

Expanding Idea Space and Declining Research Productivity

Ina Ganguli

UMass–Amherst

Jeffrey Lin

FRB Philadelphia

Vitaly Meursault

FRB Philadelphia

Nicholas Reynolds

U. Essex

February 2026

The views expressed in this presentation 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.

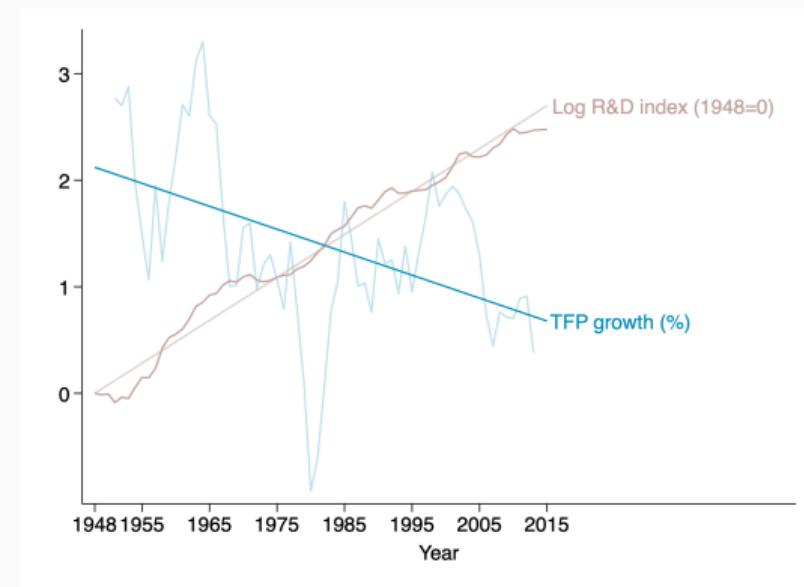
Introduction

The Puzzle: Are Ideas Getting Harder to Find?

The research productivity decline:

- Real R&D up **>20x** since 1930
- TFP growth slowed by factor of **3x**
- R&D productivity decline >-5%/yr

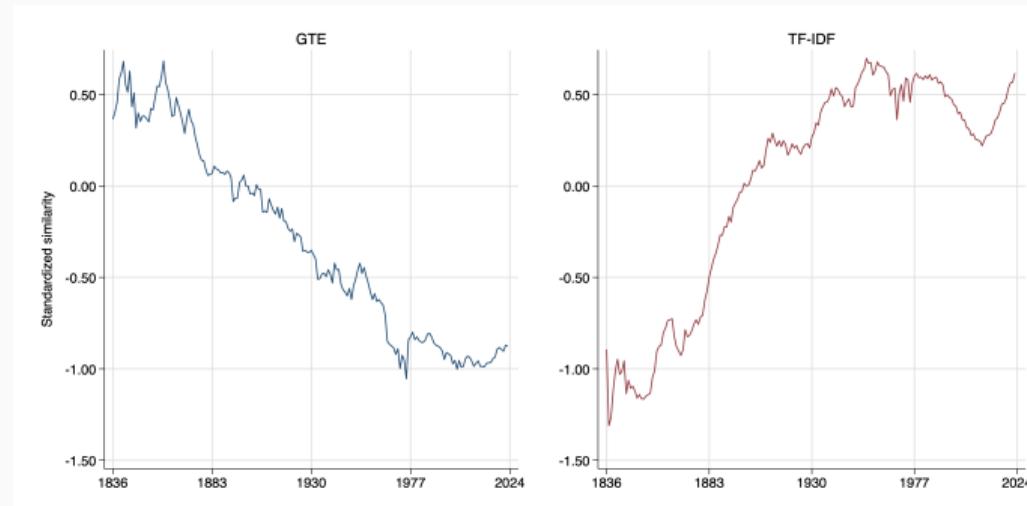
Key question: Why does it take so much **more** research effort to achieve the same rate of **slower** growth?



Bloom, Jones, Van Reenen and Webb, 2020

The Measurement Challenge

Same patent text, opposite conclusions:



- **Left (GTE):** Similarity *declining* — inventions spreading out
- **Right (TF-IDF):** Similarity *increasing* — inventions clustering

Key Question: Which “map” of idea space should we trust?

Research Questions

Q1: How do inventors position themselves across idea space?

Q2: What are the consequences for R&D productivity?

Q3: How do we measure positioning in idea space?

A1: Spatial model of positioning in idea space

- Rising “burden of knowledge” → inventors spread out to capture larger territories
- Also explains: ↑R&D intensity, ↑returns, ↑quality, ↑breadth (unifies prior findings)

A2: Spreading out over expanding idea space reduces R&D productivity

- Spillover attenuation + adaptation drag + entry & territory expansion
- **Spatial forces can explain 40–60% of R&D productivity decline**

A3: Validated measurement framework

- Systematic comparison using domain-specific tasks
- GTE embeddings outperform TF-IDF; cover 1836–2023

Related Work

Spatial Models of Innovation:

- Dasgupta & Maskin 1987; Salop 1979
- **Connect to growth**

R&D Productivity Decline: Bloom et al. 2020

- Burden of knowledge Jones 2009
- Fishing out Kortum 1997
- **Add new “spatial” explanations**

Endogenous Growth Theory:

- Howitt 1999; Peretto 1998, 2018
- **Spatially coupled quality/expansion**
- **Connect to ↓R&D productivity**

Patent Similarity Measurement:

- Jaffe et al. 1993; Kelly et al. 2021; Arts et al. 2025
- **Systematic validation**

We bridge these literatures with spatial model + validated measurement

Part I: Theory

- Spatial model of idea space
- Predictions: spreading out + declining productivity

Part II: Measurement

- Validation framework for patent similarity
- Model selection: Why GTE?

Part III: Evidence

- 188 years of spreading out
- Quantifying the spatial contribution to productivity decline

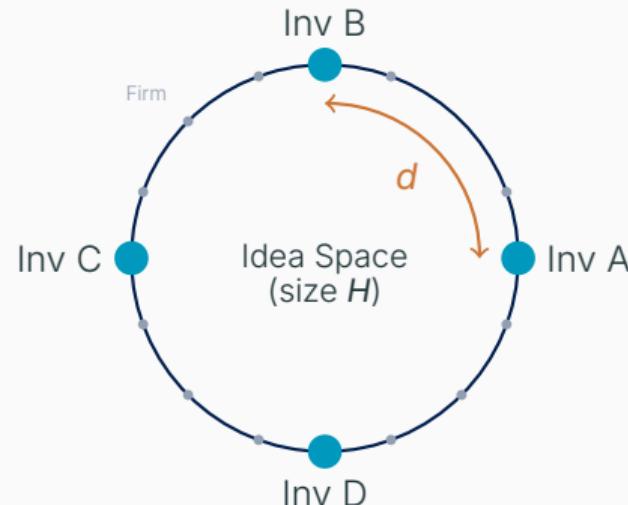
Part I: A Theory of Invention in Idea Space

Model Setup: Spatial Competition in Idea Space

Idea space: Circle of circumference H

- H = size of market for new ideas
- “Similar problems have similar solutions”
- Expands exogenously w/knowledge, demand

Inventors (idea producers):



Downstream firms (idea consumers):

Space of new ideas as Salop (1979) circle

Model Setup: Spatial Competition in Idea Space

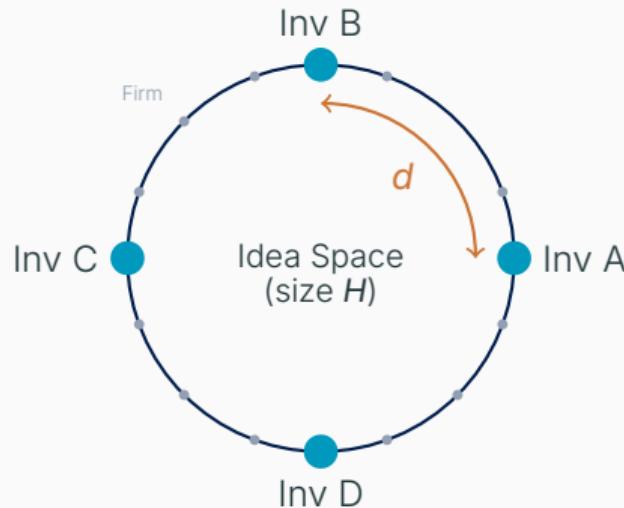
Idea space: Circle of circumference H

- H = size of market for new ideas
- "Similar problems have similar solutions"
- Expands exogenously w/knowledge, demand

Inventors (idea producers):

- Choose: **entry**, location, quality q_i , price p_i
- License non-rival ideas downstream
- "Entry" = undertaking a project (\neq firm)

Downstream firms (idea consumers):



