The Economic Value of Science*

Xugan Chen†
Yale

Ernest Liu‡
Princeton and NBER

Song Ma§
Yale and NBER

Abstract

How valuable is scientific research to economic growth? Scientific fields do not directly produce consumption goods; their contribution flows entirely through knowledge spillovers to technology and downstream sectors. We construct the first comprehensive innovation network that traces these flows, using the text of 463 million scientific articles and 81 million patents to measure knowledge diffusion across 209 scientific fields and 123 technology classes annually from 1870 to 2024. Unlike citation-based approaches, our text-based network reveals dense science-to-technology flows that are largely invisible in citation data. We embed this network into an endogenous growth model and show that the social value of R&D in any field equals its total discounted influence on downstream consumption—aggregating direct and indirect spillovers through the full network. Implementing the framework for the United States, we find that science accounts for 56% of the forward-looking economic value of R&D while receiving only 25% of R&D resources; its value-to-resource ratio of 2.25 (versus 0.58 for technology) persists throughout 1980–2019. Science's value is realized entirely downstream, with half of it one spillover step away, and it lands predominantly in patented technologies. Field-level values shift markedly over time: machine learning has risen sharply in recent years.

Keywords: Science, Economic Value, Innovation Network, R&D

* *This version:* July 19, 2026. All errors are our own. We thank seminar participants at Kellogg, NYU, Princeton, and Yale, and gratefully acknowledge financial support from the Alfred P. Sloan Foundation. Ernest Liu also acknowledges financial support from the NSF.

†Yale School of Management, xugan.chen@yale.edu
‡Princeton University and NBER, ernestliu@princeton.edu
§Yale School of Management and NBER, song.ma@yale.edu

1. Introduction

Modern economic growth is, at its root, science-based growth. For two centuries, economies have prospered by turning scientific discoveries into technologies—electromagnetism into electric power, molecular biology into biotechnology, machine learning into artificial-intelligence systems—and technologies into the goods they produce. This conviction has been central to economic policy and public discourse since Bush (1945) declared science "the endless frontier," supported by the economic argument that scientific knowledge is non-rival, so its social value far exceeds what its producers capture (Nelson, 1959; Arrow, 1962).

Yet while existing evidence makes a compelling case that science is valuable,[1] the fundamental questions remain open: how much does scientific research contribute to economic growth, from which fields does this contribution arise, and, at the margin, are we investing enough in science?

Answering these questions is challenging for two interrelated reasons. First, scientific fields—mathematics, chemistry, molecular biology, and the like—do not produce consumption goods; their contribution to the economy flows entirely through knowledge spillovers, often indirectly and with long lags. Second, these spillovers propagate through a complex web of cross-field linkages, so the full chain of science's contributions is difficult to observe, let alone measure. An ideal measure of the economic value of science must therefore account for the entire structure of knowledge flows—from science to technology and onward through the innovation network, both present and future.

This paper presents a new framework, both theoretical and empirical, for quantifying the economic value of science. Our approach rests on two pillars. First, we construct an empirical innovation network that captures how scientific knowledge diffuses into technology and across sectors, using millions of historical text documents from scientific publications and patents. Second, we embed this network into an endogenous growth framework that

[1] A large body of economic work provides micro-level evidence on how scientific discoveries create value. Science-driven R&D generates high social returns, both in case studies (Griliches, 1958; Mansfield, 1980) and in the aggregate (Jones and Williams, 1998; Jones, Summers et al., 2020); inventions that build on science are systematically more valuable (Fleming and Sorenson, 2004; Ahmadpoor and Jones, 2017; Poege et al., 2019; Arora, Belenzon and Sheer, 2021); and scientific funding fuels downstream innovation (Azoulay et al., 2019; Myers and Lanahan, 2022).

characterizes the economic value of R&D in every field as a closed-form function of the network structure, allowing us to quantify the spillover effects of science alongside standard growth metrics.

Innovation network. We construct the innovation network by exploiting the lead-lag structure of knowledge diffusion reflected in the textual content of scientific publications and patents. We represent each research field in a given year as a word-frequency vector over a shared vocabulary—using term-frequency-inverse-document-frequency (TFIDF) weighting to emphasize distinctive technical vocabulary—and measure cross-field proximity using cosine similarity. We infer knowledge diffusion from an upstream field j to a downstream field i when the downstream field's word-frequency vector becomes closer to the upstream field's lagged frequency vector.

A critical challenge is that raw lagged similarities may reflect persistent overlap between historically related fields rather than genuine flows of new ideas. To isolate true diffusion, we partial out the contemporaneous correlation structure among source fields. This operation has an elegant interpretation: the resulting innovation network corresponds to the coefficients of a vector autoregression (VAR) of the frequency vectors on their own lags. The VAR structure also yields an empirical advantage: it allows us to estimate each element of the network using a fixed-effect regression of each word's frequency in a downstream field on its lagged frequencies across all upstream fields. We introduce field-by-word fixed effects to absorb time-invariant vocabulary differences and field-by-time fixed effects to normalize the frequency vectors.

We implement this methodology on a comprehensive collection of data from multiple sources. We collect more than 463 million academic articles from 150,000 journals and 45,000 universities and research institutions dating back to the 16th century from OpenAlex, of which 299 million enter our analysis. We also use 81 million granted patents (representing 33 million unique inventions) from over 40 major patent authorities covering 1782–2025 from the European Patent Office's DOCDB and INPADOC databases. We complement these corpora with R&D expenditure data: expenditures from 110,000 global firms come from Worldscope; 6.3 million scientific grants come from 651 funding agencies across 43 countries over 1952–2021 from Dimensions; aggregate U.S. R&D by performer comes from the National Science Foundation's *National Patterns* series. Finally, U.S. production data from the Bureau of Economic Analysis's GDP-by-industry accounts over 1947–2024 supply the consumption weights and calibration inputs for our quantitative analysis.

The key input for constructing the innovation network is the text from the title and abstract of each publication and patent. We classify articles into 209 STEM-related scientific fields using the OpenAlex hierarchical classification and patents into 123 technology classes using 3-digit IPC codes, yielding a 332×332 innovation network for each year from 1870 to 2024. Combining this novel network with rich data on R&D inputs and production output allows us to quantify the economic value of science.

We validate the innovation network through several complementary approaches. First, we compare our text-based network with the citation-based network—combining patent-to-patent citations (Acemoglu, Akcigit and Kerr, 2016), paper-to-paper citations, and patent-to-paper citations (Marx and Fuegi, 2020, 2022). Within-domain elements of the two networks show a strong, positive relationship. Crucially, however, our text-based network captures cross-domain knowledge flows, particularly science-to-technology flows, that are largely absent from citation-based approaches—the heatmap of the citation-based network shows a near-empty science-to-technology quadrant, in sharp contrast to the rich cross-domain structure in our text-based network. Second, a Monte Carlo simulation confirms that our approach accurately recovers the true structure of a known innovation network from text data alone, with a correlation of 0.93 for the network elements and 0.99 for the derived Katz centrality measures, while a simple cosine similarity matrix fails entirely, producing a near-uniform and uninformative matrix. Third, we present event studies that trace the diffusion of distinctive keywords from focal innovation fields to downstream fields. Keywords spread first to closely connected fields and subsequently to more distant ones, in an order that aligns with the directional structure of our network.

Several stylized facts emerge from the innovation network. First, science is more central than technology in the knowledge diffusion process. Scientific fields such as artificial intelligence and biochemistry serve as dominant knowledge sources for a broad range of technology sectors. In 2000, scientific fields at the interface of computing, information, and the life sciences, including computational biology and data mining, occupied the top three positions in the combined centrality ranking of all 332 fields; seven of the ten most central fields were scientific, and the highest-ranked technology class entered only at rank four. Second, centrality dynamics differ markedly between the two domains. Tracking how the Katz centrality rankings of each domain's top 20 fields evolve over subsequent years, we find that technology classes experience relatively large short-run movements but stabilize over time. Scientific fields, by contrast, exhibit a larger long-term decline in centrality, with initially central fields more likely to be displaced by new entrants over time—consistent with the exploratory and breakthrough nature of scientific progress.

We further examine how breakthrough innovations, identified following Kelly et al. (2021), reshape the network. In both domains, Katz centrality rises sharply in the 5–10 years following a breakthrough, with the increase more pronounced and persistent in science; these results hold after controlling for the existing knowledge stock.

Endogenous growth model. Our measure of the economic value of science is guided by a multisector endogenous growth model in which a finite stock of R&D resources is allocated across real-economy sectors and scientific fields to generate new ideas. A key structural feature, noted at the outset, is that scientific fields do not directly produce consumption goods: by advancing the state of knowledge in science, researchers make R&D in downstream technology sectors more productive, which in turn raises the quality of goods available to consumers. This is precisely why standard market-based methods cannot price the value of science, and why the structure of the innovation network is indispensable for welfare accounting.

