencies and our exclusion of the non-BLS productivity data prior to 1948). The majority of this decline—by construction, 4.0 out of 5.6 percentage points—reflects the rising "treadmill" of required research effort. The remaining 1.6 percentage points reflects the deceleration in TFP growth itself. Our decomposition quantifies how much of this decline can be attributed to spatial forces—the spreading of inventors across idea space.

We begin by decomposing the R&D spending growth into its underlying components. The R&D regression estimates (Table 2) imply $\theta = 0.719$ and $g_H^{1991} = 2.70\%$/year. From these inputs, we can back out the implied growth rates of the various components of R&D spending. From equation (29), we have:

$$4.0\% = 2.70\%(2 - 0.719) + (2 \times 0.719 - 1)g_d \quad \Rightarrow \quad \boxed{g_d = 1.23\%/\text{year}} \tag{34}$$

From $n = H/d$:

$$g_n = g_H - g_d = 2.70\% - 1.23\% \quad \Rightarrow \quad \boxed{g_n = 1.47\%/\text{year}} \tag{35}$$

This is reasonably consistent with the average growth in unique patent assignees from 1976–2023 of 1.96%/year, providing external validation of this implied entry growth rate. We would expect assignee growth to exceed growth in new ideas if the number of ideas per patent is declining over time, which would be consistent with the rise of defensive patenting and patent thickets.

From equilibrium $q = d/\gamma$:

$$g_q = g_d \quad \Rightarrow \quad \boxed{g_q = 1.23\%/\text{year}} \tag{36}$$

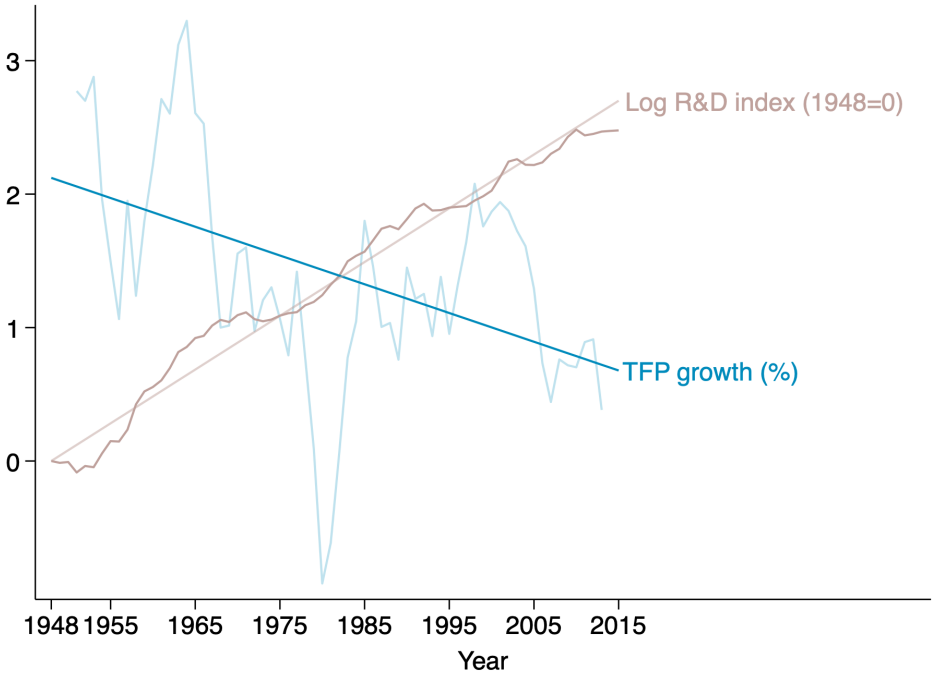


Figure 10: TFP Growth and Real R&D Growth

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.

From $g_{R\&D} = g_n + \theta(1+\eta)g_q + (1-\theta)g_f$ with $\eta = 1$:

$$4.0\% = 1.47\% + 0.719(2)(1.23\%) + 0.281 \cdot g_f \quad \Rightarrow \quad \boxed{g_f = 2.70\%/\text{year}} \tag{37}$$

To sum up, the 4.0%/year aggregate R&D spending growth can be decomposed into:

1. **Entry expansion ($g_n = 1.47\%$/yr):** Covering new technological territories. Represents $1.47/4.0 = 37\%$ of R&D growth. This is the extensive margin: more inventors, but no improvement in average TFP.

2. **Quality intensification ($\theta(1+\eta)g_q = 1.77\%$/yr):** Deepening within territories. Represents $1.77/4.0 = 44\%$ of R&D growth. Convex costs mean spending grows faster than quality ($2\times$ multiplier from $\eta = 1$).

3. **Rising fixed costs ($(1-\theta)g_f = 0.76\%$/yr):** Burden of knowledge. Represents $0.76/4.0 = 19\%$ of R&D growth. Each inventor requires more resources (education,

equipment, setup).[23]

Spatial Contribution to Research Productivity Decline We now complete the accounting by combining the TFP and R&D decompositions. On the TFP side, the regression (Table 1) identified 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. Thus, the spatial contribution to $g_{g_{TFP}}$ is $6.8\% \times -1.6\%$ = -0.11%/year. The remaining $g_{g_{TFP}}$ contribution of -1.49%/year reflects non-spatial factors outside the model, such as compositional shifts across technological domains, changes in research organization, or measurement issues.

On the R&D side, the 4.0%/year spending growth decomposes into entry expansion ($g_n = 1.47\%$/year), quality intensification ($\theta(1+\eta)g_q = 1.77\%$/year with $\eta = 1$), and rising fixed costs ($(1-\theta)g_f = 0.76\%$/year). We further decompose quality intensification into spatial and non-spatial components. The equilibrium condition $q = d/\gamma$ implies quality deepening responds one-for-one to spreading out: $g_q = g_d = 1.23\%$/year. However, convex R&D costs (with curvature parameter $\eta = 1$ representing quadratic costs) mean spending grows faster than quality. The linear component $\theta g_q = 0.88\%$/year is fundamentally spatial—quality must keep pace with distance. The convex wedge $\theta\eta g_q = 0.88\%$/year reflects fishing out, a non-spatial technological parameter capturing diminishing returns to R&D effort. This parameter would apply even without spreading out and represents a technological constraint rather than a spatial phenomenon.

Combining all channels, spatial forces contribute through: (i) TFP drag acceleration (-0.11%/year), (ii) entry expansion (-1.47%/year), and (iii) quality scaling (-0.88%/year), totaling $\Delta g_{\Pi,\text{spatial}} = -0.11\% - 1.47\% - 0.88\% = -2.46\%$ per year. This represents $2.46/5.6 = \boxed{44\%}$ of the total research productivity decline. Non-spatial forces—fishing out and rising fixed costs—contribute the remaining 56%. The spatial contribution underscores that the geography of idea space is a first-order determinant of aggregate innovation productivity, operating through both direct effects (spillover attenuation, adaptation costs) and equilibrium

[23]The baseline calibration uses $\eta = 1$ (quadratic costs), implying $(1+\eta) = 2$. However, quasi-experimental estimates from firm-level tax credit variation suggest lower curvature. Guceri and Liu (2017) estimate a user cost elasticity of -1.6 from UK corporation tax records, which maps to a cost curvature parameter of $1+\eta = 1.625$ (i.e., $\eta = 0.625$). With this well-identified parameter, the quality contribution to R&D spending growth reduces to $\theta(1+\eta)g_q = 0.719(1.625)(1.23\%) = 1.44\%$/year, implying a greater contribution from rising fixed costs: $g_f = 3.88\%$/year contributes $(1-\theta)g_f = 1.09\%$/year.

responses (R&D reallocation toward entry and quality deepening).

5 Validation Framework and Model Selection

5.1 The Representation Challenge

Figure 1 illustrates a fundamental challenge for testing our spreading-out theory: different NLP representations of the same patent text can yield opposing conclusions about invention similarity trends. Using GTE embeddings, we observe the predicted decline in similarity over nearly two centuries, consistent with our theory as inventors spread out across expanding idea space. Yet using TF-IDF representations on identical patent text, we find a dramatic *increase* in similarity over the same period, contradicting our theoretical prediction. See Online Appendix S4 for visualizations of contrasting representations.

This divergence is at the heart of our empirical contribution. Testing our theory requires measuring proximity in idea space, but ideas themselves are not directly observable. We rely on patent text and NLP methods to map inventions into measurable similarity spaces. The representation choice problem is compounded by the rapid evolution of NLP technology. Researchers now face an abundance of options, from traditional approaches like TF-IDF to sophisticated neural network models. Each method makes different assumptions about how text maps to meaning and captures different aspects of semantic similarity.

Unlike structural economic models where we can examine functional forms and behavioral assumptions, these models often operate as "black boxes" with complex engineering choices that make *a priori* evaluation difficult. Traditional validation approaches—selecting a single representation and showing it correlates with benchmark measures—cannot distinguish between methods that capture different aspects of the same underlying concept or identify which method best captures the specific concept of interest.

Our solution is systematic validation-based model selection. Rather than assuming any single representation is correct, we develop a comprehensive framework that evaluates multiple NLP approaches using external ground truth measures designed to capture different aspects of technological similarity. This approach allows us to move beyond arbitrary choice to an evidence-based selection process, identifying which methods are most reliable for measuring invention similarity and testing our spreading-out prediction.

5.2 Divergent Results from Alternative Representations

Before describing our validation framework, we illustrate why model selection matters by examining how different representations produce qualitatively different similarity trends.

Figure 2 in Section 3 compared four leading approaches: GTE, PaECTER, S-BERT, and TF-IDF. 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 1.5σ below the historical maximum—our main finding of spreading-out.

PaECTER suggests steadily declining patent similarity from 1898 through 1999 (approximately -0.8σ), followed by partial retracing through 2023.

S-BERT indicates steadily declining patent similarity from the early 20th century through 2023. The overall decline is similar in magnitude compared with GTE or PaECTER—about -0.8σ. The pattern is qualitatively consistent with spreading-out.

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 1.5σ below its maximum. This pattern directly contradicts our theoretical predictions and suggests inventors are clustering rather than spreading out.

These divergent patterns raise a critical question: which representation should we trust? All four representations technically "passed" basic validation—they beat random chance on our validation tasks. Without comparative validation, a researcher might have selected TF-IDF and reached the opposite conclusion about our theory. This underscores why validation-based model selection is essential rather than optional. The following subsections describe our systematic approach to answering this question.

5.3 Validation Overview

Our validation framework evaluates multiple NLP representations using three complementary tasks. First, patent interference cases—USPTO determinations that independent inventors made identical discoveries (Ganguli et al. 2020)—provide the strongest ground truth for technological similarity. Second, generalist similarity ratings from human annotators assess technological proximity on historical patents (1880–1920), testing temporal robustness. Third, patent office classifications assigned by examiners test whether representations capture categorical technological relationships.

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). Representations performing consistently well across multiple tasks are more likely to reliably measure the invention similarity our theory predicts should decline over time. Section 6 presents detailed results. For more discussion, see Online Appendix S5.

5.4 Representations

We compare multiple approaches for mapping patent text to numerical representations that can be used to measure similarity. We denote a representation 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 basd on method m.