Space of new ideas as Salop (1979) circle

Model Setup: Spatial Competition in Idea Space

Idea space: Circle of circumference H

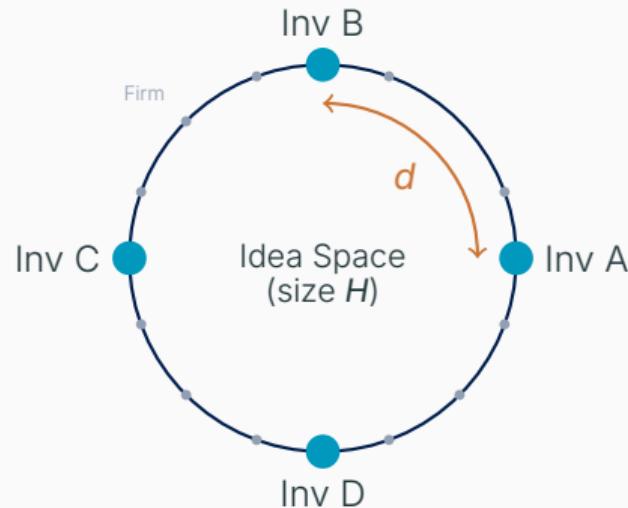
- H = size of market for new ideas
- "Similar problems have similar solutions"
- Expands exogenously w/knowledge, demand

Inventors (idea producers):

- Choose: **entry**, location, quality q_i , price p_i
- License non-rival ideas downstream
- "Entry" = undertaking a project (\neq firm)

Downstream firms (idea consumers):

- Distributed uniformly on circle
- License ideas to boost their TFP



Space of new ideas as Salop (1979) circle

Idea Consumers: Downstream Firms

Setup: Mass H of downstream firms uniformly distributed on circle

- Each firm licenses one idea to improve productivity
- Firm location = preferred technological variety

Idea Consumers: Downstream Firms

Setup: Mass H of downstream firms uniformly distributed on circle

- Each firm licenses one idea to improve productivity
- Firm location = preferred technological variety

TFP from licensing: Firm at distance h from invention i achieves **log TFP**:



$$A_i(h) = Q_i - \tau h$$

- Q_i = realized quality of invention i (including spillovers)
- τh = **adaptation cost** from technological mismatch (Bloom et al. 2013, Arora et al. 2021)

Idea Consumers: Downstream Firms

Setup: Mass H of downstream firms uniformly distributed on circle

- Each firm licenses one idea to improve productivity
- Firm location = preferred technological variety

TFP from licensing: Firm at distance h from invention i achieves **log TFP**:



$$A_i(h) = Q_i - \tau h$$

- Q_i = realized quality of invention i (including spillovers)
- τh = **adaptation cost** from technological mismatch (Bloom et al. 2013, Arora et al. 2021)

Net surplus: Firm chooses invention to maximize:

$$\text{Surplus} = \underbrace{Q_i - \tau h}_{\text{TFP gain}} - \underbrace{p_i}_{\text{license fee}}$$

- Adaptation costs create product differentiation among inventions

R&D Technology: Costs and Licensing

R&D investment: Inventor i produces idea of quality q_i at cost:

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

- $\eta > 0 \Rightarrow$ diminishing returns to R&D effort. Baseline: $\eta = 1$ (quadratic costs).
- Captures “fishing out” (Kortum 1997): harder to find new ideas

R&D Technology: Costs and Licensing

R&D investment: Inventor i produces idea of quality q_i at cost:

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

- $\eta > 0 \Rightarrow$ diminishing returns to R&D effort. Baseline: $\eta = 1$ (quadratic costs).
- Captures “fishing out” (Kortum 1997): harder to find new ideas

Non-rival licensing:

- Ideas are **non-rival**—can license to multiple firms at zero marginal cost
- Inventor charges license fee p_i to each downstream firm in territory
- Revenue = $p_i \times$ (number of firms served)

R&D Technology: Costs and Licensing

R&D investment: Inventor i produces idea of quality q_i at cost:

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

- $\eta > 0 \Rightarrow$ diminishing returns to R&D effort. Baseline: $\eta = 1$ (quadratic costs).
- Captures "fishing out" (Kortum 1997): harder to find new ideas

Non-rival licensing:

- Ideas are **non-rival**—can license to multiple firms at zero marginal cost
- Inventor charges license fee p_i to each downstream firm in territory
- Revenue = $p_i \times$ (number of firms served)

Entry cost: Fixed cost $f(H) = \phi H^\alpha$, $\alpha > 0$ captures burden of knowledge (Jones 2009)

- More education, larger teams, sophisticated equipment
- Baseline: $\alpha = 1$ (linear), but $\alpha < 2$ preserves main results. 

Knowledge Spillovers

Realized quality incorporates spillovers from neighbors:

$$Q_i = q_i + \frac{\beta}{2} \left(1 - \frac{d}{\lambda}\right) q_{i-1} + \frac{\beta}{2} \left(1 - \frac{d}{\lambda}\right) q_{i+1}$$

Parameters:

- q_i = own R&D investment
- $\beta \in (0, 1)$ = spillover intensity
- λ = spillover reach (spillovers vanish beyond distance λ)
- d = distance to nearest neighbor

Key property: Spillovers **decay with distance**

- At $d = 0$: maximum spillover βq
- At $d = \lambda$: spillovers vanish

Proximity → spillovers, but also → competition

Equilibrium Analysis

Key Mechanism: Rising Entry Costs Drive Spreading Out

As idea space expands (H grows):

1. **Burden of knowledge** rises
2. **Inventors respond by spreading out**
 - Invest more in R&D quality
 - Charge higher prices (adaptation costs create differentiation)
 - Capture larger territories (serve more downstream firms)
 - Earn sufficient revenue to cover rising entry costs

Spreading out restores zero profits as entry costs rise

Equilibrium: Pricing and Quality

Symmetric equilibrium: n inventions, equal spacing $d = H/n$, identical (p, q)

▶ Existence

Equilibrium: Pricing and Quality

Symmetric equilibrium: n inventions, equal spacing $d = H/n$, identical (p, q)

▶ Existence

Optimal pricing (standard differentiated-goods logic):

$$p^* = \tau d$$

- Price proportional to spacing
- Adaptation costs τ create pricing power through differentiation

Equilibrium: Pricing and Quality

Symmetric equilibrium: n inventions, equal spacing $d = H/n$, identical (p, q)

▶ Existence

Optimal pricing (standard differentiated-goods logic):

$$p^* = \tau d$$

- Price proportional to spacing
- Adaptation costs τ create pricing power through differentiation

Optimal quality (MR = MC for quality investment):

$$q^* = \frac{d}{\gamma}$$

- Quality proportional to spacing
- Larger territories \Rightarrow higher quality investment
- Key insight: adaptation costs make this *necessary*, not just profitable

Equilibrium: Pricing and Quality

Symmetric equilibrium: n inventions, equal spacing $d = H/n$, identical (p, q)

▶ Existence

Optimal pricing (standard differentiated-goods logic):

$$p^* = \tau d$$

- Price proportional to spacing
- Adaptation costs τ create pricing power through differentiation

Optimal quality (MR = MC for quality investment):

$$q^* = \frac{d}{\gamma}$$

- Quality proportional to spacing
- Larger territories \Rightarrow higher quality investment
- Key insight: adaptation costs make this *necessary*, not just profitable

Both price and quality rise as inventions spread out

- Validated by empirical trends [Hirshey et al. 2019](#), [Bessen et al. 2018](#), [KPSS 2017](#), [Kelly et al. 2021](#)

▶ Kelly replication

Free Entry Determines Equilibrium Spacing and Inventions

Zero-profit condition:

$$\underbrace{\tau d^2}_{\text{Revenue}} - \underbrace{\frac{d^2}{2\gamma}}_{\text{R\&D cost}} - \underbrace{\phi H}_{\text{Entry cost}} = 0$$

Solving for equilibrium spacing and number of inventions ($n = H/d$):

$$d^*(H) = \sqrt{\frac{\phi H}{\tau - \frac{1}{2\gamma}}} \Rightarrow n^* = \sqrt{\frac{1}{\phi} H \left(\tau - \frac{1}{2\gamma} \right)}$$

Key observations:

- Spacing d^* **increases** with idea space H (spreading out!)
- Requires $\tau\gamma > \frac{1}{2}$ for real solution (economic interpretation next slide)
- Number of inventions n^* **also increases** with idea space H

↑ entry costs and ↑ idea space ⇒ Spreading out and more new ideas

Prediction 1: Spreading Out

Proposition (Spreading Out)

For $\tau\gamma > \frac{1}{2}$, equilibrium spacing increases with opportunity space: $\frac{dd^*}{dH} > 0$.

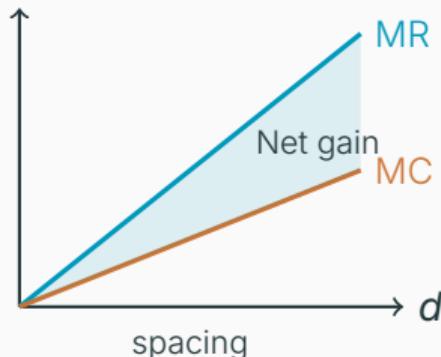
Inventions become **less similar** over time.

Proposition (Spreading Out)

For $\tau\gamma > \frac{1}{2}$, equilibrium spacing increases with opportunity space: $\frac{dd^*}{dH} > 0$.

Inventions become **less similar** over time.

Why is spreading out profitable?