Our main theoretical result provides a closed-form characterization of the economic value of R&D in each sector. A sector's total economic value aggregates all the ways its knowledge propagates through the network over time—first to directly connected downstream sectors, then through successive rounds of spillovers to sectors further away—each round discounted at rate ρ/λ, the discount rate relative to the rate at which knowledge spillovers materialize. When the innovation network is static, this infinite sum of discounted network effects is precisely the Katz centrality of the sector in the innovation network, weighted by the vector of consumer preferences across goods. This connection provides an economic foundation for a well-known network statistic. When the network is time-varying, the economic value generalizes to a forward-looking analogue in which each round of network effects uses the future network at the appropriate horizon. The marginal economic value of investing in a sector is then proportional to the ratio of that sector's total economic value to the amount of R&D it already receives. This result nests the framework of Liu and Ma (2021) as a special case in which the innovation network is time-invariant. Because scientific fields receive zero direct weight in consumer preferences, their entire economic value is derived from these indirect chains, and the social optimum calls for directing R&D to science in proportion to the strength and reach of these chains. Finally, we extend the framework to stochastic regime changes—recurrent, randomly arriving shocks that redraw how knowledge feeds innovation; the value of R&D then takes the same form, with an expectation over the network's future path.

Quantifying this in the data requires forecasting how the innovation network evolves over time, since the present value of science depends on its future role in the diffusion of knowledge. To make this tractable, we introduce a joint diagonalization method that decomposes the high-dimensional, time-varying network into 150 stable "knowledge topics"—economically interpretable clusters of related fields, such as computer science & engineering or particle physics—with time-varying importance weights. This reduces the prediction problem from forecasting each of the more than 100,000 network elements at every future date to forecasting 150 eigenvalue time series. The decomposition reproduces the observed Katz centrality closely (a correlation above 0.9), and the eigenvalues are substantially predictable: economic fundamentals—the knowledge stocks and breakthrough innovations of all fields—deliver an out-of-sample R^2 of approximately 0.64 at short horizons, well beyond what the network's own persistence achieves.

Quantitative results. We compute the forward-looking economic value of R&D for each field and year. Our central finding is that science is a quantitatively dominant—and systematically under-resourced—source of economic value. In 2019, scientific fields accounted for 56% of the forward-looking economic value of R&D in the United States while receiving only 25% of R&D resources, a value-to-resource ratio of 2.25; technology fields created the remaining 44% of value on 75% of resources, a ratio of 0.58. Science's value share exceeds half in every snapshot over 1980–2019, its value-to-resource ratio never falls below 1.7, and the gap widens over the sample as R&D resources drift toward technology fields while value remains anchored upstream in science. Put differently, science is underfunded by roughly a factor of four relative to industrial technologies. The pattern is broad-based across science: mathematics has a value-to-resource ratio above 12, and engineering, computer science, chemistry, and physics all exceed 3.

We also document heterogeneity in how the economic value of science evolves over time. Among computer science subfields, machine learning exhibits near-exponential growth in value, consistent with its role as a general-purpose tool with broad downstream applications, while the World Wide Web shows a sharp rise in the 1990s followed by sustained decline as its foundational contribution was absorbed into downstream fields. The same reallocation from foundational to applied work recurs within disciplines. In medicine, clinical fields such as oncology and physical therapy gain value, while basic areas such as immunology and pharmacology recede. In biology, biotechnology and bioinformatics rise, while genetics has fallen by more than half from its early-1990s peak as the foundational contribution of the genomic revolution was absorbed downstream. Physics is where this displacement is most pronounced, with the fundamental and observational subfields losing the most.

Why is science so valuable when it produces nothing consumers buy directly? Decomposing each field's value into successive spillover rounds shows that science's value is realized entirely downstream: in 2019, half of it arrives one spillover step away, another 28% two steps away, and the remaining fifth through longer chains. Case studies of four canonical science-origin innovations—nuclear physics in the early 1960s, molecular biology in the late 1980s, the Internet around 2000, and machine learning in the early 2020s—show that between 72% and 81% of each field's value ultimately lands in patented technologies. A direct-consumption accounting of value would therefore miss science almost entirely: science is valuable precisely through what technology does with it.

Related literature. Our work contributes to three strands of the literature. First, we add to the extensive literature estimating the private and social returns to R&D—through case studies of specific technologies (Griliches, 1958; Mansfield et al., 1977; Tewksbury, Crandall and Crane, 1980), through knowledge spillover approaches (Jaffe, 1986; Bloom, Schankerman and Van Reenen, 2013; Lucking, Bloom and Van Reenen, 2019; Eberhardt, Helmers and Strauss, 2013; Azoulay et al., 2019; Zacchia, 2020; Arqué-Castells and Spulber, 2022), and within macroeconomic growth models (Jones and Williams, 1998; Jones, Summers et al., 2020; Akcigit and Kerr, 2018; Acemoglu et al., 2018; Atkeson and Burstein, 2019; Garcia-Macia, Hsieh and Klenow, 2019; Bloom et al., 2020; Akcigit, Hanley and Serrano-Velarde, 2021; Cai and Tian, 2021; König et al., 2022); see Hall, Mairesse and Mohnen (2010) for a survey. A related strand focuses specifically on science-driven R&D (Myers and Lanahan, 2022; Arora, Belenzon and Sheer, 2021; Iaria, Schwarz and Waldinger, 2018; Arora et al., 2024; Mansfield, 1980; Fleming et al., 2019), examining how science feeds into inventions through patent-to-paper citations (Ahmadpoor and Jones, 2017; Marx and Fuegi, 2020) and documenting that science-driven R&D generates substantial private value (Agrawal and Henderson, 2002; Fleming and Sorenson, 2004; Branstetter, 2005; Poege et al., 2019; Bikard and Marx, 2020; Krieger, Schnitzer and Watzinger, 2024). We contribute to this literature in three ways: we build an endogenous growth model in which the marginal value of R&D is characterized in closed form as a function of the full innovation network, encompassing both invention and science; we estimate the economic value of science at the level of individual fields, enabling cross-field comparisons; and we construct a comprehensive innovation network spanning real-economy sectors and scientific fields from the global corpus of patents and publications, overcoming the data-availability and selection limitations of citation-based approaches.

Second, our work relates to the growing literature on knowledge linkages across sectors (Acemoglu, Akcigit and Kerr, 2016; Liu and Ma, 2021; Cai and Li, 2019; Huang and Zenou, 2020; Cai, Li and Santacreu, 2022; Guillard et al., 2021) and across countries (Caballero and Jaffe, 1993; Jaffe, Trajtenberg and Henderson, 1993; Eaton and Kortum, 1999, 2006; Coe and Helpman, 1995; Coe, Helpman and Hoffmaister, 2009; Santacreu, 2015; Buera and Oberfield, 2020; Cai et al., 2022), as well as to work on geographic knowledge diffusion from universities to local commercial innovation (Jaffe, 1989; Belenzon and Schankerman, 2013; Hausman, 2022; Andrews, 2023). While this literature focuses on knowledge linkages among real-economy sectors or between universities and nearby industry, our paper broadens the landscape by incorporating detailed scientific fields and tracing knowledge connections between science and technology within a unified framework.

Third, our work contributes to the science of science (Fortunato et al., 2018; Wang and Barabási, 2021; Liu et al., 2023; Azoulay et al., 2018). This literature examines the allocation and direction of research funding (Azoulay and Li, 2022; Myers, 2020; Yin et al., 2022; Ahmadpoor and Jones, 2017) and its effects on scientific output (Azoulay et al., 2019; Moretti, Steinwender and Van Reenen, 2019; Howell, 2017; Myers and Lanahan, 2022; Fleming et al., 2019), as well as how research is shaped by team composition (Lin et al., 2023; Ahmadpoor and Jones, 2019; Wu, Wang and Evans, 2019; Wuchty, Jones and Uzzi, 2007), intellectual property rights (Williams, 2013; Murray et al., 2016; Biasi and Moser, 2021), open access (McCabe and Snyder, 2005), international cooperation (Iaria, Schwarz and Waldinger, 2018; Yin et al., 2021), reward systems (Clauset, Arbesman and Larremore, 2015; Ma, Mondragón and Latora, 2015; Ma and Uzzi, 2018; Azoulay, Graff Zivin and Manso, 2011), scientific careers (Way et al., 2017; Sinatra et al., 2016; Liu et al., 2018), and diversity (Hofstra et al., 2020; Huang et al., 2020). We contribute a direct measure of the economic value of science at the field level—grounded in a fully specified welfare framework—that can serve as a rigorous empirical foundation for science policy.