A traditional mapping uses patent office technology classifications (Jaffe 1986), primarily administrative tools designed for prior art searches. A class-based mapping represents each patent as a binary vector indicating class membership. While straightforward, this treats all patents within a class as equally similar and all patents across classes as equally dissimilar.

NLP methods offer finer granularity.[24] We evaluate traditional frequency-based approaches (TF-IDF) and modern neural embeddings (Doc2vec, USE, S-BERT, GTE, PaECTER, OpenAI), which make different assumptions about mapping text to meaning.

Frequency-Based Representations The workhorse model TF-IDF (Sparck Jones 1972) represents patents based on word frequency, weighted by the inverse of word frequency across all patents. The TF-IDF vector for patent i has elements $c_{i,k} = TF_{i,k} \cdot IDF_{i,k}$, where k indexes unique words, $TF_{i,k} = n_{i,k} / \sum_j n_{i,j}$ (Term Frequency), and $IDF_{i,k} = \log(\frac{\text{total patents}}{\text{patents containing word k}})$ (Inverse Document Frequency). This approach captures which words are distinctive to each patent but treats words as independent and ignores semantic relationships.

Neural Network Embeddings Modern NLP methods produce distributed embeddings that capture semantic relationships between words and documents. We evaluate several approaches that differ in their training objectives and architectural choices. Doc2vec (Le and Mikolov 2014; Mikolov et al. 2013) extends word embeddings to documents using neural networks trained to predict context. Universal Sentence Encoder (USE) (Cer et al. 2018) produces sentence-level embeddings designed for semantic similarity tasks. S-BERT (Reimers and Gurevych 2019) adapts BERT (Devlin et al. 2019) for sentence similarity by fine-tuning on semantic textual similarity datasets.

More recent models include GTE (Li et al. 2023), which uses contrastive learning to explicitly separate similar and dissimilar texts, and OpenAI's proprietary embedding models. We also evaluate PaECTER (Ghosh et al. 2024), a model specifically trained on patent data using citation relationships as similarity signals.

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

The engineering choices underlying these models—training objectives, data sources, architectural decisions—significantly impact performance but are often proprietary or poorly documented (see Online Appendix S6 for technical details). This opacity makes empirical validation essential.

5.5 Validation Framework

The central challenge is that we cannot directly observe true similarity between inventions. Our solution uses external ground truth—independent measures not relying on the text representations we validate. For each validation task j, we evaluate how well similarity measures from representation m align with ground truth:

$$V^j(m) = S^j\left(1 - d^m(\mathbf{p}), g^j(\mathbf{p})\right) \tag{38}$$

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 ROC AUC or PR AUC.

For example, in the interference task, g^j indicates whether patent pairs were in interference, while representations provide continuous similarity scores. The score function measures how well these rankings predict interference status.

This framework addresses Figure 1's challenge: different representations yield different conclusions. Rather than assuming any representation is correct, we evaluate each against multiple independent benchmarks. Representations consistently aligning with external ground truth provide more reliable measures for economic analysis. When validation tasks disagree, we weight results by task relevance, ground truth reliability, and performance magnitude.

6 Validation Task Results

This section evaluates alternative NLP representations using three complementary validation tasks. Patent interferences test whether models can identify applications that patent examiners deemed legally identical, establishing a floor for acceptable performance. Human similarity judgments test whether models align with general assessments of technological relatedness, particularly for historical patents. Patent classifications test whether models capture the coarse-grained technological categories used by the patent office. Together, these tasks span different time periods (1850–2023), scales of similarity (identical claims to broad technological fields), and sources of judgment (expert examiners, generalist annotators, and

institutional classifications), ensuring our selected representations perform robustly across the range of similarity concepts relevant to our analysis.

6.1 Interferences

Our first validation task uses patent interferences, a unique feature of US patent law until March 2013 that provides a benchmark for measuring invention similarity.[25]. Patent interferences were USPTO 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). Interferences represent expert judgment by patent examiners—trained in both law and technology—that two independent applications describe identical inventions. This makes them an ideal validation task: if a similarity measure cannot identify applications that patent experts deemed identical, it is unlikely to reliably measure similarity concepts that correspond to our theory.

Data and Measurement We use 215 interference cases decided between 2001 and 2014, obtained from the USPTO's e-FOIA Reading Room and encoded by Ganguli et al. (2020).[26] Each case involves two or more independent parties, each with competing claims to the same invention contained in one or more patent applications. The 215 cases correspond with 440 distinct patent applications. This produces 96,580 application pairs, of which 322 are interfering pairs—applications from opposing parties making overlapping claims of invention.

We compute cosine similarity for every application pair using vector representations from seven NLP models. We also construct a baseline similarity measure based on shared CPC classifications (see Section 6.3). This creates a dataset where each row represents an application pair with similarity scores from each method and an indicator for true interference.

Performance Metrics and Economic Interpretation We evaluate each representation's ability to classify interfering versus non-interfering application pairs. To provide economic intuition, consider a patent examiner seeking to identify likely interferences among pending applications. The examiner can rank application pairs by similarity and investigate

[25]For more detail, see Ganguli et al. (2020)

[26]To select this sample, we conditioned on the availability of claims in interference, which depended on the level of detail in the decisions, and the availability of full application text, which eliminated interferences involving applications filed between 1998–2000.

Table 3: Rankings: Threshold-based Metrics

(a) Separate F1-max. thresh.

Rank	Repr.	TP	FP	F1
1	PaECTER	168	58	0.67
2	GTE	170	82	0.64
3	OpenAI	182	123	0.63
4	S-BERT	143	90	0.56
5	TF-IDF	110	67	0.48
6	USE	85	58	0.40
7	doc2vec	50	72	0.25
8	Class	98	792	0.17

(b) Separate F10-max. thresh.

Rank	Repr.	TP	FP	F10
1	PaECTER	265	1,862	0.90
2	GTE	259	1,222	0.90
3	OpenAI	255	1,118	0.89
4	S-BERT	250	3,001	0.82
5	TF-IDF	253	5,306	0.77
6	USE	235	4,984	0.72
7	Class	209	6,255	0.62
8	doc2vec	198	17,944	0.44

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.

those above a chosen threshold. This creates a fundamental trade-off: higher thresholds reduce investigative costs (fewer false positives) but risk missing true interferences (lower recall), while lower thresholds capture more true interferences but burden staff with unnecessary investigations (more false positives).

Different threshold choices yield classifiers with varying performance. We evaluate representations using four complementary metrics. The F1 score is the harmonic mean of precision and recall, which is appropriate when both are valued equally. The F10 score weights recall ten times more than precision, reflecting scenarios where missing interferences is costlier than unnecessary investigations. Receiver Operating Characteristic Area Under Curve (ROC AUC) and Precision-Recall Area Under Curve (PR AUC) measure true positive versus false positive rates and precision versus recall, respectively, across all thresholds.

Table 3a reports F1 scores at each representation's optimal threshold. PaECTER achieves the highest score (67%), followed closely by GTE (64%) and OpenAI embeddings (63%), significantly outperforming S-BERT (56%) and TF-IDF (48%).

When prioritizing interference detection over investigation costs (F10 score), PaECTER, GTE, and OpenAI embeddings perform nearly identically, retrieving 90%, 90%, and 89% of true interferences respectively (Table 3b). Critically, while all competitive models identify roughly 250–265 true interferences out of 322 total, the top three dramatically reduce false positives: they generate 1.6–2.7 times fewer false positives than S-BERT and 2.8–4.7 times fewer than TF-IDF. This reduction would significantly decrease unnecessary examiner investigations while maintaining high detection rates. USE and doc2vec prove uncompetitive,

Table 4: Rankings: Non-threshold-based Metrics

(a) ROC AUC

Rank	Repr.	ROC AUC
1	PaECTER	0.99
2	GTE	0.99
3	OpenAI	0.99
4	S-BERT	0.98
5	TF-IDF	0.98
6	USE	0.96
7	Class	0.85
8	doc2vec	0.84

(b) PR AUC

Rank	Repr.	PR AUC
1	PaECTER	0.65
2	GTE	0.64
3	OpenAI	0.62
4	S-BERT	0.52
5	TF-IDF	0.44
6	USE	0.36
7	Class	0.21
8	doc2vec	0.16

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

while classification-based similarity consistently lags all NLP methods except doc2vec.

The threshold-independent metrics confirm these patterns. Table 4 shows that PaECTER, GTE, and OpenAI embeddings achieve the highest ROC AUC scores. The performance gaps are more pronounced for PR AUC, as expected for an imbalanced binary prediction problem: PaECTER leads at 0.65, followed closely by GTE (0.64) and OpenAI (0.62), with S-BERT (0.52) and TF-IDF (0.44) trailing substantially.

Implications These results carry some important lessons. All models, including the worst-performing doc2vec, technically "pass" validation by predicting interferences better than random chance. This exposes a flaw in the common practice of selecting a single model *ex ante* and validating it against ground truth: such an approach would accept TF-IDF despite its 20–40% lower PR AUC compared to top models. The substantial performance differences demonstrate that comparative evaluation across multiple candidates is essential.

Based on these results, we eliminate USE, doc2vec, and classification-based measures from further validation tasks due to poor performance. We also eliminate OpenAI embeddings: while competitive, their good-but-not-best performance does not justify including a proprietary and expensive model in subsequent analyses when open-source alternatives (PaECTER, GTE) perform as well or better.

6.2 Non-Expert Human Judgment

Our second validation task tests whether similarity measures align with general human assessments of technological relatedness. This complements the interference task in two important

ways: it uses a coarser notion of similarity (related technologies rather than identical legal claims) and tests performance on historical patent text from 1880–1920. Together with the modern interference data (2001–2014), this provides evidence about temporal robustness—a critical requirement given our empirical analysis spans 1836–2023.

Task Design We evaluate the four remaining competitive models: PaECTER, GTE, S-BERT, and TF-IDF. A key challenge is that humans without specialized training struggle with absolute judgments (Carlson and Montgomery 2017). We therefore asked research assistants to make *relative* judgments: given two pairs of patents, which pair contains more similar inventions?

To ensure feasible judgments, we sampled pairs that each model ranked at least 50 percentiles apart. For instance, if one pair ranked at the 90th percentile of similarity, the comparison pair ranked no higher than the 40th percentile. This created clear distinctions for annotators while testing whether model rankings align with human intuition.

For each patent, annotators saw two text fragments: the "improvement in" statement and the first 500 characters of claims. These segments provide the essence of each invention without overwhelming annotators with technical details. Model similarity scores were computed on these identical text segments, ensuring fair comparison.

We provided detailed instructions asking annotators to consider: (i) the general technological domain, (ii) the specific problem being solved, (iii) key solution components, and (iv) other major similarities or differences. Annotators could use online resources to understand unfamiliar terminology but were instructed not to read beyond the provided text. Four annotators each completed 100 comparisons. See Online Appendix S7 for full instructions.

Results Table 5 presents results from the regression:

$$I[\text{Sim}(1) > \text{Sim}(2)]^{Emb} = \beta_0 + \beta_1 I[\text{Choice} = 1]^{Human} + \epsilon, \tag{39}$$

where $Emb \in \{\text{PaECTER, GTE, S-BERT, TF-IDF}\}$. The coefficient β_1 measures the increase in probability that the embedding ranks pair 1 as more similar when humans choose pair 1, with higher values indicating stronger human-model agreement.