Marginal revenue of expanding territory:

- Revenue $R = \tau d^2 \Rightarrow MR = 2\tau d$

Marginal cost of expanding territory:

- Need higher quality: $q = d/\gamma$
- $MC = d/\gamma$

Spreading profitable when:

$$MR > MC \Rightarrow$$

$$\tau\gamma > \frac{1}{2}$$

Adaptation costs must create sufficient pricing power

Prediction 2: Expanding Idea Space and Declining R&D Productivity

Declining R&D Productivity

Define aggregate R&D productivity (cf. Bloom et al. 2020)

$$\Pi \equiv \frac{\text{Agg TFP growth}}{\text{Agg R&D}}$$

Key insight: As idea space H expands, entry *dilutes* aggregate productivity

$$\text{Agg TFP growth} = q \left[1 + \beta \left(1 - \frac{d}{\lambda} \right) \right] - \frac{\tau d}{4}$$

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

- **Average** Δ TFP delivered downstream
- Doesn't scale with n
- **Total** R&D across n inventions
- Scales with n

Literature

Five Forces Reduce R&D Productivity

As idea space expands, five forces reduce R&D productivity:

$$\text{Agg TFP growth} = q \left[1 + \beta \left(1 - \frac{d}{\lambda} \right) \right] - \frac{\tau d}{4}$$

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

Forces reducing TFP:

1. Spillover attenuation

Knowledge flows weaken with distance

Forces raising R&D:

Five Forces Reduce R&D Productivity

As idea space expands, five forces reduce R&D productivity:

$$\text{Agg TFP growth} = q \left[1 + \beta \left(1 - \frac{d}{\lambda} \right) \right] - \frac{\tau d}{4}$$

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

Forces reducing TFP:

1. Spillover attenuation

Knowledge flows weaken with distance

Forces raising R&D:

2. Adaptation drag

Downstream firms farther from inventions

Five Forces Reduce R&D Productivity

As idea space expands, five forces reduce R&D productivity:

$$\text{Agg TFP growth} = q \left[1 + \beta \left(1 - \frac{d}{\lambda} \right) \right] - \frac{\tau d}{4}$$

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

Forces reducing TFP:

1. **Spillover attenuation**

Knowledge flows weaken with distance

2. **Adaptation drag**

Downstream firms farther from inventions

Forces raising R&D:

3. **Fishing out** Convex R&D costs

Five Forces Reduce R&D Productivity

As idea space expands, five forces reduce R&D productivity:

$$\text{Agg TFP growth} = q \left[1 + \beta \left(1 - \frac{d}{\lambda} \right) \right] - \frac{\tau d}{4}$$

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

Forces reducing TFP:

1. Spillover attenuation

Knowledge flows weaken with distance

2. Adaptation drag

Downstream firms farther from inventions

Forces raising R&D:

3. Fishing out

Convex R&D costs

4. Burden of knowledge

Rising fixed costs

Five Forces Reduce R&D Productivity

As idea space expands, five forces reduce R&D productivity:

$$\text{Agg TFP growth} = q \left[1 + \beta \left(1 - \frac{d}{\lambda} \right) \right] - \frac{\tau d}{4}$$

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

Forces reducing TFP:

1. Spillover attenuation

Knowledge flows weaken with distance

2. Adaptation drag

Downstream firms farther from inventions

Forces raising R&D:

3. Fishing out

Convex R&D costs

4. Burden of knowledge

Rising fixed costs

5. Entry and territory expansion

More inventions cover larger territories

Productivity Decline: The Decomposition

Aggregate TFP response:

$$\frac{d(\text{Agg TFP growth})}{dH} = \underbrace{\frac{dq}{dH} \left[1 + \beta \left(1 - \frac{d}{\lambda} \right) \right]}_{\text{Quality investment}} - \underbrace{\frac{\beta q}{\lambda} \frac{dd}{dH}}_{(1) \text{ Spillover attenuation}} - \underbrace{\frac{\tau}{4} \frac{dd}{dH}}_{(2) \text{ Adaptation drag}}$$

Aggregate R&D response:

$$\frac{d(\text{Agg R&D})}{dH} = \underbrace{\frac{dn}{dH} \cdot [c(q) + f(H)]}_{(5) \text{ Entry expansion}} + \underbrace{n \cdot c'(q)}_{(3) \text{ Fishing out}} \cdot \underbrace{\frac{dq}{dH}}_{(5)^*} + \underbrace{n \cdot f'(H)}_{(4) \text{ Burden of knowledge}}$$

- (5)* = territory expansion per invention
- Terms (1)–(5) correspond to five forces on previous slide

Growth Implications

Constant Growth in Fundamentals

Key question: If idea space H grows over time, what are the testable predictions?

Setup: Size of idea space expands at constant rate g_H : $\dot{H} = g_H \cdot H$

cf. Jones 1995

Fundamental variables grow at constant rates: ▶ With general cost curves

- Spacing: $d^* = C \cdot \sqrt{H} \Rightarrow g_d = \frac{1}{2}g_H$
- Quality: $q^* = d/\gamma \Rightarrow g_q = g_d = \frac{1}{2}g_H$
- Entry: $n^* = H/d \Rightarrow g_n = g_H - g_d = \frac{1}{2}g_H$
- R&D: $g_{R&D} = \frac{3}{2}g_H$ (next slide)
- Note: TFP growth rate g_{TFP} declines as d and q grow (next slide)

No knife-edge: Balanced growth holds for *any* admissible (τ, γ, ϕ) (Parallels Peretto 2018)

From Theory to Empirics: TFP and R&D Growth Equations

◀ TFP

◀ R&D

TFP growth:

$$g_{TFP} = \underbrace{g_q \left(1 + \beta - \frac{\beta d}{\lambda}\right)}_{\text{Quality (with spillovers)}} - \underbrace{\frac{\beta q}{\lambda} g_d}_{\text{Spillover attenuation}} - \underbrace{\frac{\tau}{4} g_d}_{\text{Adaptation drag}}$$

R&D growth:

$$g_{R&D} = \underbrace{g_n}_{\text{Entry}} + \underbrace{\theta \cdot g_q}_{\text{Quality scaling}} + \underbrace{\theta \cdot g_q}_{\text{Fishing out}} + \underbrace{(1 - \theta) g_f}_{\text{Burden of knowledge}}$$

where θ = variable cost share; $\alpha = 1$, $\eta = 1$

We can use these growth equations to decompose R&D productivity decline

From Theory to Empirics: Literature Evidence

Prediction	Empirical Support	Our Model
Declining aggr. R&D productivity	Bloom et al. (2020)	✓
Declining micro R&D productivity	Bloom et al. (2020)	✓
Spreading in idea space	This paper, Chopris (2024)*	$d^* = C \cdot \sqrt{H}$ ✓
Rising gross returns	Bessen et al. (2018), KPSS (2017)	$R = p \cdot d = \tau d^2$ ✓
Rising breadth/breakthroughs	Kelly et al. (2021)	$q^* = d/\gamma, d^* = C \cdot \sqrt{H}$ ✓
More firms doing R&D	Hirschey et al. (2012)	$n^* = H/d^*$ ✓
More R&D per firm	Hirschey et al. (2012)	$q^* = d/\gamma$ ✓
R&D spillovers in idea space	Bloom et al. (2013)*, Kantor & Whalley (1994)*	✓

*—Causal evidence from natural experiments

Our model reconciles rising R&D returns and declining R&D productivity plus six additional empirical patterns from a single spatial mechanism

From Theory to Empirics: Summary

Key predictions \Rightarrow testable implications:

- Spreading out: $\uparrow d$ over time $\Rightarrow \downarrow \text{similarity over time}$
- R&D productivity: $\uparrow d \rightarrow \downarrow \Pi \Rightarrow \downarrow \text{similarity} \rightarrow \uparrow \text{R&D}, \downarrow \text{TFP growth}$

Decompose R&D productivity decline into spatial and non-spatial components:

- Spatial: spillover attenuation, adaptation drag, entry & territory expansion
- Non-spatial: fishing out, burden of knowledge

Part II: Measuring Similarity in Idea Space

Data: US Patent Claims, 1836–2023

Patent text corpus:

- **Historical (1836–1975):** ProQuest Patents Core (digitized full text)
- **Modern (1976–2023):** USPTO PatentsView
- Focus on **claims** — defines legal boundaries of invention

Multiple NLP representations tested:

- Traditional: TF-IDF (word frequency)
- Modern neural embeddings: GTE, PaECTER, S-BERT, Doc2vec, USE, OpenAI

Similarity measure:

- Cosine similarity between patent representations
- Average pairwise similarity by year
- Standardized by cross-sectional standard deviation



Validation Framework: Three Complementary Tasks

Task	Time Period	Granularity	Expertise	
Patent Interferences	2001–2014	Identical	USPTO examiners	
Human Judgments	1850–1975	Continuous	Lay annotators	
Classifications	1850–2023	Categorical	Expert labels	

Why multiple tasks?