The remainder of the paper is structured as follows. Section 2 describes the data. Section 3 presents the construction of the innovation network and its validation. Section 4 documents stylized facts about centrality and knowledge diffusion. Section 5 introduces the model and derives the closed-form characterization of the economic value of R&D. Section 6 presents the quantitative estimates of the economic value of science. Section 7 concludes.

2. Data

Our empirical analysis combines data from multiple sources: global patent records, academic publications, scientific funding grants, firm-level R&D expenditures, and sectoral production data. We describe how we construct and harmonize these data below.

2.1. Data on Academic Publications

We measure global scientific output using academic publication data from OpenAlex, an open-source catalog of the world's academic literature.[2] OpenAlex aggregates bibliographic

[2] We access OpenAlex via https://openalex.org/ in November 2025.

records from major publishers, indexing services, and institutional repositories into a unified, de-duplicated graph. Its primary data sources include Microsoft Academic Graph (MAG), Crossref, PubMed, and DataCite, among others. It standardizes these records by resolving publication metadata and author identities, yielding a comprehensive record of global scientific output. OpenAlex contains metadata for more than 463 million academic articles from 150,000 journals and 45,000 universities and research institutions, with historical coverage extending back to the 16th century and increasing comprehensiveness over the 20th and 21st centuries. For our purposes, the key elements of each record are the publication year, authors and their institutional affiliations, citations to prior work, the hierarchical scientific fields assigned by OpenAlex, and the title and abstract text.

In constructing the innovation network, we rely on the raw text of publication titles and abstracts. Titles are nearly universally available, with 99% of publications containing a non-missing title. Abstracts are available for approximately 49% of publications overall, with coverage improving substantially over time—rising from roughly 23% in 1900 to over 60% in recent years. OpenAlex stores abstract data in the form of an inverted index, recording both the words and their positions in the text. We use this structure to reconstruct a bag-of-words representation for each article, providing the basis for our text-based innovation network measure. In addition to these text features, each publication record contains references to prior scientific works, allowing us to construct a citation network that we later use to validate our text-based measures of the innovation network.

The database encompasses publications across all major scientific disciplines. We use the hierarchical field classification provided by OpenAlex to assign publications to research fields. OpenAlex builds this classification based on title, abstract, and journal title, using an automated classifier trained on the Microsoft Academic Graph corpus. The taxonomy comprises 284 level-1 concepts nested under 19 level-0 root domains. We focus on the 209 STEM-related level-1 fields spanning the natural sciences, engineering, mathematics, and computer science, which form the science half of our field set. Each article is typically tagged with several fields; we weight a work across its fields by its OpenAlex concept scores, renormalized to sum to one, and fall back to equal weights when scores are unavailable so that multi-field publications contribute proportionally to each relevant field.

Given the diversity of sources OpenAlex aggregates, we retain only works with a valid publication year that fall into one of four document types: journal articles, book chapters, books, and dissertations. We exclude datasets, reference entries, preprints, peer-review records, and other ancillary types. Starting from 463 million raw works, this selection yields 299 million publications used in our analysis.

OpenAlex indexes publications in a wide range of languages, including English, German, French, and Spanish, among others. Historically, English-language publications have accounted for approximately 50% of records since the 1850s, with the share rising to more than 70% since the 1950s. To build the text corpus on a common English representation while remaining inclusive of scholarship originally published in other languages, we machine-translate non-English text into English. We first detect the language of each title and abstract separately with a fastText language-identification model.[3] We then translate every non-English title and abstract using open-source neural translation models, Helsinki-NLP Opus-MT.[4]

2.2. Data on Patents

We capture innovation activity in industrial sectors using patent data from the Worldwide Bibliographic Data (DOCDB) and the Legal Status and Patent Family Data (INPADOC) databases, both maintained by the European Patent Office (EPO).[5] These open-source databases collectively form the foundation of the widely used Worldwide Patent Statistical Database (PATSTAT). DOCDB provides comprehensive bibliographic coverage of patents published by more than 100 patent authorities worldwide, while INPADOC supplements this with extended legal status information and detailed records of international patent families. Together, these databases contain more than 165 million patent publications from over 40 major patent authorities covering leading and developing countries over 1782–2025, of which 81 million are granted patents. Restricting to granted utility patents with a valid year and

[3] fastText language-identification model: https://fasttext.cc.

[4] Helsinki-NLP Opus-MT translation model: https://github.com/Helsinki-NLP/Opus-MT.

[5] We use the August 2025 release of DOCDB and INPADOC (data vintage 2025-08-31), distributed through the EPO's open bulk-data service at https://publication-bdds.apps.epo.org/raw-data/products. We parse the raw DOCDB and INPADOC XML into PATSTAT-standard tables ourselves rather than relying on the commercial PATSTAT product.

de-duplicating patent families leaves 33 million patents, which form the starting universe for our analysis.[6] The data for each patent include the application and grant year, technology classifications, the geographic locations of patent assignees and inventors, citations to prior patents, and title and abstract text. To better reflect the timing of inventive activity, we use the application year where available and the grant year otherwise.

As with scientific publications, we rely on textual information from titles and abstracts to construct the innovation network. Title information is widely available, with over 97% of granted patents containing a non-missing title. Abstract coverage is more limited, available for approximately 78% of granted patents; it began in 1916 and has increased steadily over time. For each patent, we incorporate both title and abstract text wherever possible to maximize information content. For earlier periods and patents with missing abstracts, we rely on titles alone. To validate the robustness of this approach, we compare an innovation network constructed using both titles and abstracts against one built from titles only. The two networks are highly correlated (correlation coefficient of 0.95), confirming that our approach is not sensitive to the inclusion of abstracts and that title text alone reliably supports network construction, particularly over longer historical periods.

The EPO data include patents filed in multiple languages. While many patent offices require filings in their national language, major patent systems increasingly accept or require English translations, and DOCDB provides English translations of many titles and abstracts, particularly for patents filed through international procedures such as the Patent Cooperation Treaty (PCT). For the remaining non-English titles and abstracts, we machine-translate the original-language text into English, as with the publication data. Only English text, native or translated, enters the text corpus.

We use three-digit IPC codes as our primary technology classification, which yields 123 distinct technological categories that can be further aggregated into 22 two-digit IPC groups.[7] Although the IPC system was formally established in 1971, it is now used by more than 100

[6] To avoid double-counting, we keep a single representative per family—the earliest-granted member, breaking ties by filing date—and attribute all citations to that representative. Families are identified from the INPADOC family identifiers in DOCDB, augmented by shared priority claims across applications.

[7] There are 131 technological categories at the three-digit IPC level. We retain 123 after dropping 8 labeled as "other" within each two-digit IPC group. These dropped categories account for less than 0.0005% of patents and have extensive missing periods.

countries, and patent offices often assign IPC codes retroactively to earlier filings. Where IPC codes are missing—especially for European patents before 1920—we rely on available CPC codes and map CPC to IPC using a bridge file constructed from patents with both classifications. For patents missing both IPC and CPC codes, particularly U.S. patents before 1920, we use USPC classifications and construct a separate USPC-to-IPC bridge.

We also use patent-to-paper citation data to validate our innovation network. These citations capture references from patent documents to scientific publications. We draw on two major sources: Reliance on Science (Marx and Fuegi, 2020, 2022), which extracts citations from the front pages of patent documents, and PatCit[8], which additionally processes citations from the full text. The two sources are highly overlapping; we combine them into a unified patent-to-paper citation database to improve coverage and robustness. We link each patent to DOCDB using the patent number and each scientific article to OpenAlex using the digital object identifier (DOI) or OpenAlex unique identifier.

2.3. Data on R&D Allocation

Our empirical analysis uses data on R&D allocation across industrial sectors and scientific fields by country. Since no single comprehensive dataset captures this information globally and consistently, we draw from two complementary sources—one for sectoral (technology-focused) R&D and one for field-level (science-focused) R&D.

Our measure of R&D allocation in industrial sectors aggregates firm R&D expenditures to the country-sector-year level using Worldscope, a global firm-level database covering more than 110,000 firms located in 160 countries and representing over 95% of the world's total market capitalization. The sectoral classification of firms in Worldscope follows the Standard Industrial Classification (SIC) system. To project firm-level R&D into IPC technology classes, we construct a SIC-to-IPC crosswalk following Lybbert and Zolas (2014). Specifically, we first link patent assignees from DOCDB to firms in Worldscope using a procedure analogous to those in Kogan et al. (2017), Ma (2026), and Liu and Ma (2021), and then compute the share of each firm's patents falling within each IPC class. We use these shares to allocate each firm's R&D expenditures across IPC classes. For multinational firms, we

[8] Accessed via https://zenodo.org/record/4391095.

further distribute R&D expenditures across countries in proportion to each firm's patenting activity by country, following Griffith, Harrison and Van Reenen (2006). We then aggregate to the country-IPC-year level.