All models show statistically significant agreement with human judgments, but performance differs substantially. GTE demonstrates the strongest alignment ($\beta_1 = 0.62$), followed by S-BERT (0.54), PaECTER (0.51), and TF-IDF (0.35). The R^2 values follow the same pattern.

The relative performance ordering differs slightly from the interference task, where PaECTER led. This perhaps reflects PaECTER's fine-tuning on patent data from 1985–

Table 5: 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.

2022 (Ghosh et al. 2024), while this task uses 1880–1920 patents. GTE's strong performance across both modern (interferences) and historical (human annotations) text highlights its temporal robustness and generalizability—a particularly important property for our empirical analysis spanning nearly two centuries.

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, we do not view LLMs as substitutes for human judgment: recent research shows LLMs often fail to accurately reflect human assessments (Bisbee et al. 2024; Dominguez-Olmedo et al. 2024; Goli and Singh 2024).

The LLMs showed notable disagreement with human annotators and with each other: Claude selected GTE as best-performing (matching human annotators), while GPT-4 chose S-BERT. Both consistently ranked newer embedding models above TF-IDF, suggesting potential utility for preliminary testing before deploying human annotators. However, the disagreements underscore that LLM judgments should not replace human validation for research applications. See Appendix S8 for detailed results.

6.3 Patent Office Technology Classifications

Our third validation task uses patent classifications assigned by specialized patent examiners. This task complements the previous two: like interferences, it relies on expert judgment, but like human annotations, it focuses on coarser technological similarity. Importantly, this task

extends our validation sample back to 1850, providing the longest temporal span and testing whether models perform consistently across nearly two centuries of technological change.

Data and Task Design We use CPC assignments from the May 2023 vintage, which represents the patent office's current classification of all historical patents. We evaluate similarity at two levels of granularity: (i) eight top-level technology sections (e.g., "Human Necessities," "Physics") and (ii) 123 three-digit technology classes (e.g., "Surgery," "Optics").

For each classification level and quarter-century period from 1850 to 2023, we randomly sample 200 patents from each classification category. For every patent pair, we create indicators for common section and common class membership. We then evaluate whether similarity scores from TF-IDF, S-BERT, GTE, and PaECTER can classify pairs as belonging to the same category.[27]

An important limitation of this task is that patent classifications emphasize administrative utility—facilitating prior art searches—rather than technological similarity per se. Classifications may therefore be more accurate for recent patents (which are more relevant for current patent examination) than for historical patents. Moreover, by construction, this task considers only within-category similarity and ignores between-category similarity entirely. A model that performs well here excels at distinguishing patents within versus outside a classification boundary, but this may differ from capturing the full spectrum of technological proximity.

Results Figure 11 presents results by classification level and performance metric. TF-IDF performs uniformly worst across all specifications. The ranking among competitive models, however, differs from previous tasks.

S-BERT demonstrates notably strong performance on this task. It leads all models in predicting common top-level sections by both ROC AUC (Panel 11a) and PR AUC (Panel 11c). For three-digit classes, S-BERT leads in PR AUC (Panel 11d) and ranks a close second to PaECTER in ROC AUC (Panel 11b).

PaECTER outperforms GTE at the finer three-digit class level according to both metrics (Panels 11d and 11b), but GTE remains competitive with PaECTER at the coarser section level. This suggests PaECTER may be better at capturing fine-grained technological distinctions—consistent with its strong interference task performance—while GTE maintains robust performance across different levels of granularity.

[27]Feng (2020) uses patent classes to validate doc2vec representations. Our contribution is the systematic comparison across multiple models and classification levels.

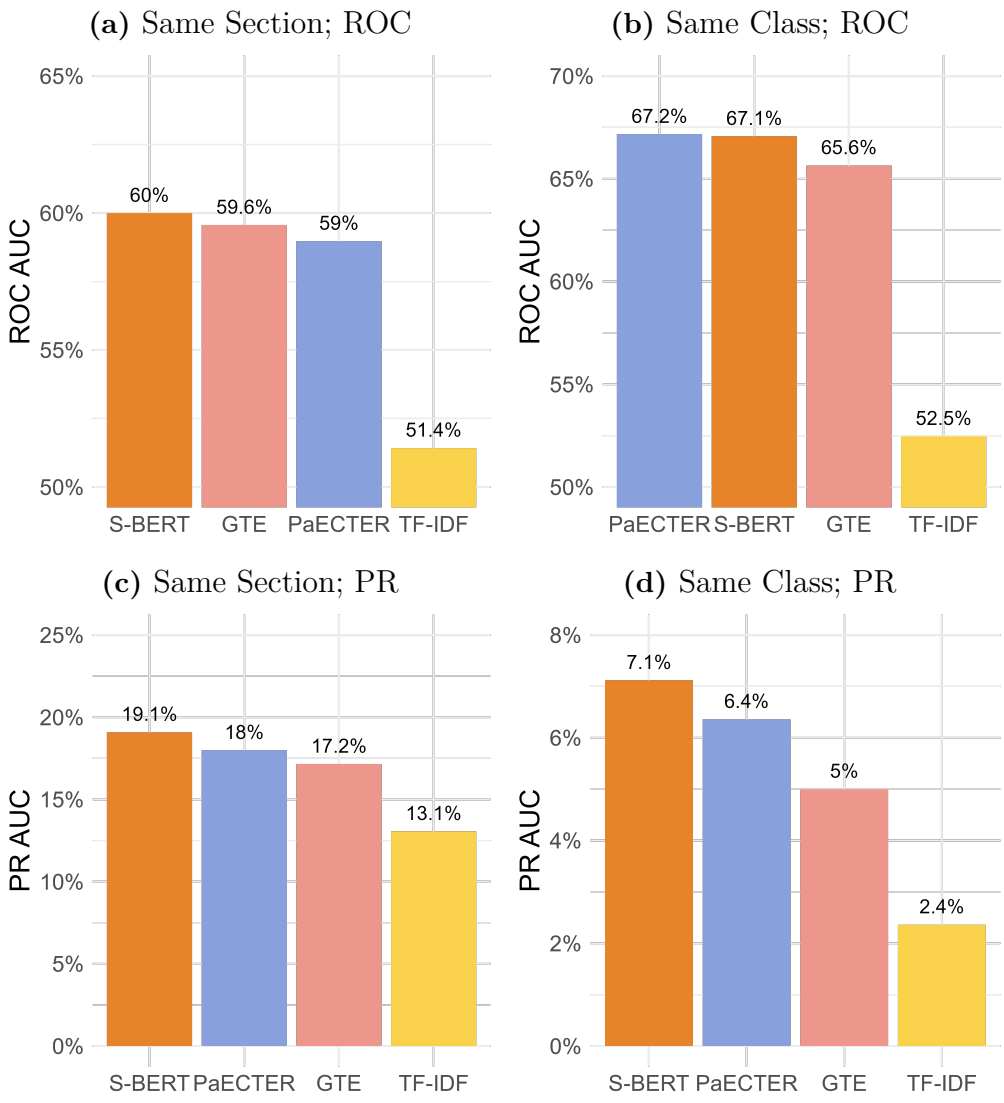


Figure 11: Representation Performance on Common Section and Class Tasks

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

Table 6: Validation Results Summary: Model Performance Across All Tasks

Model	Interferences		Human	Classifications	
	PR AUC	F10	Agreement[a]	Section[b]	Class[b]
GTE	0.64	0.90	**0.62**	0.596	0.656
	(2)	(2)	(1)	(2)	(3)
PaECTER	**0.65**	**0.90**	0.51	0.590	**0.672**
	(1)	(1)	(3)	(3)	(1)
S-BERT	0.52	0.82	0.54	**0.600**	0.671
	(3)	(3)	(2)	(1)	(2)
TF-IDF	0.44	0.77	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-digit class Note: Rankings in parentheses. Bold indicates best performance for that metric. Interference task uses 2001–2014 data, human annotations use 1880–1920 data, classifications use 1850–2023 data.

Interpretation The divergence in model rankings across validation tasks underscores the importance of multi-faceted validation. S-BERT's strong classification performance despite weaker interference and human annotation results likely reflects that patent classifications emphasize categorical boundaries rather than continuous similarity. Models optimized for such classification tasks may excel at distinguishing "inside" versus "outside" a category while performing less well at measuring proximity within the continuous space of technological ideas that our theory requires.

This pattern reinforces our validation strategy: different similarity concepts demand different validation approaches, and no single task fully captures all relevant dimensions. Section 3 demonstrates that spreading-out occurs robustly both within and between technology classes, confirming that our validated models capture continuous technological proximity rather than merely classification boundaries.

6.4 Model Selection and Validation Synthesis

Table 6 summarizes model performance across all validation tasks. Several clear patterns emerge from the comprehensive evaluation spanning 1850–2023, testing both fine-grained (legal identity) and coarse-grained (technological categories) similarity, and incorporating expert, generalist, and institutional judgments.

Consistent top performers. GTE and PaECTER demonstrate consistently strong performance across all validation tasks. PaECTER achieves the highest scores on two critical

metrics: interference detection (PR AUC of 0.65) and three-digit class prediction (ROC AUC of 0.672), suggesting superior performance at fine-grained technological distinctions. However, the performance gap with GTE on the interference task—our most demanding validation—is minimal: PaECTER's PR AUC of 0.65 versus GTE's 0.64, with identical F10 scores of 0.90. In contrast, GTE substantially outperforms PaECTER on human annotations of historical patents (1880–1920), with an agreement coefficient of 0.62 versus 0.51. This 22% performance advantage on historical text is particularly important for our empirical analysis spanning 1836–2023, as it demonstrates GTE's ability to capture technological similarity across dramatic shifts in patent language and terminology.

Task-specific variation. S-BERT performs well on classification tasks but trails on interferences and human judgments, suggesting it captures categorical boundaries better than continuous similarity. This highlights why comparative validation across multiple tasks is essential: relying solely on classification validation would overestimate S-BERT's suitability for our analysis.

TF-IDF's consistent weakness. Despite being a widely-used workhorse in economics (Kelly et al. 2021), TF-IDF consistently ranks last or near-last across all validation tasks. The performance gaps are substantial: 20–40% lower PR AUC in interferences, 40% lower agreement coefficients in human annotations, and uniformly worst classification performance. These differences are economically significant and would lead to systematically different conclusions about invention similarity trends.

In Online Appendix S9, we explore differences between TF-IDF and newer deep learning models. We find that TF-IDF tends to overweight period-specific language, leading it to assign low similarity to pairs that might describe related ideas using different words.

Model selection. For our main empirical analysis spanning 1836–2023, we select GTE as our primary representation. This choice reflects three key considerations. First, GTE demonstrates exceptional temporal robustness, substantially outperforming all alternatives including PaECTER on historical patent text from 1880–1920. Second, GTE performs nearly identically to the top-ranked PaECTER on our most demanding validation task (interference detection), with only a one-percentage-point difference in PR AUC and identical F10 scores. Third, GTE maintains strong performance across all other validation dimensions, ranking first or second on four of five metrics. Given our empirical application requires measuring similarity consistently across nearly two centuries of evolving patent language, GTE's demonstrated historical robustness makes it the preferred choice despite PaECTER's marginally higher performance on modern patents. We use PaECTER and S-BERT as robustness checks. We explicitly avoid TF-IDF despite its transparency and widespread use, as our validation demonstrates it would produce misleading results.