- No single ground truth for “similarity”
- Different aspects: legal identity vs. technological relatedness
- Temporal robustness across 175+ years

Models performing well across all tasks are most reliable

Validation Results: Model Performance

Model	Interferences		Human Agreement	Classifications	
	PR AUC	F10		Section	Class
GTE	0.64 (2)	0.90 (1)	0.62 (1)	0.596 (2)	0.656 (3)
PaECTER	0.65 (1)	0.90 (2)	0.51 (3)	0.590 (3)	0.672 (1)
S-BERT	0.52 (3)	0.82 (3)	0.54 (2)	0.600 (1)	0.671 (2)
TF-IDF	0.45 (4)	0.77 (4)	0.35 (4)	0.514 (4)	0.525 (4)

- **GTE and PaECTER** consistently top performers
- **TF-IDF** consistently worst (20–40% lower performance)
- All beat **random chance** — but **magnitudes differ dramatically**

Model Selection: Why We Use GTE

GTE selected for main results because:

1. **Temporal robustness** — best on historical patents (1880–1920)
2. **Near-identical performance on interferences** — our most demanding test
3. **Consistent across all tasks** — ranks 1st or 2nd on 4/5 metrics

Why TF-IDF fails: 

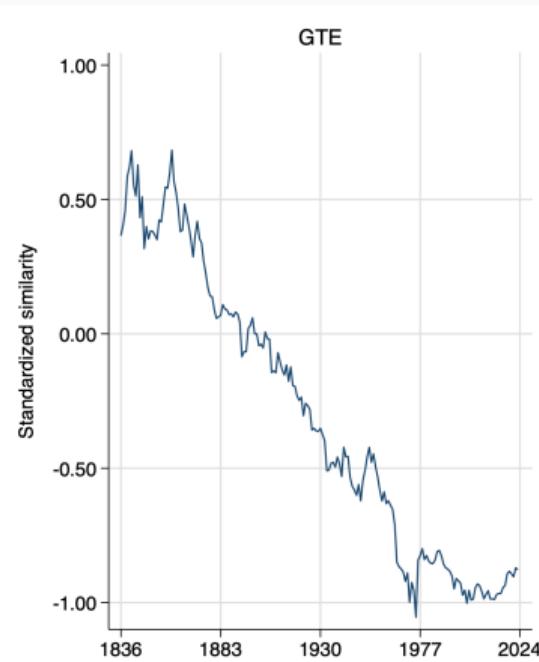
- Overweights period-specific language
- Treats synonyms as unrelated ("velocipede" ≠ "bicycle")
- Would lead to *opposite* conclusions about our theory

Robustness checks with PaECTER, S-BERT, and ensemble measures

Part III: Evidence

Prediction 1: Are Inventions Spreading Out?

Main Finding: Secular Decline in Patent Similarity



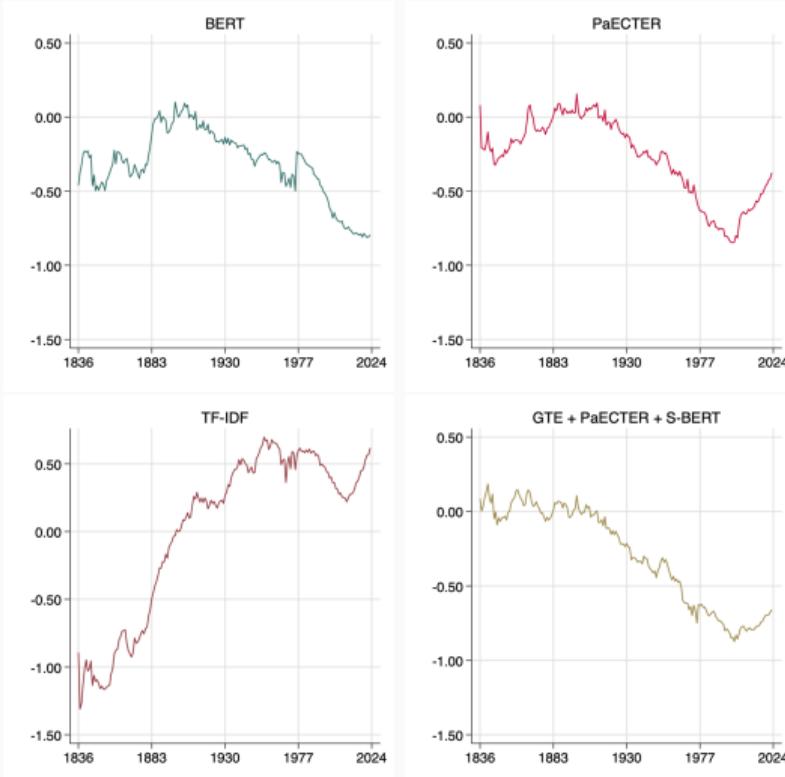
Using validated GTE embeddings:

~ 1.5σ decline in patent similarity, 1836–2023

- Consistent with theory: inventions spreading out
- Spreading out ($d \uparrow$) = Declining similarity (Sim \downarrow)
- Multi-patent entity effect post-2000 (to come)

Confirms Prediction 1: Spreading Out

Why Validation Matters: Comparing Representations



TF-IDF (worst performer):

- $\sim 1.5\sigma$ increase—opposite conclusion!
- Validation correctly discards

PaECTER, S-BERT (cf. GTE):

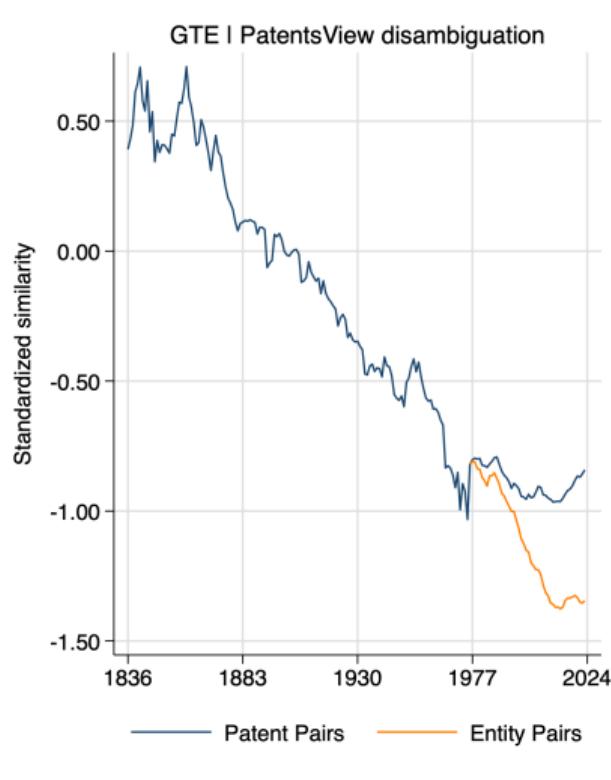
- Similar $\sim 0.8\sigma$ decline, 1880–2000
- Diverge pre-1880 & post-2000

Ensemble (avg of top models):

- $\sim 1.0\sigma$ decline, 1836–2023

**Validated methods agree; unvalidated
TF-IDF misleads**

Robustness: Accounting for Multi-Patent Entities



Concern: Post-2000 dynamics coincide with:
business method patents, non-practicing entities,
increased defensive patenting.

- **Multiple patents from same entity may be similar but not independent.**

Strategy: Sample 1 patent/entity-year

Result:

- Decline persists after correction
- **Independent inventions** still spreading out

Robustness: Spreading Out Within Technology Classes

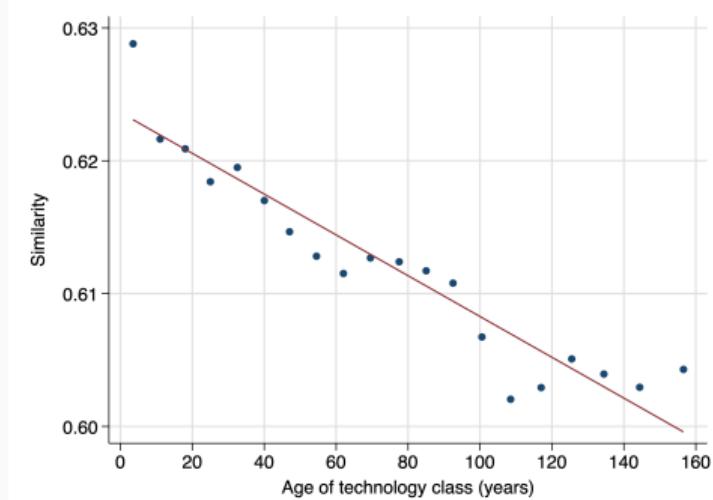
Alternative explanations: Changing patent office practice over time? Shifts across major technology areas?