Our second source, covering R&D allocation within scientific fields, is the scientific grant data from Dimensions AI. The Dimensions funding database contains 6.3 million grants from 651 funding agencies across 43 countries from 1952 through 2021. For each grant, the data record the funding amount and duration, the funding agency, researchers and their institutional affiliations, the hierarchical scientific fields assigned by Dimensions, and in many cases the resulting outputs in the form of publications, patents, and policy mentions. Dimensions classifies scientific fields using a proprietary hierarchical taxonomy comprising 171 fields organized under 22 knowledge domains. To harmonize this with our OpenAlex field definitions, we construct a field-to-field crosswalk based on the joint appearance of DOIs across the two databases: for academic articles with both OpenAlex and Dimensions metadata, we match their respective field assignments to map Dimensions fields to OpenAlex fields. We then project each grant's R&D allocation into the OpenAlex field structure using this mapping. Geographic assignment of grants follows the country of researchers' institutional affiliations. For multi-year grants, we distribute total funding equally across years to capture the actual duration of supported research activity.

Because our firm-based technology R&D (from Worldscope) and grant-based science R&D (from Dimensions) are measured on different scales, we calibrate the relative scale of science versus technology R&D to the U.S. aggregate series from the National Science Foundation's *National Patterns of R&D Resources* report, which documents national R&D by performer and by character of work over 1953–2024. We define technology R&D as business-performed R&D and science R&D as the sum of R&D performed by higher-education, non-profit, federal, and nonfederal-government performers. Under this definition, the science share of U.S. R&D ranges from 22% to 34% over the period.

2.4. Data on Production

Our core production panel is the U.S. GDP-by-industry accounts from the Bureau of Economic Analysis (BEA), which report value added, gross output, and intermediate inputs by industry and year over 1947–2024. We map BEA industries into IPC technology classes using the same assignee-based crosswalk logic described above for the R&D data (through a NAICS-to-IPC bridge). The production data serve two purposes in our analysis: the observed sectoral value-added and output shares provide the consumption weights β_i across technology sectors (with $\beta_i = 0$ for science fields), and sectoral value-added supplies the input for calibrating the knowledge-to-productivity parameter ψ.

3. Constructing an Innovation Network with Science

This section describes the construction of the innovation network connecting scientific publications and patents. The basic intuition is to exploit the lead-lag structure of knowledge diffusion embedded in the textual content of academic articles and patent documents: our method tracks how new knowledge from one field (upstream) is subsequently adopted in another field (downstream).

The construction proceeds in several steps. We first provide a parsimonious model of knowledge diffusion using text representations, which motivates our network construction framework. We then describe the empirical approach derived from the model and its implementation. Although the idea of tracking knowledge diffusion through textual terms is straightforward, the method must carefully account for a complex set of confounders to identify the underlying diffusion structure. Finally, we validate our innovation network using a Monte Carlo simulation and by comparison with citation-based networks in subsamples where both approaches are feasible. We also provide case studies of emerging technologies drawn from the network. Appendix A.1 provides more details.

3.1. Modeling the Innovation Network Using Textual Information

Textual representation and word-frequency vectors. Let there be N research fields indexed by i. Assume that the whole knowledge space of all fields is represented using text, with a fixed vocabulary of W distinct terms. Each research field i at time t, which comprises all documents newly created in that year (e.g., new academic papers and patents), can be viewed as a bag-of-words. Field i in year t is represented as a probability distribution $\boldsymbol{v}_{it} \in \mathbb{R}^{W\times 1}$ over words, with $\mathbf{1}'\boldsymbol{v}_{it} = 1$. We refer to $\boldsymbol{v}_{it}$ as the word-frequency vector

Table 1. Summary of Notation

Symbol	Dimension	Description	Economic Interpretation
N	Scalar	Number of Fields	The nodes of the innovation network (e.g., IPC classes or OpenAlex codes).
W	Scalar	Vocabulary Size	The dimensionality of the knowledge space (number of distinct terms).
$\boldsymbol{v}_{it}$	$W \times 1$	Word-Frequency Vector	The "genetic code" of field i at time t; a probability distribution over words.
$\boldsymbol{V}_t$	$N \times W$	Frequency Matrix	The aggregate state of knowledge across all fields at time t.
$\boldsymbol{P}_{t,t-1}$	$N \times N$	Lagged Similarity	The raw correlation of ideas between fields at $t-1$ and fields at t (Gross Similarity).
$\boldsymbol{P}_t$	$N \times N$	Contemporaneous Similarity	The static technological proximity between fields (used for partialling out).
$\boldsymbol{\Omega}_t$	$N \times N$	Diffusion Matrix	The network of knowledge spillovers; coefficients of the VAR system.
ω_{ijt}	Scalar	Diffusion Coefficient	The net flow of unique ideas from field j to field i, conditional on all other fields.
$\boldsymbol{\epsilon}_{it}$	$W \times 1$	Innovation Term	New technical vocabulary or random noise unique to field i at time t.

representing field i at time t. Let $\boldsymbol{V}_t \equiv [\boldsymbol{v}_{1t}, \cdots, \boldsymbol{v}_{Nt}]'$ denote the frequency matrix across fields at time t. In what follows, we use boldface letters to denote column vectors (lower case) and matrices (upper case).

Knowledge diffusion. Our notion of knowledge diffusion is based on the lead-lag information embedded in text similarities. We infer knowledge diffusion from field j (the upstream field) to field i (the downstream field) if the downstream field i's frequency vector becomes closer to the upstream field j's lagged frequency vector.

For illustrative purposes, consider the case of knowledge diffusion within a one-year window. Define the cosine similarity between the frequency vector of field i at time t and that of field j at time $t-1$ as:

$$[\boldsymbol{P}_{t,t-1}]_{ij} \equiv \boldsymbol{v}'_{it}\boldsymbol{v}_{jt-1}.$$

Correspondingly, $\boldsymbol{P}_{t,t-1} = \boldsymbol{V}_t\boldsymbol{V}'_{t-1}$ denotes the entire matrix of bilateral cosine similarities in frequency vectors with a one-period time lag.

It would be erroneous to infer knowledge diffusion solely from lagged cosine similarities, which may also reflect persistent overlap between fields rather than genuine knowledge spillovers. For example, two fields could be historically related—such as "Semiconductors" and "Computer Hardware"—sharing a large portion of technical vocabulary (e.g., "circuit," "voltage," "impedance") that remains stable over time. A high cosine similarity between the "Semiconductors" frequency vector at $t-1$ and the "Computer Hardware" vector at t might simply reflect this static overlap—the persistent technological proximity—rather than a genuine flow of new ideas from one to the other in that specific year. If diffusion were inferred from raw lagged similarity alone, one would mistakenly conclude that mature, stable fields are constantly "diffusing" knowledge to each other even in periods of stagnation. Such a measure captures the stock of shared knowledge rather than the flow of new knowledge.

This distinction is critical: a true measure of diffusion must identify when a downstream field adopts vocabulary from an upstream field that it did not previously possess, or when its trajectory is altered by the upstream field's recent innovations. We aim to capture this dynamic influence—how the past state of field j predicts the future composition of field i—rather than their static intellectual kinship.

To address this issue, we partial out the similarities between fields that do not represent knowledge diffusion. Define the contemporaneous cosine similarity matrix as

$$[\boldsymbol{P}_t]_{ij} \equiv \boldsymbol{v}_{it}' \boldsymbol{v}_{jt}, \qquad \boldsymbol{P}_t = \boldsymbol{V}_t \boldsymbol{V}_t'.$$

We measure the cross-field knowledge diffusion matrix at time t as

$$\boldsymbol{\Omega}_t = \boldsymbol{P}_{t,t-1} \boldsymbol{P}_{t-1}^{-1}. \tag{1}$$

We interpret the ij-th entry of $\boldsymbol{\Omega}_t$, denoted ω_{ijt}, as the knowledge spillover from field j to field i.

To see why this approach is helpful, note that mathematically, the operation $\boldsymbol{\Omega}_t = \boldsymbol{P}_{t,t-1} \boldsymbol{P}_{t-1}^{-1}$ corresponds to the coefficients of a multivariate orthogonal projection of the current state vectors onto the space spanned by the lagged state vectors. By multiplying the cross-temporal similarity matrix $\boldsymbol{P}_{t,t-1}$ by the inverse of the contemporaneous similarity matrix $\boldsymbol{P}_{t-1}^{-1}$, we effectively perform a generalized linear regression of $\boldsymbol{V}_t$ on $\boldsymbol{V}_{t-1}$.