Critically, all models—including TF-IDF—beat random chance on every validation task, meaning they would all "pass" traditional single-model validation. This underscores our central methodological contribution: comparative evaluation across multiple candidates using domain-specific validation tasks is essential for reliable text-based measurement in economics. The choice of representation is not innocuous, and selecting poorly-performing methods can fundamentally alter economic conclusions.

7 Conclusion

How do inventors navigate an expanding landscape of technological possibilities? This paper develops a theoretical framework and provides novel empirical evidence demonstrating that inventors spread out over an expanding knowledge frontier, making inventions less similar over time. As the burden of knowledge grows and entry costs increase, the equilibrium number of inventors grows more slowly than the space of potential inventions, causing inventors to position themselves at greater distances in idea space. This spreading-out reduces knowledge spillovers between inventors, providing a new mechanism for understanding declining research productivity.

Our empirical analysis confirms this theoretical prediction using nearly two centuries of US patent data. Applying validated text representations to patent claims from 1836 to 2023, we document a substantial and consistent decline in contemporaneous invention similarity. This pattern appears robustly across different spatial scales (from global to very local), within and between technology class boundaries, corrections for multi-patent entities, and an independent data source: patent interference rates spanning 1864–2014. The convergence of evidence from these diverse approaches provides strong support for our theory's central prediction.

These findings have important implications for innovation economics and policy. The spreading-out of inventors suggests that modern inventors increasingly work on dissimilar problems, potentially reducing opportunities for the knowledge spillovers that have historically driven technological progress. While we cannot directly measure the magnitude of spillover effects, our results are consistent with both theory and empirical evidence (Bloom et al. 2013) demonstrating that technological proximity strongly affects knowledge transfer, innovation output, and firm productivity. If declining similarity indeed reflects reduced spillovers, this trend may help explain the widely-documented productivity slowdown in research and development. Policies that either promote collaboration across fields or reduce entry costs may counteract spreading-out.

Our analysis also makes a methodological contribution to the empirical study of innova-

tion and the broader use of text-as-data methods in economics. We demonstrate that representation choice fundamentally affects economic conclusions about invention similarity, with widely-used methods sometimes producing misleading results. This finding exposes a critical flaw in common practice: validating a single model by showing it beats random chance. All models we tested, including poorly-performing ones, passed this minimal standard, yet yielded dramatically different conclusions about similarity trends. Some representations suggested inventors are clustering—the opposite of what validated models reveal.

To address this challenge, we develop and implement a validation-based pipeline for constructing and selecting text-based economic measures. Our approach emphasizes three principles. First, researchers should systematically compare multiple candidate models rather than selecting one *ex ante*. Second, validation tasks should be designed specifically for the domain and research question, using expert judgments, behavioral data, or institutional classifications relevant to the concept being measured. Third, validation should span multiple dimensions—in our case, different time periods, similarity scales, and judgment sources—to ensure robustness. This framework provides a template for economists seeking to extract meaningful measures from text data while avoiding the pitfalls of unvalidated or single-validated representations.

Our findings open several avenues for future research. While our theory emphasizes the burden of knowledge, investigating the causes of spreading-out beyond the burden of knowledge—such as changes in the organization of research, funding structures, or technological opportunities—would deepen understanding of innovation dynamics. Exploring whether spreading-out varies across technological fields or firm types could reveal heterogeneity in how the burden of knowledge shapes innovation strategies.

More broadly, our approach demonstrates the value of integrating economic theory with modern computational methods for text analysis. The combination of a theoretical framework, comprehensive validation, and long-run empirical evidence provides a model for using text-as-data methods to address fundamental questions in economics. As digitized text becomes increasingly available for historical periods and diverse contexts, validation-based approaches will be essential for ensuring that the measures we construct reliably capture the economic concepts they claim to represent.

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. Finally, we received useful feedback from

Matt Clancy, Darya Davydova, Gaétan de Rassenfosse, Luise Eisfeld, Deanna James, Semyon Malamud, Roxana Mihet, participants of the seminar at EPFL, and participants of the NBER Innovation Information Initiative Technical Working Group Meeting, TADA 2023, and the NBER Summer Institute. First version: December 21, 2023.

References

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. URL: https://economics.harvard.edu/files/economics/files/kerr-william_mechanics_of_endogenous_innovation_patents_sbbi-2-3-17_0.pdf.

Arora, Ashish, Sharon Belenzon, and Lia Sheer (Mar. 2021). "Knowledge Spillovers and Corporate Investment in Scientific Research". In: *American Economic Review* 111.3, pp. 871–98. DOI: 10.1257/aer.20171742. URL: https://www.aeaweb.org/articles?id=10.1257/aer.20171742.

Arts, Sam, Bruno Cassiman, and Juan Carlos Gomez (2018). "Text Matching to Measure Patent Similarity". In: *Strategic Management Journal* 39.1, pp. 62–84. DOI: 10.1002/smj.2699.

Arts, Sam, Jianan Hou, and Juan Carlos Gomez (2021). "Natural Language Processing to Identify the Creation and Impact of New Technologies in Patent Text: Code, Data, and New Measures". In: *Research Policy* 50.2, p. 104144. DOI: 10.1016/j.respol.2020.104144.

Arts, Sam, Nicola Melluso, and Reinhilde Veugelers (Jan. 2025). "Beyond Citations: Measuring Novel Scientific Ideas and their Impact in Publication Text". In: *The Review of Economics and Statistics*, pp. 1–33. ISSN: 0034-6535. DOI: 10.1162/rest_a_01561. eprint: https://direct.mit.edu/rest/article-pdf/doi/10.1162/rest_a_01561/2500051/rest_a_01561.pdf. URL: https://doi.org/10.1162/rest_a_01561.

Ash, Elliott and Stephen Hansen (2023). "Text Algorithms in Economics". In: *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". In: *The Quarterly Journal of Economics* 132.3, pp. 1101–1164. DOI: 10.1093/qje/qjx010.

Azoulay, Pierre, Christian Fons-Rosen, and Joshua S. Graff Zivin (Aug. 2019). "Does Science Advance One Funeral at a Time?" In: *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. URL: https://doi.org/10.3386/w33821.

Berkes, Enrico and Ruben Gaetani (Sept. 2020). "The Geography of Unconventional Innovation". In: *The 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". In: *International Review of Law and Economics* 56, pp. 53–69. ISSN: 0144-8188. 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". In: *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?" In: *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". In: *Econometrica* 81.4, pp. 1347–1393. DOI: https://doi.org/10.3982/ECTA9466. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA9466. URL: https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA9466.

Bochkay, Khrystyna, Stephen V. Brown, Andrew J. Leone, and Jennifer Wu Tucker (2023). "Textual Analysis in Accounting: What's Next?" In: *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". In: *Journal of Economic Theory* 172, pp. 247–272. ISSN: 0022-0531. DOI: https://doi.org/10.1016/j.jet.2017.09.005. URL: https://www.sciencedirect.com/science/article/pii/S0022053117300972.

Butler, John P. (1993). *Patent Interference Case Files: 1838-1900*. Special List 59. 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". In: *Journal of the Patent Office Society* 64, p. 699.

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

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

Calvert, Ian A. and Michael Sofocleous (1992). "Interference Statistics for Fiscal Years 1989 to 1991". In: *Journal of the Patent & Trademark Office Society* 74, p. 822.

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

Carlson, David and Jacob M. Montgomery (2017). "A Pairwise Comparison Framework for Fast, Flexible, and Reliable Human Coding of Political Texts". In: *American Political Science Review* 111.4, pp. 835–843. DOI: 10.1017/S0003055417000302.

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

Carnehl, Christoph and Johannes Schneider (2025). "A Quest for Knowledge". In: *Econometrica* 93.2, pp. 623–659. DOI: https://doi.org/10.3982/ECTA22144. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA22144. URL: https://onlinelibrary.wiley.com/doi/abs/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 (Nov. 2018). "Universal Sentence Encoder for English". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Brussels, Belgium: Association for Computational Linguistics, pp. 169–174. DOI: 10.18653/v1/D18-2029.

Cheng, Zhaoqi, Dokyun Lee, and Prasanna Tambe (2022). *InnoVAE: Generative AI for Understanding Patents and Innovation*. Working Paper. SSRN. DOI: 10.2139/ssrn.3868599.

Chiopris, Caterina (2024). "The Diffusion of Ideas". In: *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". In: *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". In: *Management Science* 65.12, pp. 5461–5486. DOI: 10.1287/mnsc.2018.3147. eprint: https://doi.org/10.1287/mnsc.2018.3147. URL: https://doi.org/10.1287/mnsc.2018.3147.

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

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

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (June 2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.

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

Dixit, Avinash K and Joseph E Stiglitz (1977). "Monopolistic competition and optimum product diversity". In: *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 (July 2020). "The Proximity of Ideas: An Analysis of Patent Text Using Machine Learning". In: *PLOS ONE* 15.7, pp. 1–19. DOI: 10.1371/journal.pone.0234880.

Fleming, Lee (2001). "Recombinant Uncertainty in Technological Search". In: *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". In: *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". In: *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?" In: *Marketing Science* 43.4, pp. 709–722. DOI: 10.1287/mksc.2023.0306.

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.

Hall, Bronwyn H. (2009). "BUSINESS AND FINANCIAL METHOD PATENTS, INNOVATION, AND POLICY". In: *Scottish Journal of Political Economy* 56.4, pp. 443–473. DOI: https://doi.org/10.1111/j.1467-9485.2009.00493.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9485.2009.00493.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9485.2009.00493.x.

Hall, Bronwyn H., Adam Jaffe, and Manuel Trajtenberg (2005). "Market Value and Patent Citations". In: *The RAND Journal of Economics* 36.1, pp. 16–38. ISSN: 07416261. URL: http://www.jstor.org/stable/1593752 (visited on 12/17/2025).

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". In: *RAND Journal of Economics* 32.1, pp. 101–128.

Hippel, Eric von (1994). ""Sticky Information" and the Locus of Problem Solving: Implications for Innovation". In: *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". In: *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". In: *Journal of Political Economy* 129.7, pp. 1991–2022. DOI: 10.1086/714093. eprint: https://doi.org/10.1086/714093. URL: https://doi.org/10.1086/714093.

Howitt, Peter (1999). "Steady Endogenous Growth with Population and R&D Inputs Growing". In: *Journal of Political Economy* 107.4, pp. 715–730. ISSN: 00223808, 1537534X. URL: http://www.jstor.org/stable/10.1086/250076 (visited on 01/13/2026).

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". In: *The American Economic Review* 76.5, pp. 984–1001. URL: http://www.jstor.org/stable/1816464.

Jaffe, Adam B., Manuel Trajtenberg, and Rebecca Henderson (Aug. 1993). "Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations". In: *The 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?" In: *The Review of Economic Studies* 76.1, pp. 283–317. DOI: 10.1111/j.1467-937X.2008.00531.x.