Test: Within-class similarity by class "age"

- Birth = Class first issued 50 patents
- e.g., Combinatorial Chemistry 2001
- Addresses compositional concerns

▶ Between

Finding: Within-class similarity declines as classes mature



Spreading out is a dynamic process tied to field evolution

Independent Corroboration: Declining Interference Rates

Patent interferences:

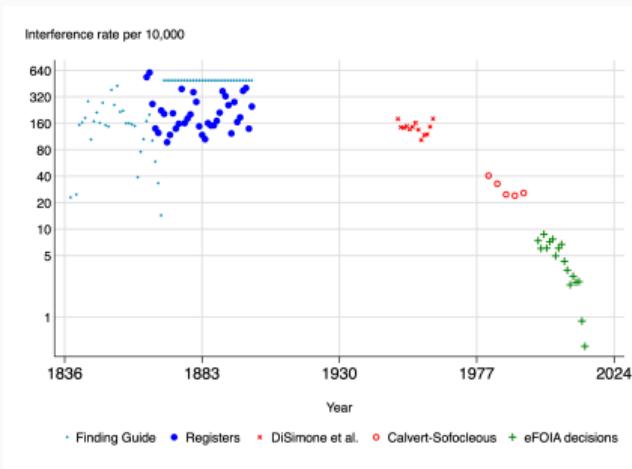
- USPTO determination that two independent inventors made *identical* inventions
- Direct measure of multiple invention ($d = 0$)

Data: Purpose - digitized from 5 sources

- Nat'l Archives files & Registers (1838–1900)

- Published statistics (1950–1994)
- eFOIA decisions (1998–2014) Ganguli et al. 2020

Finding: Interference rate declined over 150 years



**Same conclusion from
completely different data source**

Summary: Inventions Are Spreading Out

Robust evidence of spreading out:

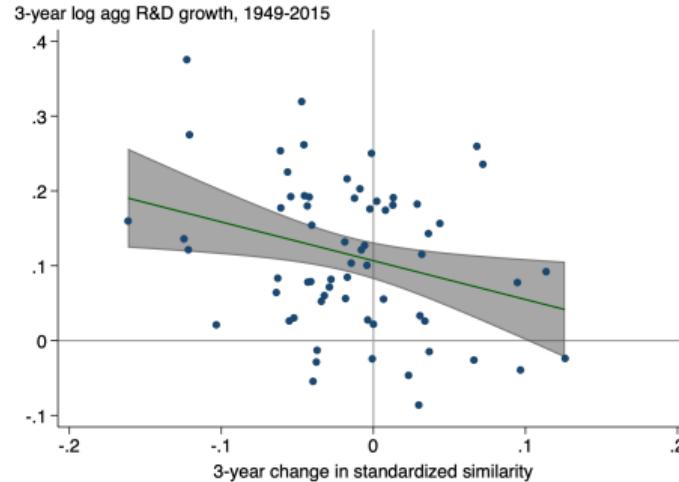
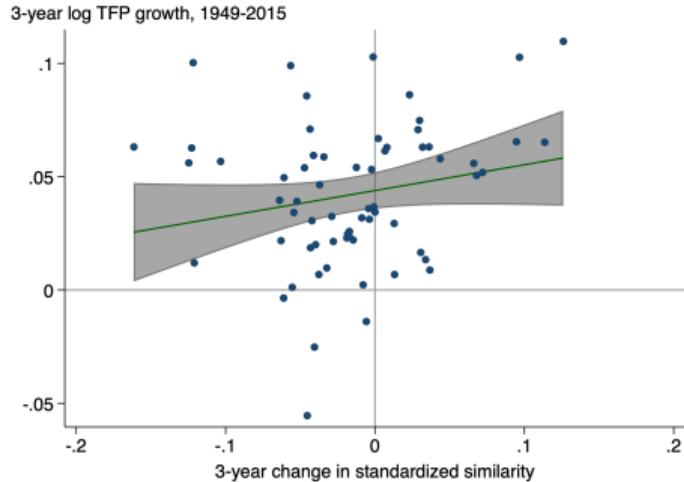
- ✓ Main finding: 1.5σ decline in similarity, 1836–2023
- ✓ Decline extends after 2000 for independent inventions
- ✓ Robust to spatial scale (local and global)
- ✓ Robust to within vs. between class decomposition
- ✓ Appears within classes as they age
- ✓ Corroborated by interference rates (150 years)



Next: What are the consequences for research productivity?

Prediction 2:
Does Spreading Out Reduce R&D Productivity?

Timing: Similarity Predicts TFP and R&D Growth



- **Left:** Declining similarity → **lower TFP growth**
- **Right:** Declining similarity → **higher R&D growth**

Both patterns confirm Prediction 2

TFP and Spreading Out

From Theory to Estimation: TFP

TFP growth equation (from BGP): 

$$g_{TFP} = \underbrace{g_q \left(1 + \beta - \frac{\beta d}{\lambda} \right)}_{\text{Quality (with spillovers)}} - \underbrace{\frac{\beta q}{\lambda} g_d}_{\text{Spillover attenuation}} - \underbrace{\frac{\tau}{4} g_d}_{\text{Adaptation drag}}$$

Substitute equilibrium relationships for unobservables:

- $q^* = d/\gamma$ and $dq^*/dt = (1/\gamma)(dd/dt) \Rightarrow g_q = g_d$

$$g_{TFP} = \underbrace{\left(1 + \beta - \frac{\tau}{4} \right) \cdot g_d}_{b_1} - \underbrace{\frac{2\beta}{\lambda} \cdot d \cdot g_d}_{b_2}$$

Suggests the regression:

- Observable proxy: $g_d \approx -\Delta \text{Sim}$ (small annual changes in standardized measure)

$$\Delta \log(\text{TFP})_t = b_0 + b_1 \cdot (-\Delta \text{Sim})_t + b_2 \cdot (-\Delta \text{Sim}) \cdot (-\text{Sim}_{t-1}) + \epsilon_t$$

From Theory to Estimation: TFP

Regression Specification:

$$\Delta \log(\text{TFP})_t = b_0 + b_1 \cdot (-\Delta \text{Sim})_t + b_2 \cdot (-\Delta \text{Sim}) \cdot (-\text{Sim}_{t-1}) + b_3 \cdot t + \epsilon_t$$

Data:

- TFP and Real R&D Inputs, 1948–2015 (Bloom et al., 2020)

Predictions and interpretation:

- $b_1 \leq 0$: Effect on TFP growth from ↑ quality scaling net of ↓ adaptation costs
- $b_2 < 0$: Spillover attenuation and reduced marginal return to R&D
- b_3 : Time trend controls for factors not explicit in the model

TFP Growth and Technological Distance

	Annual		3-Year	5-Year
$b_1 : -1 \times \Delta \text{Sim}$	−0.169*** (0.057)	−0.171*** (0.083)	−0.278*** (0.095)	−0.269*** (0.098)
$b_2 : (-1 \times \Delta \text{Sim}) \times (-1 \times \text{Sim}_{t-1})$	—	−0.015 (0.342)	−0.408 (0.320)	−0.571* (0.312)
<i>Implied TFP drag from spreading out ($\bar{\Delta \text{Sim}}$, %/yr):</i>				
1948 (Sim = 0.35)	−0.08	−0.08	−0.07	−0.04
1991 (Sim = 0, baseline)	−0.08	−0.09	−0.14	−0.16

Validation: Implied drag (−0.16%/yr) from ΔSim consistent with quasi-experimental cross-sectional elasticity ✓ Bloom et al. (2013) (−0.14 to −0.16%/yr, 1981–2001)

Contribution to TFP deceleration: Drag worsened −0.04%/yr (1948) → −0.14%/yr (2015). Change = 0.10 pp = **7% of 1.4 pp total TFP deceleration.**

► Decomposition

R&D and Spreading Out

From Theory to Estimation: R&D

R&D growth equation (from BGP): 

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

Substitute equilibrium relationships:

$$g_{R\&D} = \underbrace{[1 + \alpha(1-\theta)]g_H}_{a_0} + \underbrace{[\theta(1+\eta)-1]g_d}_{a_1}$$

Regression specification:

$$g_{R\&D,t} = a_0 + a_1 \cdot (-\Delta \text{Sim})_t + a_2 \cdot t + \epsilon_t$$

- a_2 captures (unmodeled) acceleration in idea space growth (but: $\hat{a}_2 \approx 0$)

Identification of Structural Parameters

Identification of structural parameters:

$$a_1 = \theta(1 + \eta) - 1$$



$$\theta = \frac{a_1 + 1}{1 + \eta}$$

(variable cost share)

$$a_0 = [1 + \alpha(1 - \theta)]g_H$$



$$g_H = \frac{a_0}{1 + \alpha(1 - \theta)}$$

(idea space growth)

Baseline: $\alpha = 1$, $\eta = 1$. **Later:** Calibration w/ quasi-experimental $\hat{\eta}$ and estimate of α .