This process partials out the contemporaneous correlation structure among the source fields at time $t-1$. In the absence of this adjustment (i.e., using only $\boldsymbol{P}_{t,t-1}$), if an upstream field j is highly correlated with another field k—for instance, Artificial Intelligence and Data

Science often share terms such as "learning" and "neural"—simple cosine similarity would attribute influence to both j and k redundantly. The inclusion of $\boldsymbol{P}_{t-1}^{-1}$ decorrelates the predictors, isolating the unique predictive power of each field j on field i's future trajectory. Consequently, the element ω_{ijt} captures the net knowledge spillover from field j to field i, conditional on the information contained in all other fields.

VAR Interpretation of the Model The diffusion matrix defined in (1) not only has an elegant mathematical expression but also offers an empirical advantage. Specifically, (1) implies that the network can be interpreted as the coefficient matrix of a vector autoregression (VAR), with the collection of frequency vectors across fields at time t as the dependent variable and their lagged counterparts as the explanatory variable:

$$\boldsymbol{v}_{it} = \sum_{j=1}^{N} \omega_{ijt} \boldsymbol{v}_{jt-1} + \boldsymbol{\epsilon}_{it}, \tag{2}$$

where $\boldsymbol{\epsilon}_{it}$ is a $W \times 1$ vector of mean-independent error terms. To see this, note that the coefficient vector $\boldsymbol{\omega}_{it} \equiv [\omega_{i1t}, \ldots, \omega_{iNt}]'$ from the regression is

$$\boldsymbol{\omega}_{it} = \left(\boldsymbol{V}_{t-1}\boldsymbol{V}_{t-1}'\right)^{-1} \boldsymbol{V}_{t-1}\boldsymbol{v}_{it},$$

and in matrix format, with $\boldsymbol{\Upsilon}_t \equiv [\boldsymbol{\epsilon}_{1t}, \cdots, \boldsymbol{\epsilon}_{Nt}]'$, equation (2) implies

$$\boldsymbol{V}_t = \boldsymbol{\Omega}_t \boldsymbol{V}_{t-1} + \boldsymbol{\Upsilon}_t, \qquad \boldsymbol{\Omega}_t' = [\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_N] = \left(\boldsymbol{V}_{t-1}\boldsymbol{V}_{t-1}'\right)^{-1} \boldsymbol{V}_{t-1}\boldsymbol{V}_t'.$$

Classical Connections The construction of $\boldsymbol{\Omega}_t$ connects to three classical results in statistics and econometrics, each illuminating a different facet of our approach. First, the formula $\boldsymbol{\Omega}_t = \boldsymbol{P}_{t,t-1}\boldsymbol{P}_{t-1}^{-1}$ is precisely the solution to the *Yule-Walker equations* for a first-order VAR: given a stationary process $\boldsymbol{V}_t = \boldsymbol{\Omega}\boldsymbol{V}_{t-1} + \boldsymbol{\epsilon}_t$, pre-multiplying both sides by $\boldsymbol{V}_{t-1}'$ and taking expectations yields $\Gamma(1) = \boldsymbol{\Omega}\,\Gamma(0)$, so the coefficient matrix is recovered as $\boldsymbol{\Omega} = \Gamma(1)\Gamma(0)^{-1}$. Our matrices $\boldsymbol{P}_{t,t-1}$ and $\boldsymbol{P}_{t-1}$ are exactly the empirical counterparts of the lag-1 and lag-0 autocovariance matrices, making our estimator the classical moment estimator for VAR coefficients. Second, the right-multiplication by $\boldsymbol{P}_{t-1}^{-1}$ implements the logic of the *Frisch-*

Waugh-Lovell theorem: the coefficient on an upstream field j in the regression of $\boldsymbol{v}_{it}$ on all lagged field vectors is identical to the coefficient obtained after first removing from $\boldsymbol{v}_{jt-1}$ the linear projection onto all other upstream fields. This is why ω_{ijt} captures the *net* influence of field j on field i—it represents what j contributes that no other field can substitute for. Third, the matrix $\boldsymbol{P}_{t-1}^{-1}$ is the *precision matrix* of the contemporaneous field distribution. In Gaussian graphical models, the off-diagonal entries of a precision matrix encode partial correlations—the association between two nodes after conditioning on all others. Our construction exploits this property in the time-series direction: by multiplying the lagged cross-similarity by the contemporaneous precision matrix, we isolate the genuinely novel predictive content each upstream field contributes to its downstream neighbors, purging any influence that operates only through correlated third fields.

3.2. Estimation Procedure

We now estimate equation (2) in the data. We measure the frequency vector $\boldsymbol{v}_{it}$ using term-frequency-inverse-document-frequency (TFIDF) weighting, a standard approach in natural language processing. The vocabulary consists of unigrams, bigrams, and trigrams. TFIDF rescales each component of the raw term-frequency vector by an inverse-document-frequency weight, thereby down-weighting words that appear frequently across many documents while amplifying terms that are distinctive to a particular field. To prevent the far larger publication corpus from dominating the term weights, we use a domain-balanced inverse-document-frequency that assigns equal weight to the patent and publication document shares. After computing TFIDF, we further rescale each field-year vector to have unit ℓ_2 norm so that the measure reflects relative term composition rather than the overall length of the document.

To estimate the diffusion matrix $\boldsymbol{\Omega}_t$ for year t, we first collapse the lagged knowledge vectors into a single multi-year average and then perform the regression. We average over lags because knowledge spillovers unfold over several years and the precise diffusion horizon is unknown a priori. The average captures the cumulative effect of the preceding decade while remaining agnostic about the exact timing. Concretely, for each upstream field j we form the mean of its lagged word-frequency vectors,

$$\bar{v}_{jwt} = \frac{1}{L}\sum_{\tau=1}^{L} v_{jw,t-\tau}, \qquad L = 10,$$

and estimate, for each downstream field i, the single regression

$$v_{iwt} = \sum_{j} \omega_{ijt}\, \bar{v}_{jwt} + \alpha_{it} + \alpha_{iw} + \varepsilon_{iwt}. \tag{3}$$

The dependent variable is the TFIDF frequency of word w for field i in year t. The explanatory variables are the ten-year mean lagged frequencies $\bar{v}_{jwt}$ of word w across all fields j (including field i itself). We estimate this equation for each field i using data from the ten-year window ending in year t, stacking every word across the ten years to obtain $10 \cdot W$ observations with N right-hand-side variables (and N coefficients) per regression. This baseline specification (a ten-year estimation window with a ten-lag mean) is used throughout the paper, and our estimates are robust to alternative specifications.

We introduce field-by-word fixed effects α_{iw} to absorb time-invariant vocabulary differences across fields. We also include field-by-time fixed effects α_{it} to normalize the TFIDF vector for each field in a given year. The regression generates one diffusion coefficient ω_{ijt} per field pair, summarizing the cumulative ten-year spillover from field j to field i.

Because each equation (3) involves a large number (N) of estimated coefficients for each field i, we regularize the estimator. After partialling out the fixed effects, we estimate a ridge fixed-effects VAR with a penalty parameter of 0.2.[9] We then apply a five-year calendar rolling mean to the estimated coefficient series to smooth year-to-year sampling noise. Finally, since the knowledge diffusion matrix $\mathbf{\Omega}_t$ is assumed to be non-negative with rows summing to one, we winsorize negative entries to zero and row-normalize so that $\mathbf{1}'\boldsymbol{\omega}_{it} = 1$.

[9]As is common in the statistics literature, the penalty is scaled relative to the magnitude of the median diagonal of the design's second-moment matrix (Hoerl and Kennard, 1970; Hastie, Tibshirani and Friedman, 2009).