Kantor, Shawn and Alexander Whalley (Mar. 2014). "Knowledge Spillovers from Research Universities: Evidence from Endowment Value Shocks". In: *The Review of Economics and Statistics* 96.1, pp. 171–188. ISSN: 0034-6535. DOI: 10.1162/REST_a_00357. eprint:

https://direct.mit.edu/rest/article-pdf/96/1/171/1974551/rest_a_00357.pdf. URL: https://doi.org/10.1162/REST_a_00357.

Kelly, Bryan, Dimitris Papanikolaou, Amit Seru, and Matt Taddy (Sept. 2021). "Measuring Technological Innovation over the Long Run". In: *American Economic Review: Insights* 3.3, pp. 303–20. DOI: 10.1257/aeri.20190499.

Kogan, Leonid, Dimitris Papanikolaou, Amit Seru, and Noah Stoffman (Mar. 2017). "Technological Innovation, Resource Allocation, and Growth*". In: *The Quarterly Journal of Economics* 132.2, pp. 665–712. ISSN: 0033-5533. DOI: 10.1093/qje/qjw040. eprint: https://academic.oup.com/qje/article-pdf/132/2/665/30637466/qjw040.pdf. URL: https://doi.org/10.1093/qje/qjw040.

Kortum, Samuel S. (1997). "Research, Patenting, and Technological Change". In: *Econometrica* 65.6, pp. 1389–1419. ISSN: 00129682, 14680262. URL: http://www.jstor.org/stable/2171741 (visited on 09/17/2025).

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". In: *Managerial and Decision Economics* 37.7, pp. 485–494. DOI: https://doi.org/10.1002/mde.2734. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mde.2734. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/mde.2734.

Le, Quoc and Tomas Mikolov (June 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. Bejing, China: 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?" In: *Fiscal Studies* 40.4, pp. 561–590. DOI: 10.1111/1475-5890.12195.

McInnes, L., J. Healy, and J. Melville (Feb. 2018). "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction". In: *ArXiv e-prints*. 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 (July 2021). *PatentsView: Disambiguating Inventors, Assignees, and Locations*. Tech. rep. Making Research Relevant. Arlington, VA: 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". In: *Political Analysis* 16.4, pp. 372–403. DOI: 10.1093/pan/mpn018.

Murata, Yasusada, Ryo Nakajima, Ryosuke Okamoto, and Ryuichi Tamura (Dec. 2014). "Localized Knowledge Spillovers and Patent Citations: A Distance-Based Approach". In: *The 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". In: *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 (n.d.). *Business Enterprise Research and Development (BERD) Survey*. URL: https://ncses.nsf.gov/surveys/business-enterprise-research-development/2023.

Olsson, Ola (2000). "Knowledge as a Set in Idea Space: An Epistemological View on Growth". In: *Journal of Economic Growth* 5, pp. 253–275. DOI: 10.1023/A:1009829601155.

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

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

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

Salop, Steven C (1979). "Monopolistic competition with outside goods". In: *The Bell Journal of Economics*, pp. 141–156.

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

Smith, Noah A. (May 2020). "Contextual Word Representations: Putting Words into Computers". In: *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". In: *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". In: *The Economic Journal* 87.346, pp. 242–261. DOI: 10.2307/2232084.

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

U.S. Patent and Trademark Office (Feb. 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". In: *Research Policy* 45.3, pp. 707–723. ISSN: 0048-7333. DOI: https://doi.org/10.1016/j.respol.2015.11.010. URL: https://www.sciencedirect.com/science/article/pii/S0048733315001857.

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. New Orleans, LA, USA: Curran Associates Inc.

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

Youn, Hyejin, Deborah Strumsky, Luis M. A. Bettencourt, and José Lobo (2015). "Invention as a combinatorial process: evidence from US patents". In: *Journal of The Royal Society Interface* 12.106, p. 20150272. DOI: 10.1098/rsif.2015.0272. eprint: https://royalsocietypublishing.org/doi/pdf/10.1098/rsif.2015.0272. URL: https://royalsocietypublishing.org/doi/abs/10.1098/rsif.2015.0272.

A Methodological Applications: Consequences of Representation Choice

Having established GTE's superiority through systematic validation (Section 5) and demonstrated its robustness across multiple dimensions of our main finding (Section 3), 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 12, 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).[28]

However, this TF-BIDF-based pattern proves highly sensitive to methodological choices (Figure 12, Panels B-D). Normalizing by total patents rather than population reveals that

[28]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.

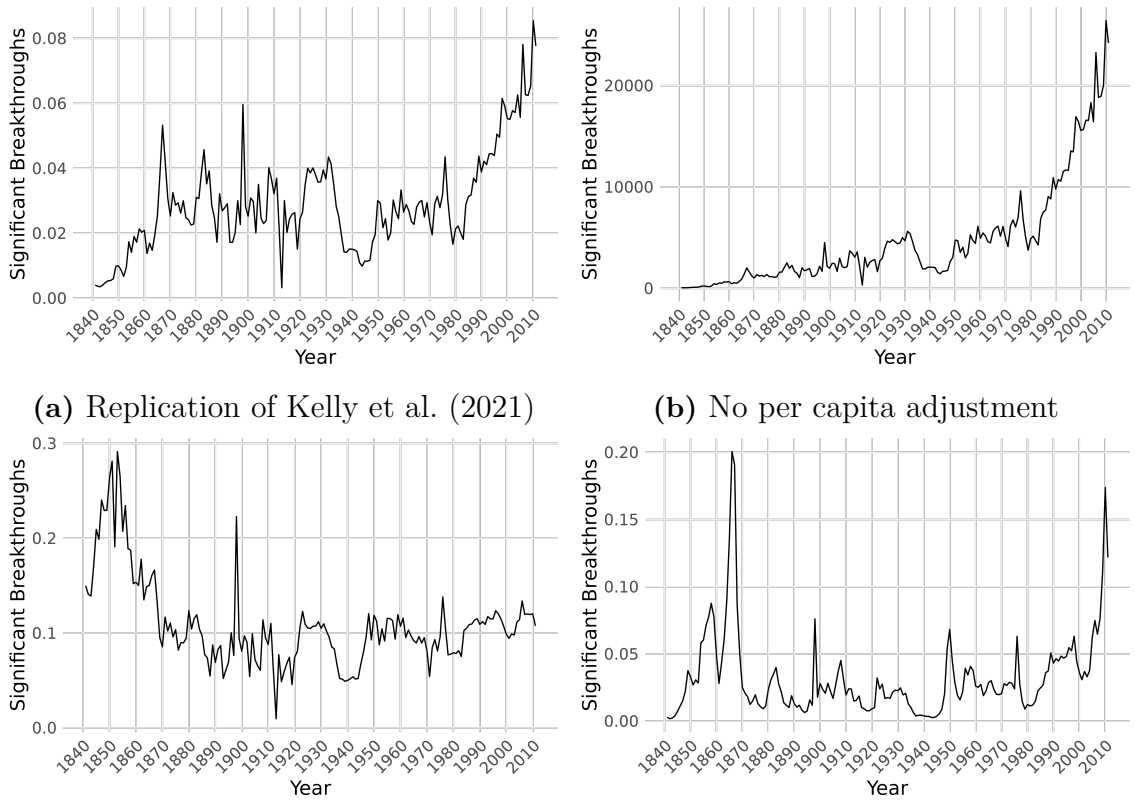
the peak breakthrough rate occurred before 1870, not recently (Panel C). Omitting year fixed effects produces a qualitatively different historical pattern with two distinct peaks (Panel D). This sensitivity raises concerns about the robustness of conclusions drawn from unvalidated representations.

Figure 13 presents the same analysis using our validated GTE representations. Several important patterns emerge. First, GTE confirms the qualitative finding of elevated breakthrough rates in recent decades, lending support to Kelly et al. (2021)'s central conclusion. Second, however, GTE reveals that the recent increase is less exceptional historically—similar booms appear in the late 19th century, the 1920s–1930s, and the 1950s–1970s. Third, and most importantly, 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 in two respects. First, GTE provides important historical context—the recent surge appears less unprecedented when earlier booms become visible. Second, 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)

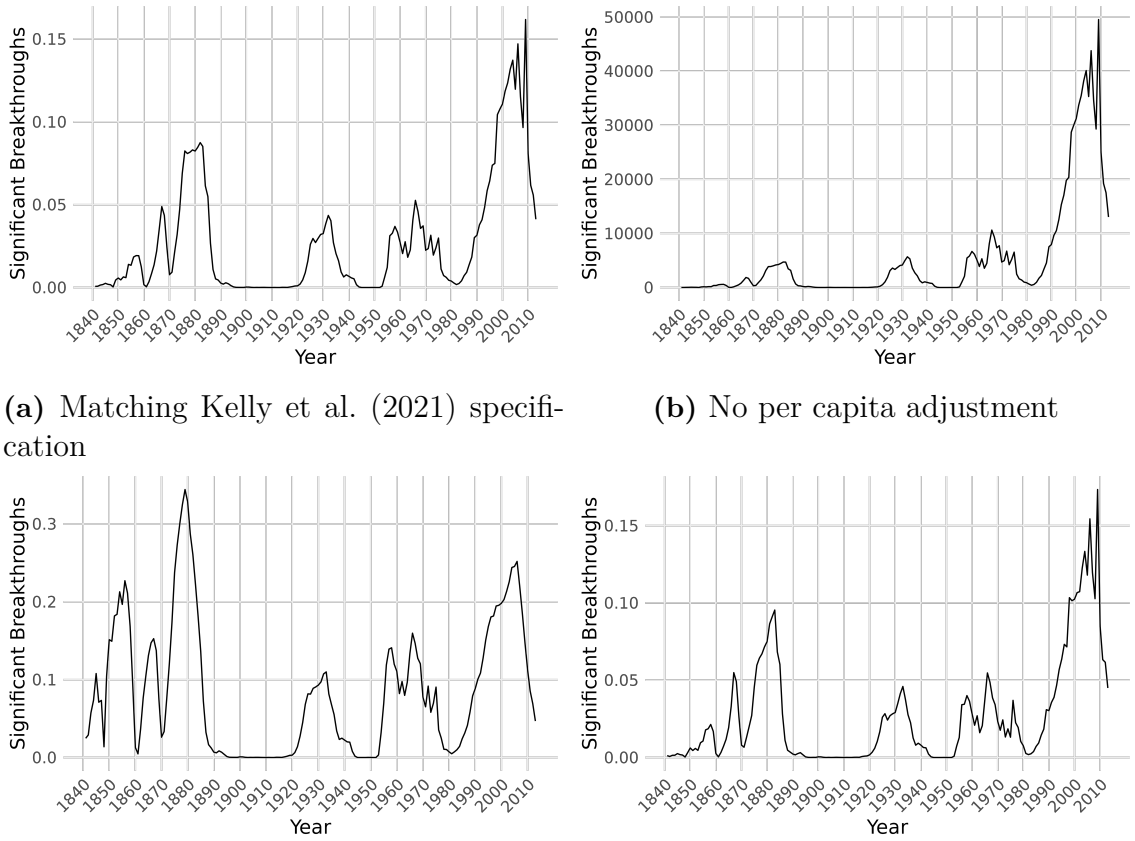
(b) No per capita adjustment

(c) Normalized by total patents

(d) No year fixed effects

Figure 12: 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 13: 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 Similarity Methods and Results