Regression coefficients → structural parameters (θ, g_H)

R&D Growth and Technological Distance

	Annual	3 - Year	5 - Year
$a_1: -1 \times \Delta \text{Sim}$	0.165 (0.177)	0.448** (0.219)	0.438* (0.244)
$a_0: \text{Constant}$	0.034*** (0.006)	0.102*** (0.013)	0.173*** (0.018)
Implied θ (variable cost share)	0.58	0.72	0.72
Implied g_H (idea space growth)	2.4%/yr	2.7%/yr	2.7%/yr

Validation:

- $\theta = 72\%$ aligns with NSF survey data (labor = 69% of R&D) ✓
- $g_H = 2.7\%/\text{yr}$ consistent with patent embedding volume growth ✓ 
- BGP consistency: Model predicts $g_d/g_{R&D} = 1/3$; In data, $-\Delta \text{Sim}/g_{R&D} = 0.31$ ✓

Growth Accounting

The Research Productivity Decline

Research productivity:

$$\Pi \equiv g_{TFP}/\text{Agg R&D} \text{ (TFP growth per unit R&D)}$$

The decline (1948–2015):

- TFP growth fell: $2.1\%/\text{yr} \rightarrow 0.7\%/\text{yr}$ ($g_{g_{TFP}} = -1.6\%/\text{yr}$)
- R&D spending grew: $4.0\%/\text{yr}$

$$g_{\Pi} = g_{g_{TFP}} - g_{\text{R&D}} = -1.6\% - 4.0\% = \boxed{-5.6\%/\text{yr}}$$

Goal: Decompose this decline into spatial and non-spatial components

From Regressions to Parameters

What we estimated from R&D regression:

Parameter	Value	Source
θ (variable cost share)	0.72	R&D regression coefficient a_2 ($\eta = 1$)
g_H (idea space growth)	2.7%/yr	R&D regression constant a_0 ($\alpha = 1$)

What we assume (baseline):

Parameter	Value	Interpretation
α (entry cost curvature)	1.0	Entry costs scale linearly with H
η (R&D cost curvature)	1.0	Quadratic R&D costs

Decomposing the R&D Productivity Decline

Model implies: $g_d = \frac{\alpha}{2} g_H = 1.35\%/\text{yr}$ (spreading rate if $\alpha = 1$)

Component	Contribution	Classification	Comment
TFP deceleration	-1.6%/yr		
Spatial drag worsened	-0.11%/yr	Spatial	7% of deceleration
Unmodeled factors	-1.49%/yr	Non-spatial	

Decomposing the R&D Productivity Decline

Model implies: $g_d = \frac{\alpha}{2} g_H = 1.35\%/\text{yr}$ (spreading rate if $\alpha = 1$)

Component	Contribution	Classification	Comment
TFP deceleration	-1.6%/yr		
Spatial drag worsened	-0.11%/yr	Spatial	7% of deceleration  TFP regression
Unmodeled factors	-1.49%/yr	Non-spatial	
<i>R&D growth</i>	+4.0%/yr		
Entry expansion $(1 - \frac{\alpha}{2})g_H$	+1.35%/yr	Spatial	(new inventions)
Quality scaling $(\theta \frac{\alpha}{2} g_H)$	+0.97%/yr	Spatial	(larger territories; TFP units)
Fishing out $(\theta \eta \frac{\alpha}{2} g_H)$	+0.97%/yr	Non-spatial	(convex costs)
Burden of knowledge $(1 - \theta)(\alpha g_H)$	+0.76%/yr	Non-spatial	(rising fixed costs)
Unmodeled factors	-0.05%/yr	Non-spatial	

Decomposing the R&D Productivity Decline

Model implies: $g_d = \frac{\alpha}{2} g_H = 1.35\%/\text{yr}$ (spreading rate if $\alpha = 1$)

Component	Contribution	Classification	Comment
TFP deceleration	-1.6%/yr	Spatial	7% of deceleration ▶ TFP regression
Spatial drag worsened	-0.11%/yr		
Unmodeled factors	-1.49%/yr		
R&D growth	+4.0%/yr		
Entry expansion $(1 - \frac{\alpha}{2})g_H$	+1.35%/yr	Spatial	(new inventions)
Quality scaling $(\theta \frac{\alpha}{2} g_H)$	+0.97%/yr	Spatial	(larger territories; TFP units)
Fishing out $(\theta \eta \frac{\alpha}{2} g_H)$	+0.97%/yr	Non-spatial	(convex costs)
Burden of knowledge $(1 - \theta)(\alpha g_H)$	+0.76%/yr	Non-spatial	(rising fixed costs)
Unmodeled factors	-0.05%/yr	Non-spatial	
Total decline	-5.6%/yr		
Spatial contribution	-2.43%/yr		43%
Non-spatial contribution	-3.17%/yr		57%

Robustness: Spatial Share Increases with Better Calibration

Baseline assumptions: $\alpha = 1$, $\eta = 1$

Alternative calibration:

- $\eta = 0.625$: Guceri-Liu (2019)
- $\theta = 0.89$: From R&D regression a_2
- $\alpha = 0.76$: Constrain sum to 4.0%
- $g_H = 3.2\%/\text{yr}$: From R&D regression a_0
- $g_d = \frac{\alpha}{2} g_H = 1.2\%/\text{yr}$

Robustness: Spatial Share Increases with Better Calibration

Baseline assumptions: $\alpha = 1$, $\eta = 1$

Alternative calibration:

- $\eta = 0.625$: Guceri-Liu (2019)
- $\theta = 0.89$: From R&D regression a_2
- $\alpha = 0.76$: Constrain sum to 4.0%
- $g_H = 3.2\%/\text{yr}$: From R&D regression a_0
- $g_d = \frac{\alpha}{2} g_H = 1.2\%/\text{yr}$

Alternative decomposition:

	Baseline	Alternative
Entry expansion	1.35%/ yr	1.98%/ yr
Quality scaling	0.97%/ yr	1.08%/ yr
Fishing out	0.97%/ yr	0.67%/ yr
Burden of knowledge	0.76%/ yr	0.27%/ yr
Sum	4.05%/ yr	4.00%/ yr
Spatial share	43%	57%

Robustness: Spatial Share Increases with Better Calibration

Baseline assumptions: $\alpha = 1$, $\eta = 1$

Alternative calibration:

- $\eta = 0.625$: Guceri-Liu (2019)
- $\theta = 0.89$: From R&D regression a_2
- $\alpha = 0.76$: Constrain sum to 4.0%
- $g_H = 3.2\%/\text{yr}$: From R&D regression a_0
- $g_d = \frac{\alpha}{2} g_H = 1.2\%/\text{yr}$

Alternative decomposition:

	Baseline	Alternative
Entry expansion	1.35%/ yr	1.98%/ yr
Quality scaling	0.97%/ yr	1.08%/ yr
Fishing out	0.97%/ yr	0.67%/ yr
Burden of knowledge	0.76%/ yr	0.27%/ yr
Sum	4.05%/ yr	4.00%/ yr
Spatial share	43%	57%

Conservative baseline; higher spatial share with alternative calibration

- $\eta = 0.625 < 1$: R&D costs grow sub-quadratically with q
- $\hat{\alpha} = 0.76 < 1$: Entry costs grow sub-linearly with H
- Entry expansion (1.98%) < patent growth (3.9%) $\Rightarrow \downarrow$ ideas per patent (-1.9%/ yr)

Conclusion

Summary

1. **Theory:** Spatial model predicts:

- As idea space expands, inventions spread out
- R&D productivity declines through spillover attenuation, adaptation drag, and entry & territory expansion

2. **Measurement:** Validated NLP methods using domain-specific tasks

- Representation choice fundamentally affects conclusions
- GTE outperforms traditional workhorse TF-IDF

3. **Empirics:** Nearly 2 centuries of spreading out

- Robust across multiple tests and data sources
- **Spatial forces can explain 40–60% of R&D productivity decline**

Backup Slides

Backup: Model Equations

Fixed cost (burden of knowledge): $f(H) = \phi H, \quad \phi > 0$

R&D cost: $c(q_i) = \frac{1}{2} \gamma q_i^2$

Realized quality (with spillovers): $Q_i = q_i + \frac{1}{2} \beta \left(1 - \frac{d}{\lambda}\right) (q_{i-1} + q_{i+1})$

Equilibrium pricing and quality: $p^* = \tau d, \quad q^* = \frac{d}{\gamma}$

Equilibrium spacing: $d^*(H) = \sqrt{\frac{\phi H}{\tau - \frac{1}{2}\gamma}}$

Equilibrium entry: $n^* = \frac{H}{d^*}$

Equilibrium revenue: $R^* = p^* \cdot d^* = \tau d^2$

Backup: Robustness to Entry Cost Curvature

◀ Return

With $f(H) = \phi H^\alpha$, equilibrium spacing satisfies $d \propto H^{\alpha/2}$:

Condition	Entry Growth	Prediction
$0 < \alpha < 2$	$g_n = (1 - \frac{\alpha}{2})g_H > 0$	Entry grows ✓
$\alpha = 2$	$g_n = 0$	Entry stagnates
$\alpha > 2$	$g_n < 0$	Entry declines ✗

Main results robust for $\alpha < 2$:

- Spreading out: $g_d = \frac{\alpha}{2}g_H > 0$ for any $\alpha > 0$
- Declining R&D productivity: Holds throughout range
- Higher $\alpha \rightarrow$ faster spreading, but lower spatial share of productivity decline

Counterfactual boundary: Patent counts grow over time, ruling out $\alpha \geq 2$

Backup: Model with General Cost Curvatures

[► Growth equations](#)[◀ Decomposition](#)

Component	General	Baseline ($\alpha = 1, \eta = 1$)
Entry cost	$f(H) = \phi H^\alpha$	ϕH
R&D cost	$c(q) = \frac{1}{2} \gamma q^{1+\eta}$	$\frac{1}{2} \gamma q^2$
Spacing growth	$g_d = \frac{\alpha}{2} g_H$	$\frac{1}{2} g_H$
Quality growth	$g_q = g_d$	$\frac{1}{2} g_H$
Entry growth	$g_n = (1 - \frac{\alpha}{2}) g_H$	$\frac{1}{2} g_H$
Fixed cost growth	$g_f = \alpha g_H$	g_H

R&D growth equation:

$$g_{R\&D} = \underbrace{(1 - \frac{\alpha}{2}) g_H}_{\text{Entry}} + \underbrace{\frac{\theta \alpha}{2} g_H}_{\text{Quality scaling}} + \underbrace{\frac{\theta \eta \alpha}{2} g_H}_{\text{Fishing out}} + \underbrace{(1 - \theta) \alpha g_H}_{\text{Burden of knowledge}}$$

Log TFP Specification

[◀ Return](#)

Why linear in log TFP? $A_i(h) = Q_i - \tau h$

Standard in spatial competition (Salop 1979):

- Idea consumers have preferences linear in quality net of distance costs
- $A_i(h)$ interpreted as log TFP \Rightarrow firms care about *proportional* productivity gains

Microfoundation: Each downstream firm has one unit of fixed input ℓ and produces:

$$y = e^A \cdot \ell$$

With output price = 1 and $\ell = 1$, profit is $\pi = e^A$. Willingness to pay for technology delivering incremental log TFP A (relative to baseline $e^0 = 1$):

$$WTP = e^A - 1 \approx A \quad (\text{first-order Taylor approximation})$$

Accuracy: For annual TFP increments ($A \approx 0.015/\text{year}$), approximation error < 0.01%

Advantage: Predictions directly comparable to empirical TFP elasticities (Bloom et al. 2013) and growth accounting (Bloom et al. 2020)

Spreading-out condition:

$$\tau\gamma > \frac{1}{2}$$

Marginal revenue of expanding territory exceeds marginal cost.