3.3. Network Description and Validation

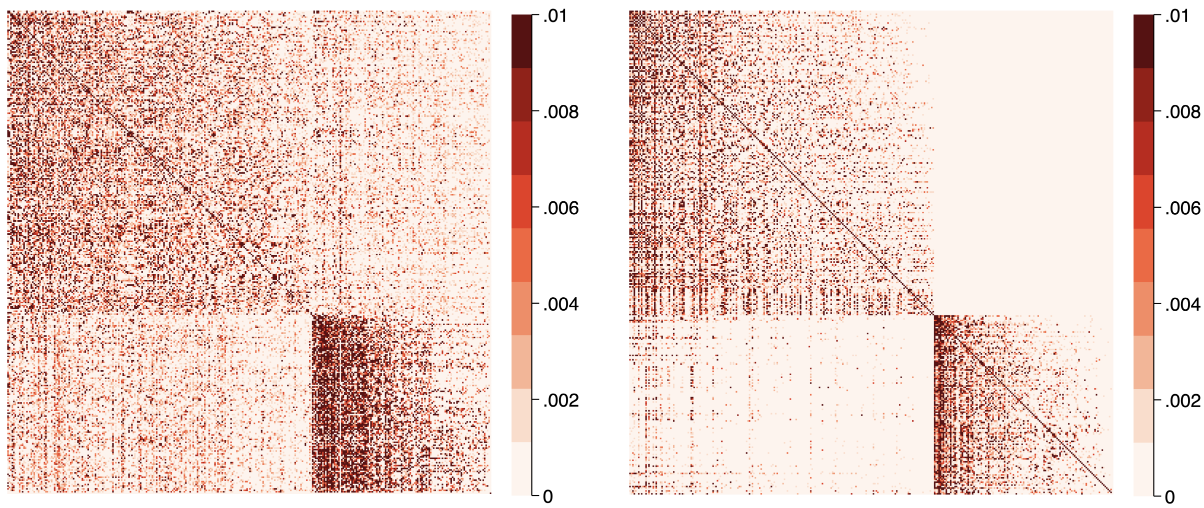
The Advantage of the Text-Based Network Over Citation-Based Networks We first compare our innovation network with existing network measures in the literature. Prior work relies heavily on patent citation data to construct knowledge spillover networks (Acemoglu, Akcigit and Kerr, 2016; Liu and Ma, 2021), typically focusing on within-technology citations and consequently missing cross-domain linkages. Recent work by Marx and Fuegi (2020, 2022) makes important progress by linking patents to scientific articles, but patent-to-paper links remain incomplete, as we document below. Comprehensive paper-to-patent citation data are generally unavailable. Moreover, because the units of knowledge in scientific papers and patents differ fundamentally, citation-based networks are inherently limited in capturing the full cross-domain structure of knowledge flows.

Figure 1, Panel (a) visualizes our text-based innovation network $\mathbf{\Omega}_{2010}$ as a heatmap. The first 209 rows and columns correspond to the 209 science fields, and the last 123 rows and columns correspond to the 123 three-digit IPC technology classes, with fields and classes sorted by decreasing Katz centrality. The color in the i-th row and j-th column corresponds to $\omega_{ij,t}$ using the colormap shown to the right. The heatmap reveals a clear block structure. Within each domain, innovation-central sectors account for a disproportionate share of knowledge flows to other sectors while receiving relatively little in return from non-central ones. Moreover, the off-diagonal science-to-patent quadrant (bottom-left corner) exhibits active knowledge flows, consistent with the finding that knowledge primarily flows from science into technology. In contrast, the patent-to-science quadrant (upper-right corner) is considerably sparser.

Figure 1, Panel (b) visualizes the citation-based network as a heatmap, combining patent-to-patent citations, paper-to-paper citations, and the patent-to-paper citations described in the previous section. Both networks exhibit similar within-domain block structure; however, our text-based network has the key advantage of capturing knowledge flows between science and technology that are largely absent from citation-based approaches. The science-to-patent quadrant of the citation-based heatmap is quite sparse, and the patent-to-science quadrant is effectively empty—in sharp contrast to the rich cross-domain structure visible in the text-

Figure 1. Text-Based and Citation-Based Innovation Network: Heatmap

Panel (a): Text-Based Network   Panel (b): Citation-Based Network



Notes. This figure visualizes the text-based network in Panel (a) and the citation-based network in Panel (b). Darker colors represent larger matrix entries. Sectors are first ordered by technological class and scientific field, and then ordered within each group by decreasing Katz centrality.

based network. The citation-based network also displays a much darker diagonal, indicating that citations are predominantly within-field, a tendency reinforced by the publication and patent review processes.

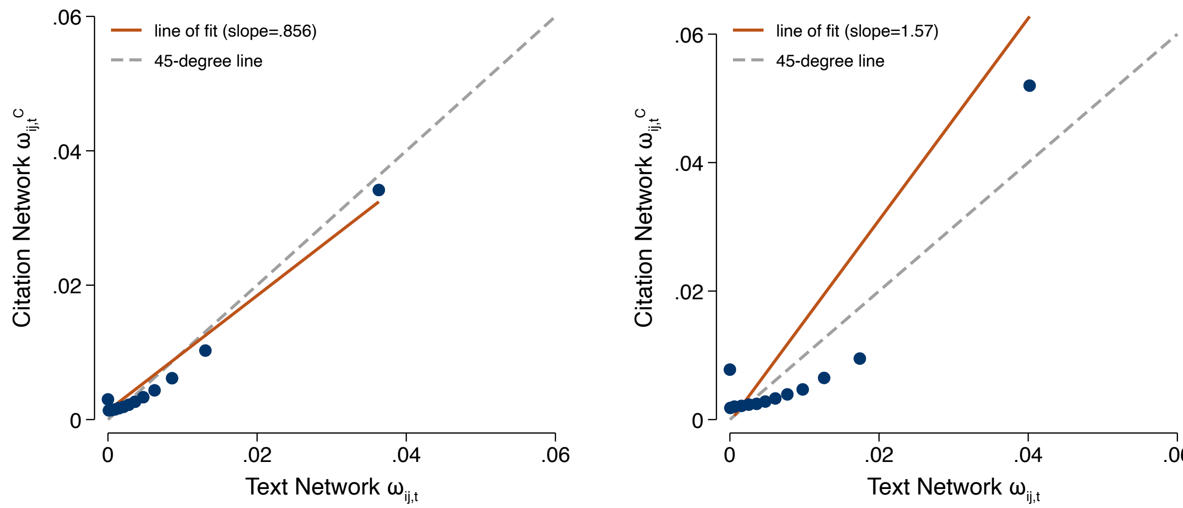
Validation: Alignment Between Text-Based and Citation-Based Networks We perform a series of tests to validate our measure's ability to capture knowledge flows. Figure 2 compares our text-based network with the citation-based network for within-science fields (Panel a) and within-technology fields (Panel b). Each panel plots binned values from both networks alongside a linear fit. In both cases, we observe a strong positive relationship closely aligned with the 45-degree line, indicating that our text-based innovation network successfully captures knowledge flows that are consistent with citation evidence.

Table A.1 further quantifies the relationship between the two networks by reporting Pearson and Spearman correlations separately for patent fields and science fields across decades. The correlation is consistently positive in both domains, supporting the validity of our text-based network, and it increases over time: the Pearson correlation within science rises from 0.21 in 1950 to 0.58 in 2020, with a similar pattern within technology. This upward

Figure 2. Alignment Between the Text-Based and Citation-Based Networks

Panel (a): Scientific Field

Panel (b): Technology Class



Notes. This figure shows a binned scatterplot of elements from the text-based and citation-based innovation networks using data from 1940 to 2024. Panel (a) considers the network within scientific fields and Panel (b) considers the network within technological classes.

trend likely reflects both improvements in citation data quality and the growth in document volume, which makes the citation-based network increasingly informative about underlying knowledge flows. These findings further support the conclusion that our text-based network captures meaningful knowledge diffusion that is consistent with, but extends well beyond, citation-based evidence.

Monte Carlo Exercise A key challenge in validating a text-based innovation network is the absence of a ground truth in real-world settings. To address this, we implement a Monte Carlo simulation to assess whether our method can recover the true structure of an innovation network from text representations, under the assumption that the knowledge diffusion model holds and the data-generating process is known. This exercise also helps rule out simpler alternatives, such as raw cosine similarity. Details of the simulation exercise are provided in Appendix Section A.1.3.

In the simulation, documents for each field are generated via a multinomial process based on a known innovation network and the knowledge diffusion model introduced above. We simulate the process over 1,000 periods and use data from the final 100 periods to estimate the network, allowing us to assess whether our estimation procedure accurately reconstructs the true network structure from text alone.

Figure A.1 shows that our method effectively recovers both the overall network structure and the relative importance of fields. The estimated innovation network closely matches the ground-truth network, with a correlation of 0.93. Moreover, the derived Katz centrality measures align closely with those from the true network, achieving a correlation of 0.99. In contrast, a simple cosine similarity matrix fails to capture the directional structure of diffusion, producing a nearly uniform and uninformative network.

3.4. Event Studies: Tracing Knowledge Diffusion Using Words

To further validate our text-based innovation network, we conduct event studies that trace how distinctive keywords diffuse from focal innovation fields to related fields over time. This approach allows us to directly assess whether terminology associated with major technological innovations spreads to the fields our network identifies as downstream recipients.