S2.1 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:

$$\text{Average Pairwise Cosine Similarity} = \frac{\left\|\sum_{i=1}^{N} \mathbf{v}_i\right\|^2 - N}{N(N-1)} \tag{S2.16}$$

Or equivalently:

$$\text{Average Pairwise Cosine Similarity} = \frac{\|\text{sum}(\mathbf{V})\|^2 - N}{N(N-1)} \tag{S2.17}$$

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:

$$\text{Avg} = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (\mathbf{v}_i \cdot \mathbf{v}_j) \tag{S2.18}$$

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

$$\left\|\sum_{i=1}^{N} \mathbf{v}_i\right\|^2 = \left(\sum_{i=1}^{N} \mathbf{v}_i\right) \cdot \left(\sum_{j=1}^{N} \mathbf{v}_j\right) \tag{S2.19}$$

Step 2: Expand the dot product:

$$\left\|\sum_{i=1}^{N} \mathbf{v}_i\right\|^2 = \sum_{i=1}^{N} \sum_{j=1}^{N} (\mathbf{v}_i \cdot \mathbf{v}_j) \tag{S2.20}$$

Step 3: Separate diagonal and off-diagonal terms:

$$\left\|\sum_{i=1}^{N} \mathbf{v}_i\right\|^2 = \sum_{i=1}^{N} (\mathbf{v}_i \cdot \mathbf{v}_i) + \sum_{i \neq j} (\mathbf{v}_i \cdot \mathbf{v}_j) \tag{S2.21}$$

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

$$\left\| \sum_{i=1}^{N} \mathbf{v}_i \right\|^2 = N + \sum_{i \neq j} (\mathbf{v}_i \cdot \mathbf{v}_j) \tag{S2.22}$$

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

$$\sum_{i \neq j} (\mathbf{v}_i \cdot \mathbf{v}_j) = 2 \times \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (\mathbf{v}_i \cdot \mathbf{v}_j) \tag{S2.23}$$

Step 6: Solve for the sum of unique pairs:

$$\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (\mathbf{v}_i \cdot \mathbf{v}_j) = \frac{\left\| \sum_{i=1}^{N} \mathbf{v}_i \right\|^2 - N}{2} \tag{S2.24}$$

Step 7: Apply the averaging factor:

$$\text{Avg} = \frac{2}{N(N-1)} \times \frac{\left\| \sum_{i=1}^{N} \mathbf{v}_i \right\|^2 - N}{2} \tag{S2.25}$$

$$= \frac{\left\| \sum_{i=1}^{N} \mathbf{v}_i \right\|^2 - N}{N(N-1)} \tag{S2.26}$$

S2.2 Multi-Patent Entitites

To address the rise in multi-patent entities in Section 3.2, 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 entity for Figure 4. Varying the random seed multiple times yielded nearly identical quantitative results.

S2.3 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 S2.5.

S2.4 Alternative Normalizations

Figure 14 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.

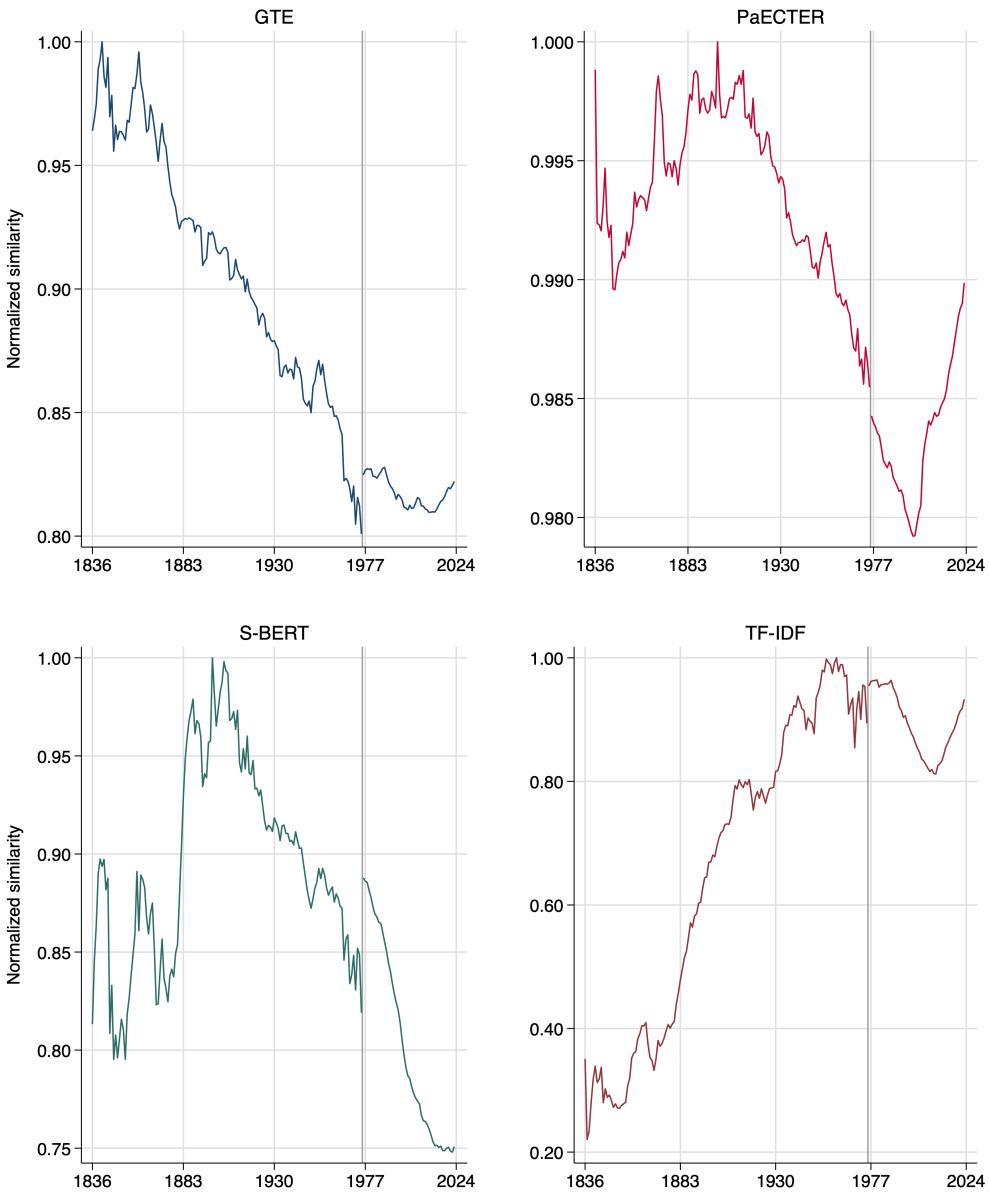


Figure 14: Similarity by Year and by Representation

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.

S2.5 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 3.3 provides some evidence for increased local clustering, especially since 2000, as indicated by the high-γ weighted similarity dynamics. Figure 15 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.

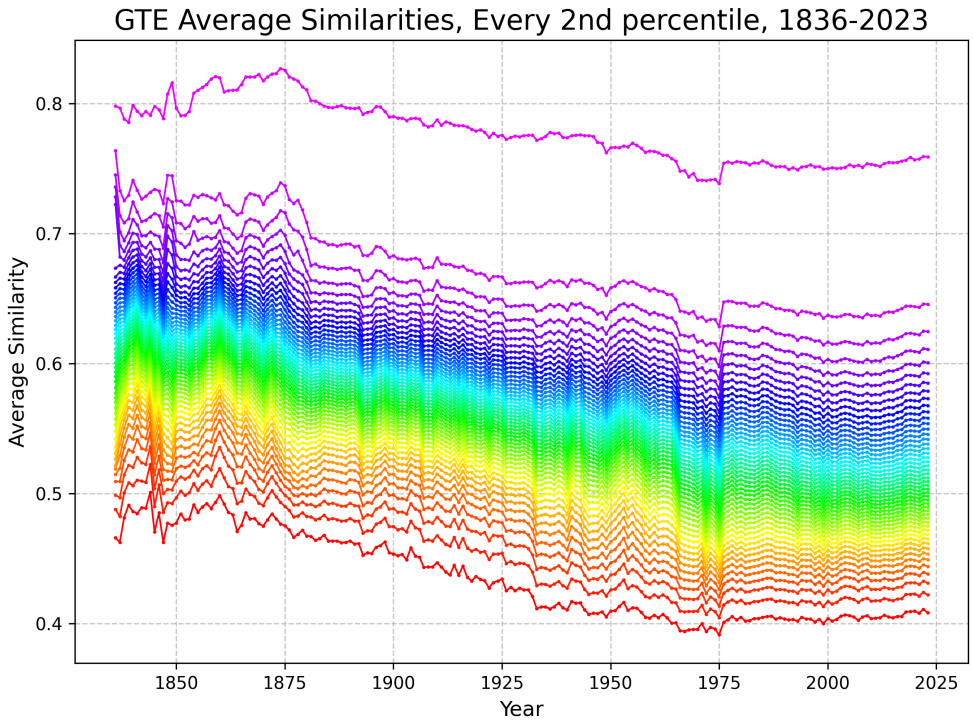


Figure 15: Similarity at Different Quantiles

This figure shows GTE similarity trends for 50 quantiles of pairwise similarity. The secular decline in similarity is robust across all quantiles.

Appendix S3 Register of Interferences

Figure S3.1 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.