Second-order conditions:

- Pricing: $\partial^2 R / \partial p^2 < 0$ (satisfied)
- Quality: $\partial^2 \pi / \partial q^2 = -\gamma < 0$ (satisfied)
- No spatial deviation (verified in paper)

Additional conditions:

- Spillover reach: $d < \lambda$ (spillovers active)
- Full coverage: All downstream firms adopt some technology

Backup: Comparative Statics Derivations

◀ Return

Spreading out: From zero-profit condition $d^2(\tau - \frac{1}{2\gamma}) = \phi H$:

$$\frac{dd}{dH} = \frac{\phi}{2d(\tau - \frac{1}{2\gamma})} = \frac{\phi}{dR/dd - dc/dd} > 0$$

Rising quality and prices:

$$\frac{dq}{dH} = \frac{1}{\gamma} \frac{dd}{dH} > 0, \quad \frac{dp}{dH} = \tau \frac{dd}{dH} > 0$$

Rising entry:

$$\frac{dn}{dH} = \frac{1}{d} - \frac{H}{d^2} \frac{dd}{dH} > 0 \text{ under spreading-out condition}$$

Declining productivity:

$$\frac{d\rho}{dH} < 0, \quad \frac{d\Pi}{dH} < 0$$

MR > MC Derivation

◀ Return

Question: When is spreading out profitable?

Marginal revenue of expanding spacing:

$$R = p \cdot d = \tau d \cdot d = \tau d^2 \Rightarrow \frac{dR}{dd} = 2\tau d$$

Marginal cost of expanding spacing (need higher quality to serve larger territory):

$$c(q) = \frac{1}{2}\gamma q^2, \quad q = \frac{d}{\gamma} \Rightarrow c = \frac{d^2}{2\gamma} \Rightarrow \frac{dc}{dd} = \frac{d}{\gamma}$$

Spreading out profitable when MR > MC:

$$2\tau d > \frac{d}{\gamma} \Rightarrow 2\tau > \frac{1}{\gamma} \Rightarrow \boxed{\tau\gamma > \frac{1}{2}}$$

Interpretation: Adaptation cost intensity (τ) \times R&D cost parameter (γ) must exceed $\frac{1}{2}$.
Pricing power from differentiation must outweigh quality investment costs.

Backup: Connection to Endogenous Growth Literature

Scale effects debate (1990s):

[◀ Related Work](#) [◀ Productivity](#)

- Early models: bigger economy → faster growth (Romer 1990)
- Empirically rejected: R&D workforce grew, growth rates didn't (Jones 1995)
- Resolution: hybrid models with horizontal + vertical innovation

Howitt (1999), Peretto (1998):

- Horizontal R&D expands varieties but **doesn't improve average quality**
- New entrants match (not exceed) existing productivity levels
- Entry dissipates scale effects — addressed 1990s debate
- **Our contribution:** Connect this to *declining R&D productivity* (Bloom et al. 2020)

No knife-edge assumptions:

- Endogenous spacing $d(H)$ adjusts for any admissible parameters
- Balanced growth emerges as equilibrium outcome, not assumption
- Parallels Peretto (2018) where market mechanisms achieve steady-state

Backup: Interference Validation Task

[◀ Return](#)

Patent interferences (2001–2014):

- **First to invent:** USPTO proceeding for multiple applicants w/ identical claims
- Provides ground truth for “identical” similarity
- 322 true interfering pairs among 96,580 application pairs

Economic intuition: Examiner ranks pairs by similarity, investigates above threshold

- Higher threshold → fewer false positives but miss true interferences
- Lower threshold → catch more but burden staff with unnecessary investigations

Metrics:

- F10: Weights recall 10x more than precision (missing interferences is costly)
- PR AUC: Precision-Recall area under curve across all thresholds

Key result: GTE, PaECTER, OpenAI retrieve ~90% of true interferences with 2–5x fewer false positives than TF-IDF/S-BERT

Example: Register of Interferences (1890)

◀ Return

INTERFERENCES.			
NAME OF PARTIES	REPORT	DAY OF HEARING	DECISION
Ehrlich, Leo Lawton, Jas. B. - 14131 -	Roll Paper Cutters. Statement Jan 7, 1890. Decided in favor of Lawton. May 11, 1890.		
	Statement of Lawton Dec 29, 1889.		L. A. B. Feb 15, 1890.
	Statement of Ehrlich Jan 6, 1890.		Decided in favor of Ehrlich. May 11, 1890.
Blaine, David W. Hadley, Artemus H. - 14124 -	Corn Harvesters. Statement Jan 7, 1890. Decided in favor of Hadley. April 27, 1890.		Hadley, April 27, 1890.
	Motion by Blaine to amend his application April 27, 1890.		Artemus H. May 11, 1890.
	Brief for Hadley See 82, 1890.		Decided in favor of Hadley. June 1, 1890.
	Statement of Hadley Jan 6, 1890.		
	Statement of Blaine Jan 7, 1890.		
	Motion by Hadley for leave to amend his application April 28, 1890.		
	Statement of Blaine Jan 7, 1890.		
	Motion by Hadley for leave to amend his application April 6, 1890.		
	Brief for Hadley April 6, 1890.		
	Motion by Blaine April 6, 1890.		
	Hadley April 6, 1890.		

Purpose - digitized from National Archives:

- USPTO Registers of Interferences, 1864–1900
- 19,388 interference cases documented
- Average 504 annual terminations

Example cases (Jan 7, 1890):

- Ehrlich v. Lawton: Roll paper cutters
- Blaine v. Hadley: Corn harvesters

Backup: Human Annotation Task

[◀ Return](#)

Historical patents (1850–1975):

- Sampled patent pairs that each model ranked at least 50 percentiles apart
- Annotators rank **relative** similarity of 2 patent pairs
- Tests temporal robustness (historical language): Oversample 1880–1920

Task: Do model rankings agree with human rankings?

- For each patent, rank others by similarity
- Compare model ranking to human ranking

Metric: Agreement coefficient from regression

$$\text{Human Rank} = \alpha + \beta \cdot \text{Model Rank} + \epsilon$$

Higher β = better agreement

Backup: Classification Validation Task

[◀ Return](#)

USPTO Classifications (1850–2023):

- CPC technology codes assigned by examiners
- Section level (8 categories) and Class level (120+ categories)
- Captures expert judgment of technological relatedness

Task: Predict whether patent pair shares classification

- Same Section (coarse): 8 top-level categories
- Same Class (fine): 3-digit classification

Metric: ROC AUC

- Area under Receiver Operating Characteristic curve
- 0.5 = random, 1.0 = perfect

TF-IDF overweights period-specific language:

- Treats “velocipede” (1880s) and “bicycle” (modern) as unrelated
- Period-specific terminology dominates similarity scores
- Creates spurious correlation with time

Example: 1880 velocipede patent

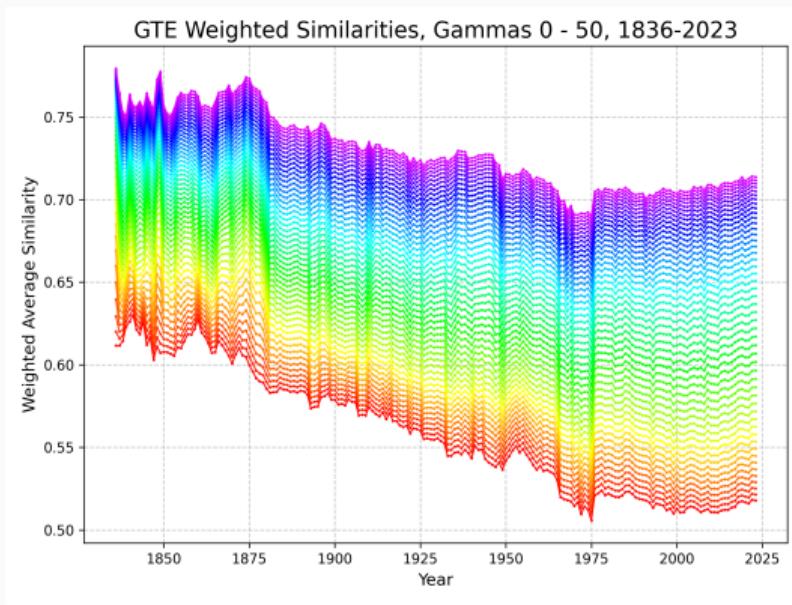
- TF-IDF: High similarity to other 1880s patents (shared vocabulary)
- GTE: High similarity to modern bicycle patents (shared concepts)