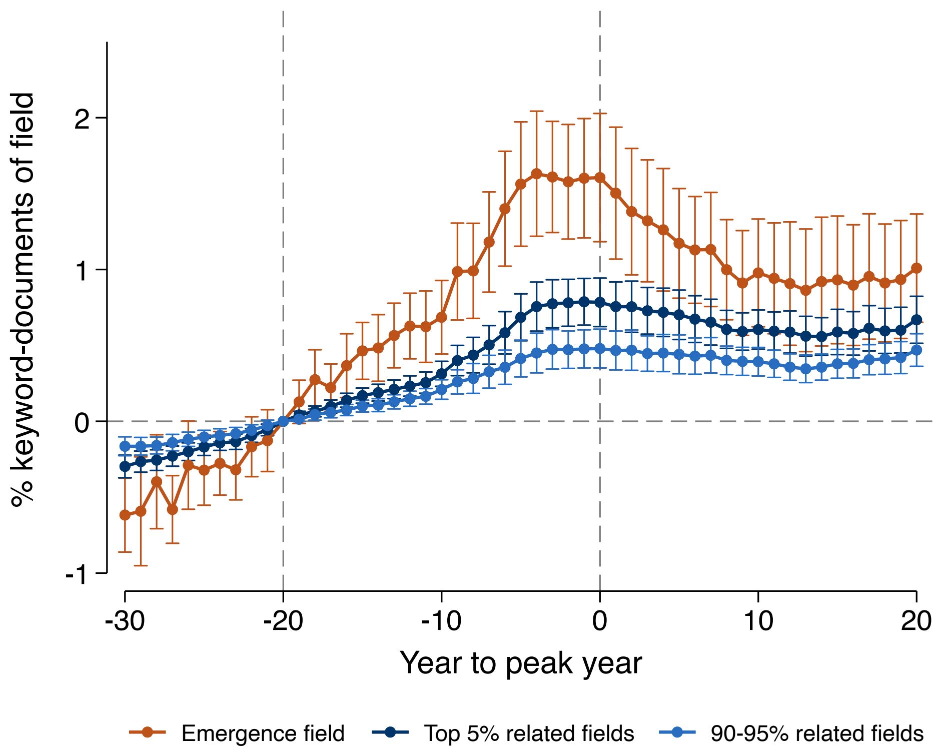
We implement this analysis in three steps. First, we identify major innovation episodes by applying a Markov regime-switching model to the Katz centrality time series for each field, detecting sustained intervals during which a field is unusually central within the innovation network. We retain an episode as "major" only if it emerges from a low base, screening out fields that were already central. This procedure identifies 95 distinct innovation episodes (71 in science and 24 in technology), each characterized by an emergence year, peak year, and decline year. Appendix Table A.2 lists all episodes along with their maximum Katz centrality and associated keywords. Appendix Figure A.3 visualizes the Katz centrality trajectories for the top 20 fields by maximum centrality, with gray shading indicating the major innovation waves identified by our procedure.

Second, for each innovation episode, we identify the most important keywords characterizing the focal field during its peak period using R^2 decomposition, which quantifies how much each word contributes to explaining variation in the field's word-frequency vector relative to other fields. This approach focuses on terminology that is distinctive to the innovation episode rather than generic technical language. The resulting keywords are also reported in

Appendix Table A.2.

Third, we trace the diffusion of these keywords from the focal field to other fields over time. For each episode, we identify the top 50 most important keywords and compute for each field in each year the average share of documents containing at least one of these keywords. We then group recipient fields by their dependence on the focal field as measured by our innovation network, enabling us to examine whether fields more closely connected to the focal field in the network are also more likely to adopt its terminology.

Figure 3. The Diffusion of Important Keywords Across Fields Over Time



Notes. This figure shows the diffusion of important keywords across fields over time. For each episode, we identify the top 50 most important keywords and compute for each field and year the average share of documents containing at least one keyword. Fields are grouped by their network-based dependence on the focal emergence field.

Figure 3 presents the results, tracing keyword adoption over time for fields grouped by their network proximity to the emerging innovation field. The patterns reveal clear evidence of directional knowledge diffusion: keywords associated with major innovations first appear in the focal field, then spread to closely connected fields, and eventually diffuse to more distant fields. This temporal ordering aligns with the directional structure of our innovation network, providing direct validation of our text-based approach. Appendix Figure A.4 provides additional examples illustrating this diffusion process for specific innovation episodes.

4. Knowledge Evolution and Diffusion: Stylized Facts

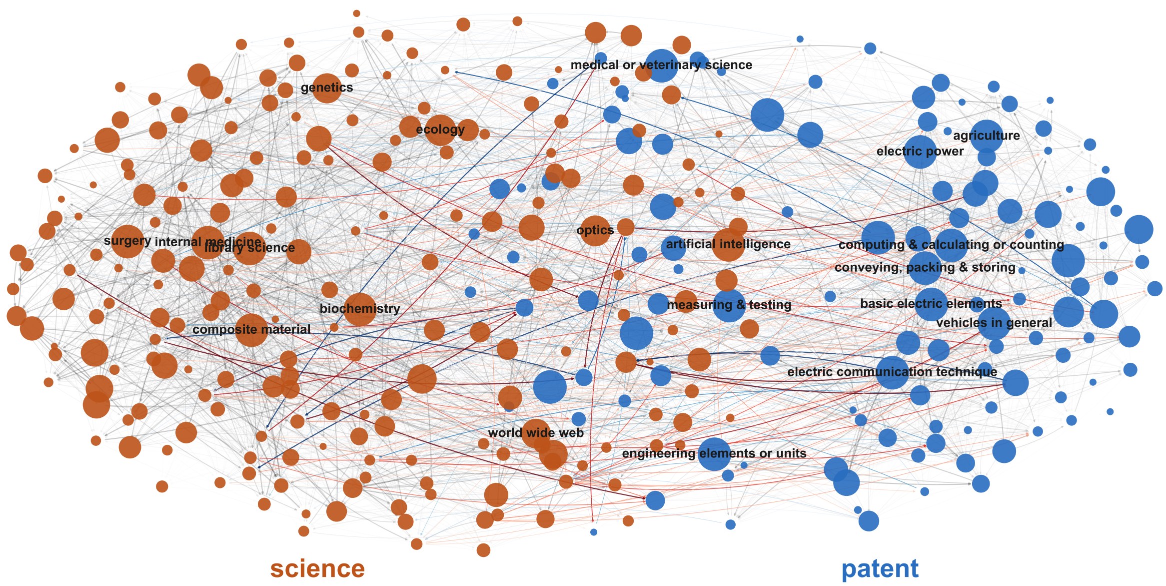
In this section, we document three stylized facts drawn from the innovation network that jointly motivate the key mechanisms in our theoretical framework and quantitative analysis: the structural dominance of science over technology in knowledge centrality; the heterogeneous and cyclical dynamics of centrality rankings over time; and the systematic role of breakthrough innovations in reshaping a field's network position.

4.1. Fact #1: Science Is More Central

Figure 4 visualizes the innovation network, aggregating knowledge flows by averaging the diffusion matrix $\mathbf{\Omega}_t$ over our full sample period. We plot the knowledge diffusion links between patent fields (blue nodes) and science fields (red nodes). Each node represents a distinct field, and node size is proportional to the total number of documents in our sample; we label the 10 largest patent and science fields by document volume. Directed arrows represent the flow of knowledge from an upstream field j to a downstream field i: gray arrows indicate within-domain flows, red arrows indicate science-to-patent flows, and blue arrows indicate patent-to-science flows. Arrow width reflects the magnitude of the knowledge flow coefficient ω_{ij}.

Cross-domain flows exhibit strong asymmetry. There are substantially more science-to-patent flows (red arrows) than patent-to-science flows (blue arrows), indicating that knowledge diffusion runs predominantly from science into technology rather than the reverse. Scientific fields such as the World Wide Web, artificial intelligence, and biochemistry serve as key knowledge sources across a broad range of patent fields, consistent with their foundational role in enabling applied innovation. The reverse channel—patent fields diffusing knowledge into science—is far more limited and concentrated in a handful of the largest technology classes, such as computing & calculating or counting and electric communication technique.