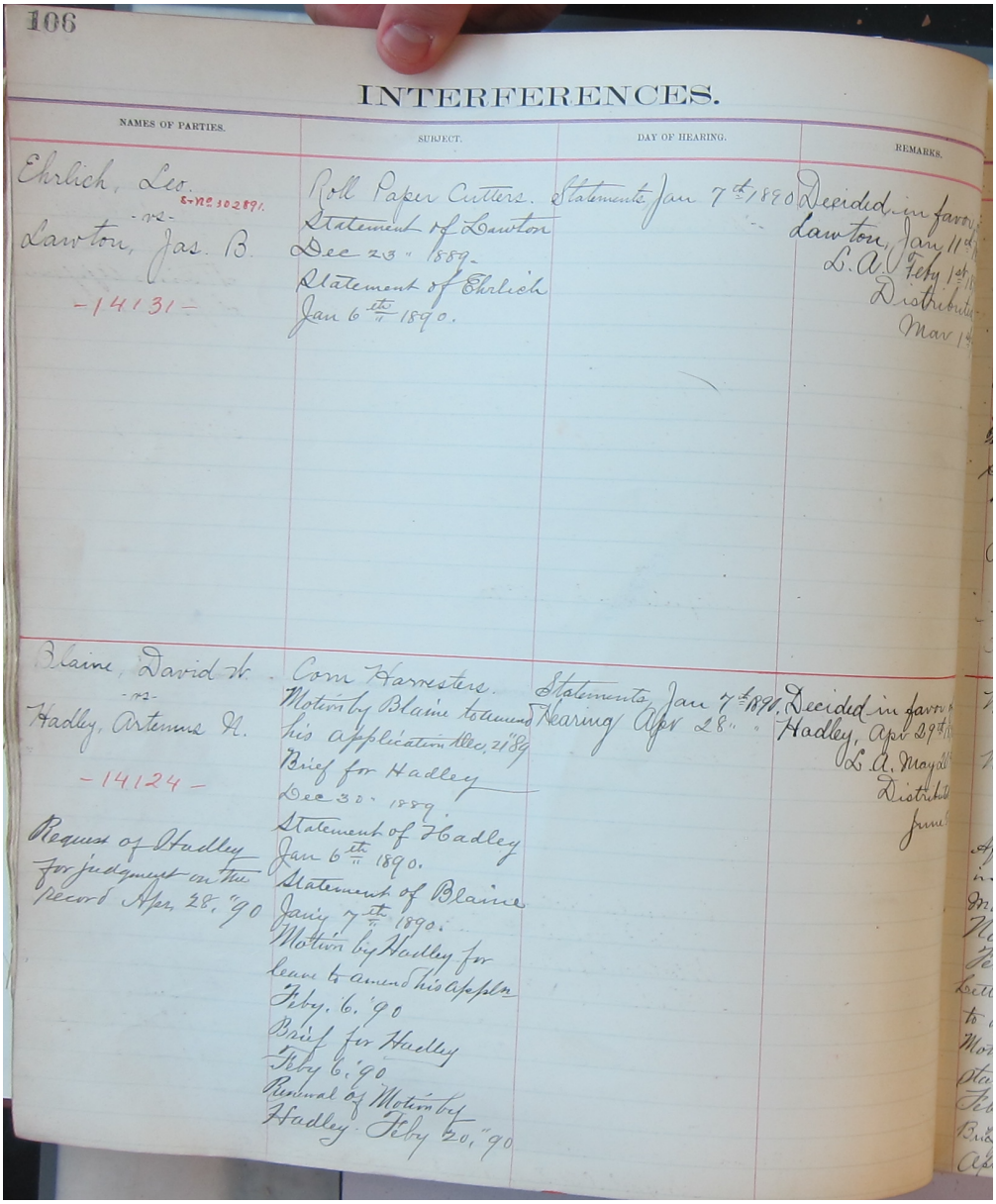


Figure S3.1: Example page from Register of Interferences

Appendix S4 Visualizing Representation Differences

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

S4.1 Methodology

The raw data are obtained using the same sampling strategy outlined in the classification validation task (6.3). 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.

S4.2 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. This jittering results in a boxier scatter plot, which is a compromise for better clarity.

The plots (refer to Figure **??**) 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 6.3.

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.

S4.3 Results and Interpretation

Figure S4.2 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.

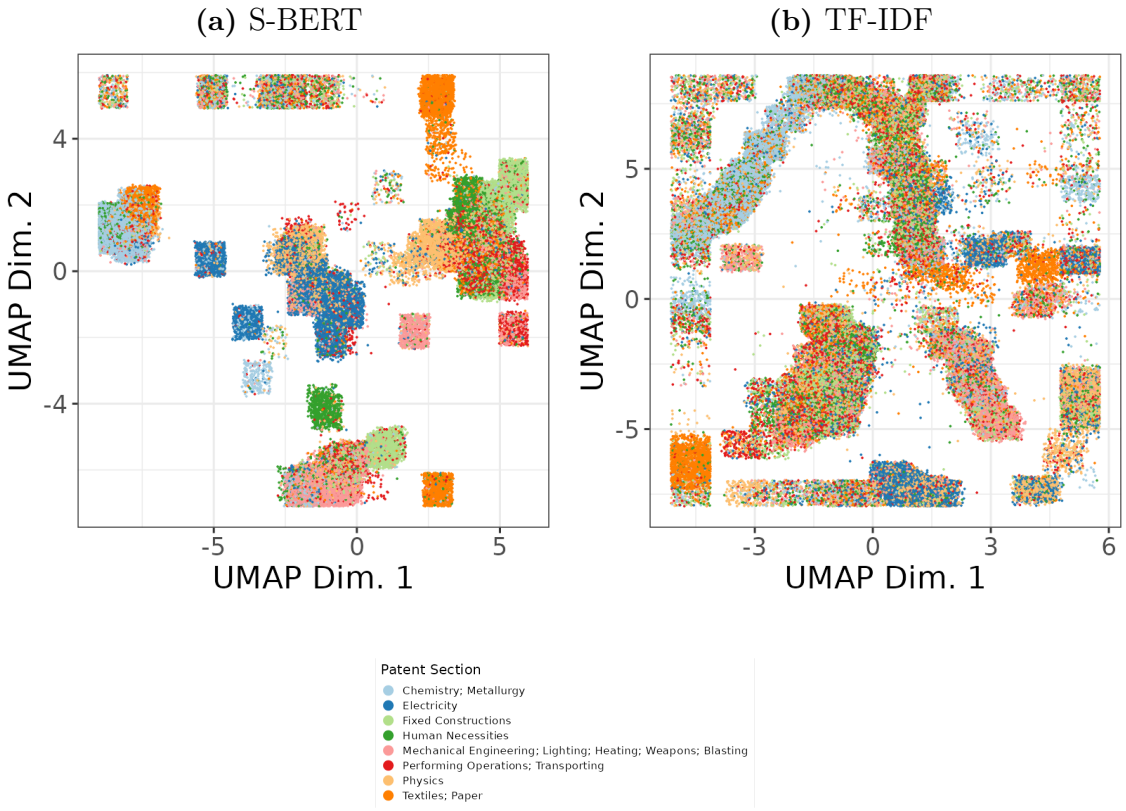


Figure S4.2: Visualizations of S-BERT and TF-IDF Representations

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 S5 Validation Framework Details

This section provides more discussion about the validation framework outlined in Section 5.

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 S5.3 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.

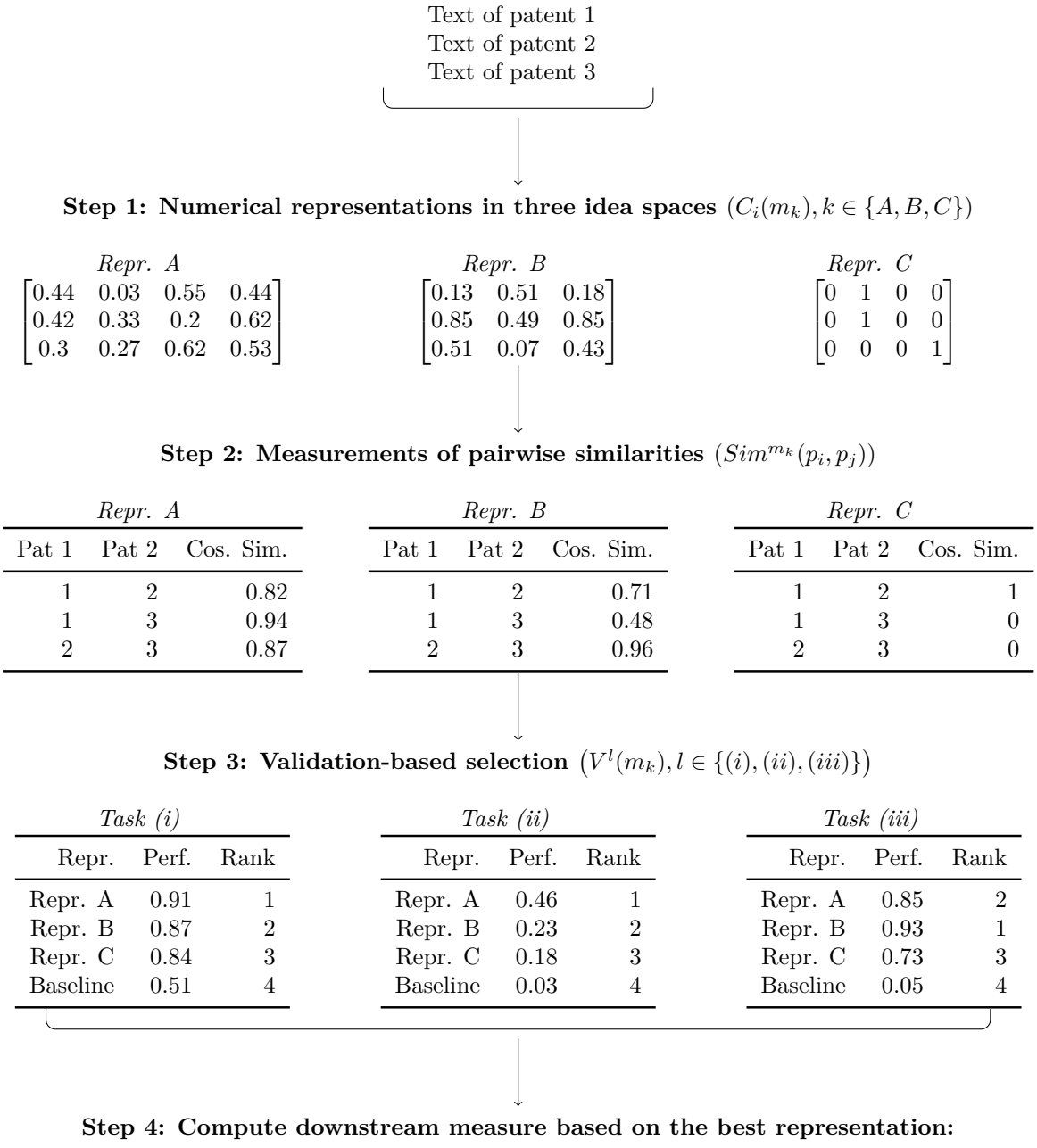
The pipeline's Steps 1 and 2 can produce different similarity measures from the same patent text—as Figure S5.3 illustrates with representations A, B, and C yielding different similarity patterns. Step 3 addresses this multiplicity by evaluating each representation using external validation:

$$V^j(m) = S^j\left(1 - d^m(\mathbf{p}), g^j(\mathbf{p})\right) \tag{S5.27}$$

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 $1 - d^m \equiv \frac{C_i^m \cdot C_j^m}{||C_i^m|| ||C_j^m||}$ computes cosine similarity 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.

his 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.



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 S5.3: 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.[29] 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.

[29]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 S6 Technical Details of NLP Models

This section provides technical details about the NLP models evaluated in Section 5.

S6.1 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. The training of doc2vec models is based on a skip-gram approach, predicting context words when given an input word. In contrast, USE and subsequent models use a two-stage training process: an initial unsupervised stage followed by supervised fine-tuning on downstream tasks. 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 S7 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 S8 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 S8.1 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.

S8.1 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.

S8.2 LLM Results

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

$$I[Sim(2) > Sim(1)]^{Emb} = \beta_0^{LLM} + \beta_1^{LLM} I[Response = 2]^{LLM} + \epsilon \tag{S8.28}$$

where LLM $\in$ {Claude, GPT} and Emb $\in$ {PaECTER, GTE, BERT, 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 S8.1. The ranking of representations differs between the two LLMs and from our human annotation results. For Claude, the ranking is GTE >S-BERT >PaECTER >TF-IDF, while for GPT-4o, it's S-BERT >GTE >PaECTER >TF-IDF. This contrasts with the human annotation ranking of GTE >S-BERT >PaECTER >TF-IDF. Despite these differences, both LLMs consistently

Table S8.1: 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.

show that newer embedding models (PaECTER, GTE, S-BERT) outperform the traditional TF-IDF approach, aligning with our human annotation findings in this crucial aspect.