Evidence:

- TF-IDF similarity correlates with word overlap
- GTE similarity correlates with conceptual similarity
- Google Ngrams shows vocabulary shifts over time

Backup: Similarity at Different Spatial Scales

[◀ Return](#)

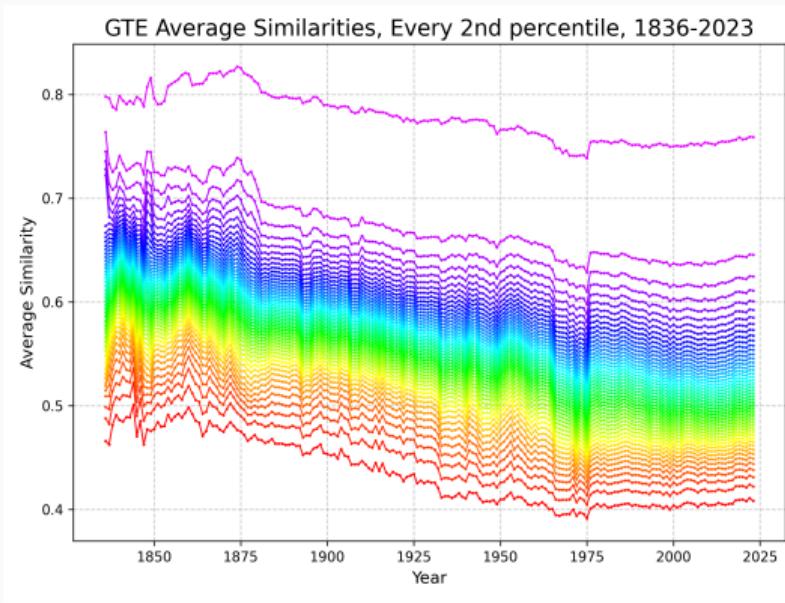


Weighted average: $\equiv \frac{1}{n} \sum_{i=1}^n \frac{\sum_{j \neq i} (1-d_{ij}) e^{-\gamma d_{ij}}}{\sum_{j \neq i} e^{-\gamma d_{ij}}}$
where γ from 0 (global) to 50 (local)

- **Key finding:** Similar declining trends across all spatial scales
- Model predictions concern averages — important to verify pattern holds across distribution
- Post-2000 arrest slightly stronger at local scales (consistent with entity correction)

Backup: Similarity at Different Quantiles

[◀ Return](#)

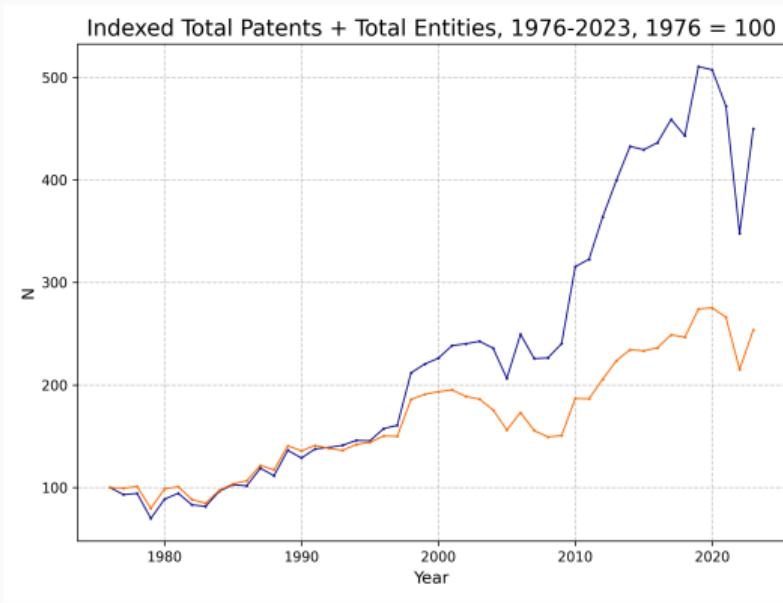


Similarity at Different Quantiles:

- 50 quantiles of pairwise similarity in each year
- Secular decline is robust across all quantiles
- Post-2000 increase in similarity is slightly faster for higher quantiles

Backup: Growth in Patents vs. Patenting Entities

[◀ Return](#)

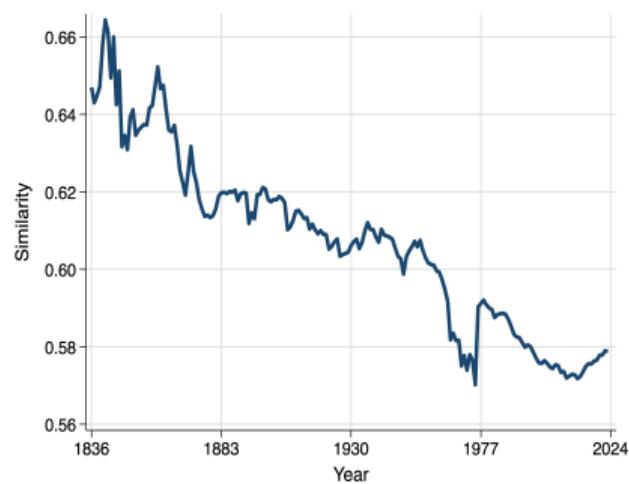


- Number of issued utility patents and unique patenting entities per year
- Divergence after 1999: substantial growth in patents per entity
- Driven by business method patents and non-practicing entities
- Motivates sampling 1 patent per entity per year for robustness

Backup: Within vs. Between Technology Classes

[◀ Return](#)

Within-class similarity



Between-class similarity



- **Addresses compositional concern:** Decline not driven by shifts across technology fields — spreading out occurs *within* established classes

Main specification: Standardize by annual cross-sectional SD

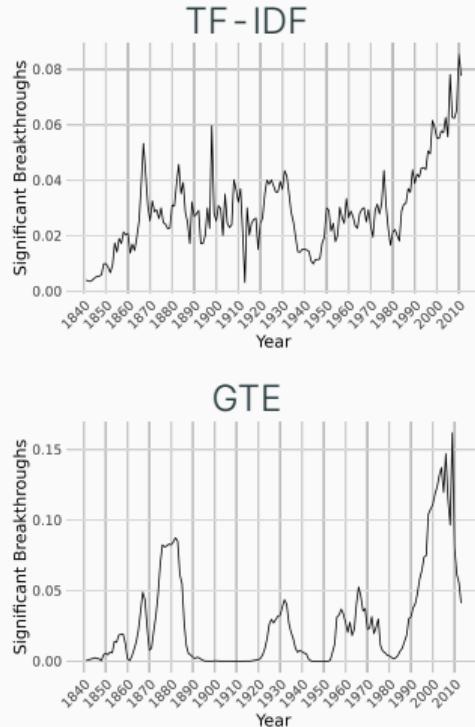
Robustness checks:

1. Time-invariant global SD → Nearly identical results
2. Raw similarity (no standardization) → Same qualitative pattern
3. Different sample sizes per year → Robust

Why standardize?

- Different representations have different scales
- No intrinsic economic interpretation of raw similarity
- SD provides meaningful units for comparison

Backup: Kelly et al. Breakthrough Replication



Kelly et al. (2021): Identify "breakthrough" patents using similarity to future patents

Our replication with GTE:

- Qualitative conclusions align (more breakthroughs today)
- Quantitative results more robust (less sensitivity to methodological choices)
- TF-IDF produces noisier breakthrough classification

Implication: Validated similarity measures improve downstream analyses

[◀ Return](#)

Historical (1836–1975): ProQuest Patents Core

- OCR-digitized patent images
- Full text of claims extracted
- Quality varies with original document condition

Modern (1976–2023): USPTO PatentsView

- Machine-readable full text
- Structured data with claim parsing
- Consistent quality

Potential discontinuity at 1976:

- Some evidence of break in levels
- Trends consistent across periods
- Results robust to excluding transition years

Backup: Computing Similarity Efficiently

[◀ Return](#)

Challenge: $O(N^2)$ pairwise comparisons infeasible for millions of patents

Solution: For unit-normalized vectors, average cosine similarity reduces to:

$$\bar{S} = \frac{1}{N(N-1)} \sum_{i \neq j} \cos(v_i, v_j) = \frac{\|\sum_i v_i\|^2 - N}{N(N-1)}$$

Complexity: $O(N \cdot d)$ where d = embedding dimension

Implementation:

1. Normalize all vectors to unit length
2. Sum vectors: $S = \sum_i v_i$
3. Compute $\|S\|^2$
4. Apply formula

Cross-sectional SD: Subsample up to 10,000 patents/year

Backup: Convex Hull

◀ Return

Log volume of convex hull (7 principal components)