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 S9 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.

S9.1 Example: Bicycle versus Velocipede

Figure S9.1 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

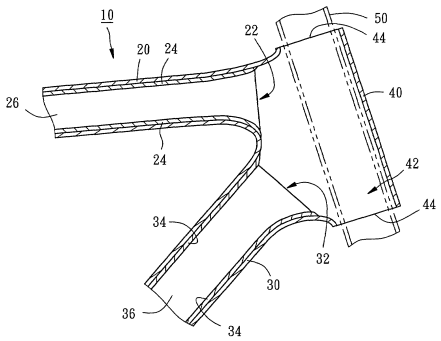
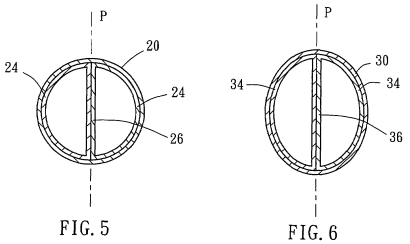
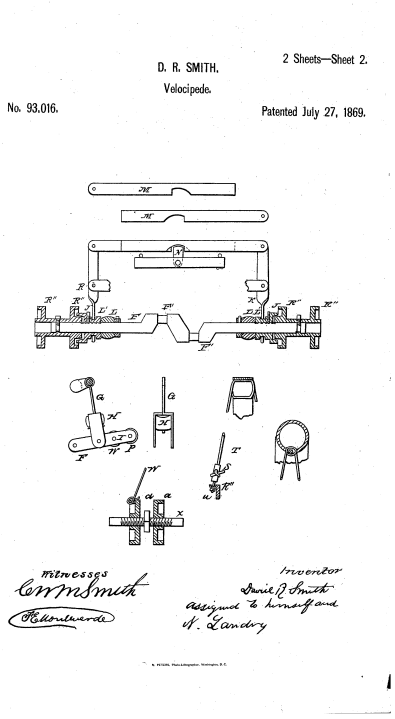
Patent 1: US7562890B2 (2009)	Patent 2: US93016A (1869)
Front frame for a bicycle. 1. A front frame for a bicycle, comprising: two first inner tubes abutted together; two second inner tubes abutted together; an upper tube of cured multiple layers of fiber reinforced rein material wound around the two 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 S9.1: 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.

S9.2 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 S9.2 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.

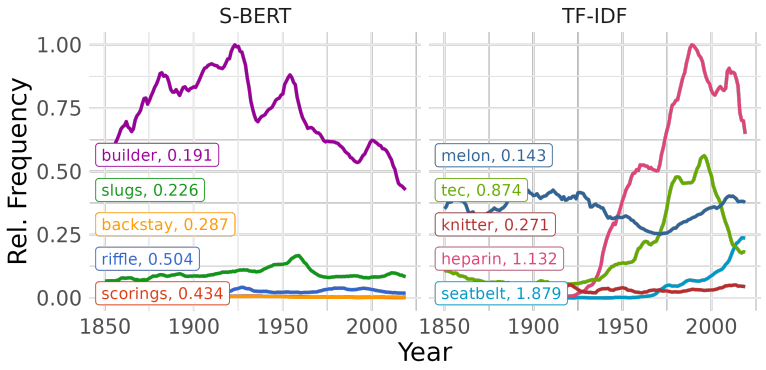
We further hand-picked examples of conceptually-similar words in panel (b). "Dresser," characteristic of S-BERT similar pairs, exhibits moderate use with little variation until the 2000s. In contrast, "vanity," characteristic of TF-IDF similar pairs, exhibits more volatility, steadily dropping in usage 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.

S9.3 Google Ngrams Analysis

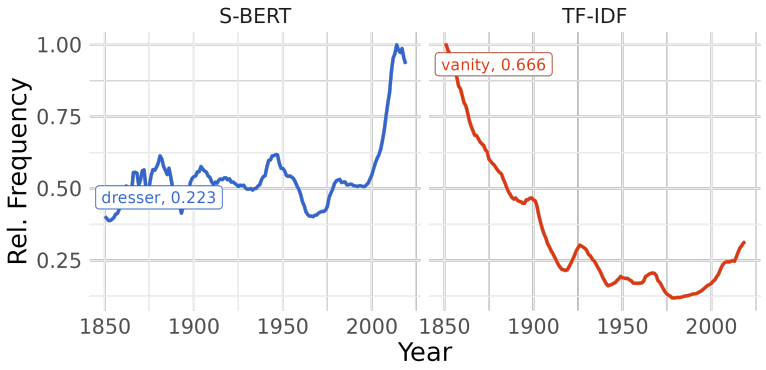
To gain insights into the time-specific nature of the words that TF-IDF focuses on, we turn to examining the tokens characteristic of patent pairs located closely in the TF-IDF space through the lens of Google Ngrams data. We identify characteristic tokens that differentiate patent pairs based on their 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

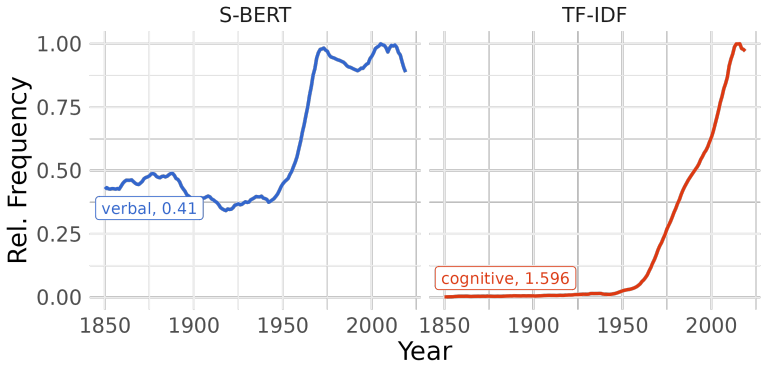


Figure S9.2: Frequency of characteristic unigrams of the pairs of patents classified as similar by S-BERT and TF-IDF

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,[30] 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.

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

Figure S9.3 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.

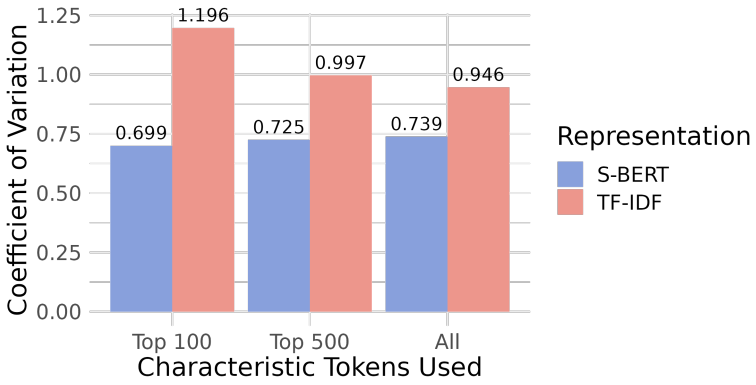


Figure S9.3: Average over-time coefficient of variation of the frequency of characteristic unigrams of the pairs of patents classified as similar by S-BERT and TF-IDF

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.

S9.4 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 two 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 S9.1: 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 S9.1 and Figure S9.4, exhibiting expected patterns. Table S9.1 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 is not surprising, given our sampling strategy's focus on patent pairs where the two models disagree). 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 S9.4 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 `same_period` task on the sub-sample used in this analysis, conforming with the full sample results discussed in Section 6.3.

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.

S9.5 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.

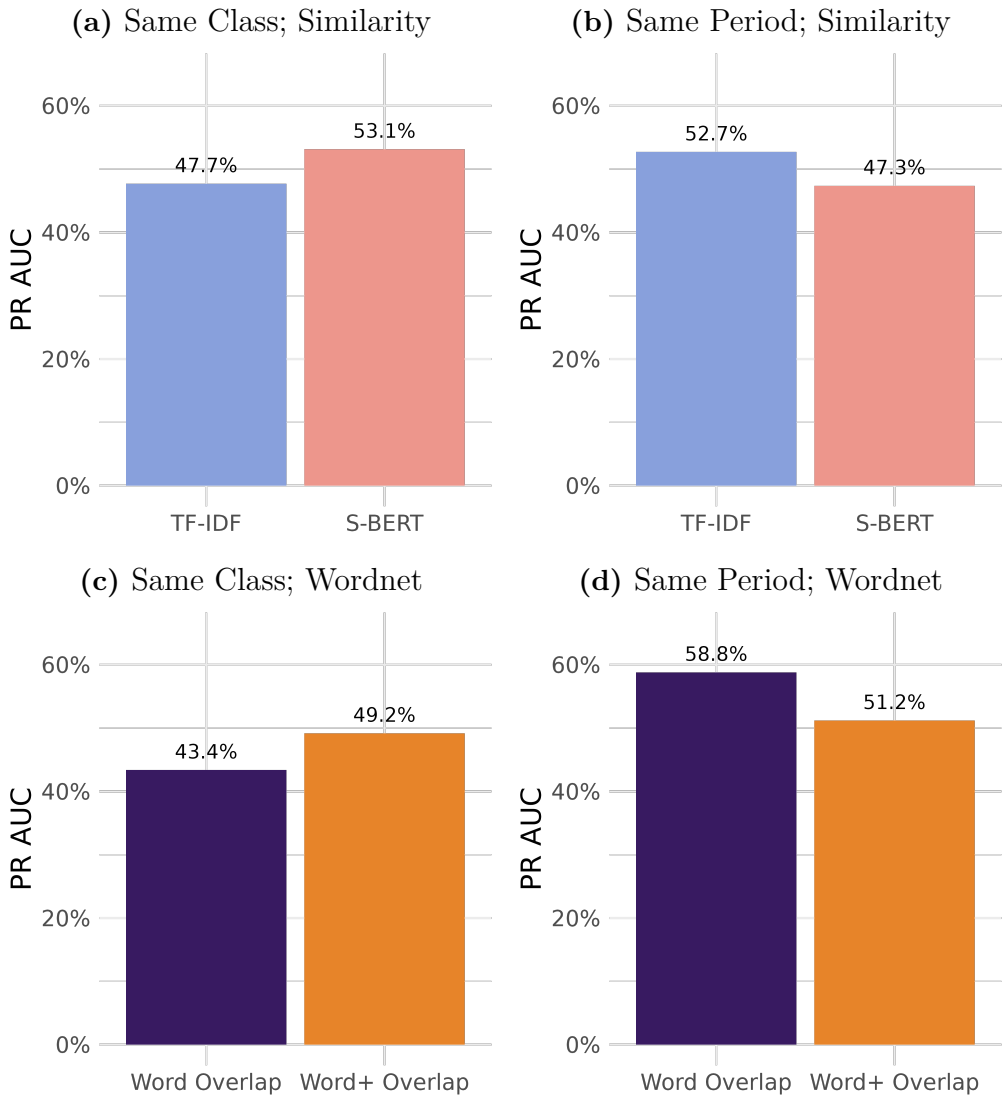


Figure S9.4: Similarity scores based on the S-BERT and TF-IDF representations and Wordnet-based measures for categorizing patent pairs as belonging to the same class and period

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 S10 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 3.3) and quantiles of average similarity (Section S2.5).

- 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.