Figure 4. Visualization of Innovation Network

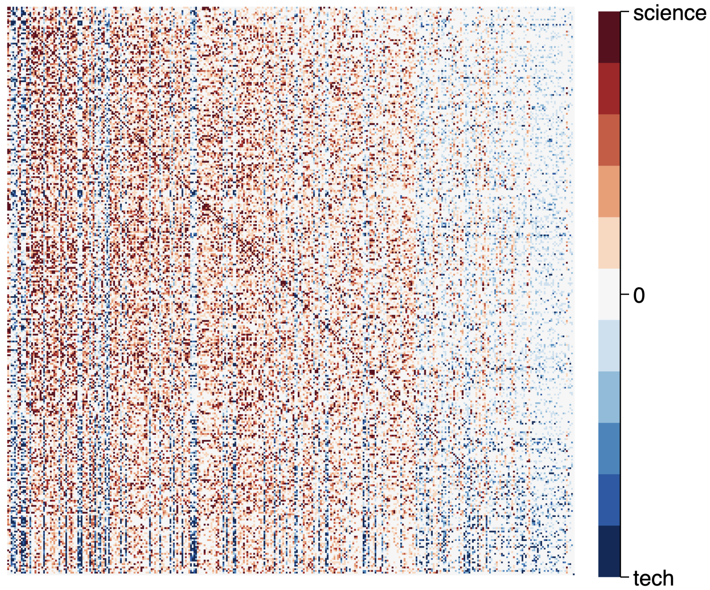


Notes. This figure visualizes the average innovation network, computed by averaging the diffusion matrix $\mathbf{\Omega}_t$ over the full sample period (1870–2024). Blue nodes represent 3-digit IPC patent (technology) classes and red nodes represent OpenAlex scientific fields. Node size is proportional to the total number of documents in that field over the sample period, and the 10 largest fields in each domain are labeled. Directed arrows indicate knowledge flows from the source (upstream) field to the destination (downstream) field: gray arrows represent within-domain flows, red arrows represent science-to-patent flows, and blue arrows represent patent-to-science flows. Arrow width is proportional to the strength of the knowledge flow coefficient ω_{ij}.

Figure 5 presents the same 332×332 innovation network as a heatmap, sorting all fields by Katz centrality within each domain. Each cell reflects the intensity of knowledge flow ω_{ij} between an upstream field j (column) and a downstream field i (row): red shading indicates flows originating in science fields, and blue shading indicates flows originating in patent fields. Darker colors represent larger matrix entries. Two patterns are immediately visible. First, the network is highly concentrated: a small number of fields in the upper-left corner of each domain block account for a disproportionately large share of knowledge outflows, while the majority of fields receive little diffusion from non-central sources. Second, the science-to-patent quadrant (bottom-left) is substantially denser than the patent-to-science quadrant (upper-right), corroborating the directional asymmetry seen in Figure 4: science is structurally more upstream in the knowledge diffusion process.

This pattern is further quantified in Table 2, which lists the top 10 science and patent fields by Katz centrality in 2000. A field's Katz centrality aggregates all the direct and

Figure 5. Visualization of Innovation Network: Heatmap



Notes. This figure presents the average innovation network (1870–2024) as a 332 × 332 heatmap. The first 209 rows and columns correspond to scientific fields (OpenAlex classification) and the last 123 rows and columns correspond to 3-digit IPC technology classes; within each domain, fields are ordered by decreasing Katz centrality, so the most central fields appear nearest the upper-left of each block. Cell color represents the magnitude of the knowledge flow coefficient ω_{ij} from upstream field j (column) to downstream field i (row). Red shading indicates flows from upstream science fields; blue shading indicates flows from upstream patent fields; darker colors indicate larger entries.

indirect influence it exerts on the rest of the network: a field with high Katz centrality not only diffuses knowledge to many fields directly, but also transmits it further through the network via intermediate fields. The Rank column in both panels reflects each field's position in the *combined* ordering across all 332 fields (209 science and 123 technology). Among the highest-ranking fields overall, science disciplines dominate: internet privacy, computational biology, and data mining occupy the top three positions (internet privacy leads the entire network with a Katz centrality of 0.0111), and seven of the ten most central fields are scientific. The highest-ranked technology class—electric communication technique—appears at overall rank 4 with a Katz centrality of 0.0078, and other leading technology classes such as hand tools, nuclear physics, and basic electric elements fall between overall ranks 8 and 13. This pattern arises because science fields are more upstream: they influence both other science fields and downstream technology sectors, whereas technology sectors primarily influence one another. Appendix Table A.3 shows that this dominance of science in overall centrality rankings is a persistent feature of the post-WWII innovation network; Appendix Figure A.3 traces the Katz centrality trajectories of the top fields in each domain over the full sample.

Table 2. Top 10 Science and Technology Fields by Katz Centrality in 2000

Scientific Fields			Technological Classes		
Rank	Katz Centrality	Field Name	Rank	Katz Centrality	Field Name
1	0.0111	internet privacy	4	0.0078	electric communication technique
2	0.0083	computational biology	8	0.0061	hand tools
3	0.0081	data mining	10	0.0060	nuclear physics
5	0.0063	molecular biology	13	0.0056	basic electric elements
6	0.0063	computer network	17	0.0052	biochemistry, beer, microbiology & enzymology
7	0.0062	cell biology	23	0.0050	medical or veterinary science
9	0.0061	human-computer interaction	25	0.0050	coating metallic material
11	0.0058	world wide web	28	0.0047	vehicles in general
12	0.0057	information retrieval	32	0.0046	specific information & communication tech.
14	0.0056	gerontology	33	0.0045	computing & calculating or counting

Notes. This table lists the top 10 scientific fields and the top 10 technology classes by Katz centrality in 2000, each selected from within their own domain. The Rank column reflects each field's position in the *combined* ordering across all 332 fields (209 science fields and 123 technology classes). Science fields hold the top three positions in the combined ranking, and seven of the ten most central fields overall are scientific; the highest-ranked technology class enters at overall rank 4. Katz centrality is computed from the text-based innovation network $\mathbf{\Omega}_{2000}$ as described in Section 3; higher values indicate greater upstream influence in the knowledge diffusion process.

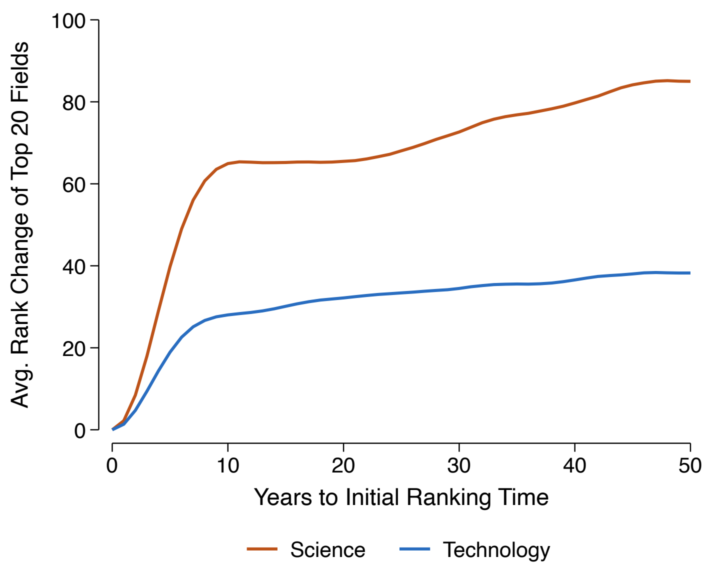
4.2. Fact #2: Heterogeneous Centrality Dynamics Across Domains

Figure 6 summarizes how centrality rankings evolve over time for the most central fields in science and technology. For each year, we identify the top 20 fields by Katz centrality within each domain—science (red) and technology (blue)—and track how their centrality rankings within the full 332-field network change over subsequent years. The figure plots the average change in the combined-domain ranking for these initially top-ranked fields as a function of years elapsed since they entered the top 20.

The rank changes reveal a modest level of instability, with distinct dynamics between science and technology. Patent fields experience relatively large short-run movements in their rankings. These fluctuations are likely driven by product cycles, shifting market demands, or regulatory changes that temporarily elevate a technology class and then fade as the wave passes. After this initial volatility, however, the average rank change stabilizes, suggesting that once a technology field secures a relatively central position, it tends to persist there over longer horizons—consistent with the path-dependent and cumulative nature of applied technological knowledge.

In contrast, science fields exhibit a steeper long-term decline in centrality. While their initial rank change is comparable to or smaller than that of technology fields, their rankings continue to drift steadily over time rather than stabilizing. This persistent downward drift suggests that initially central scientific fields are more likely to be displaced by new entrants, as foundational breakthroughs in one area of science open entirely new research directions that quickly become more influential. This instability reflects the exploratory and paradigm-shifting character of scientific progress, whereby a major breakthrough—genomics, nanotechnology, or machine learning—can rapidly reorganize the entire knowledge diffusion structure, pushing established central fields further down the hierarchy.

Figure 6. Change in Rankings of Top Central Fields



Notes. This figure tracks how the centrality rankings of the most central fields in each domain evolve over time. In each year, we identify the top 20 fields by Katz centrality within the science domain (red) and the technology domain (blue). The horizontal axis measures the number of years elapsed since the field entered the top 20; the vertical axis shows the average change in combined-domain Katz centrality rank (across all 332 fields) relative to the entry year. Positive values indicate rank deterioration (the field has become less central relative to others). Results are averaged across all entry cohorts in the sample.

Long-Run Cyclicality of Network Centrality We next characterize the long-run dynamics of Katz centrality at the aggregate level for both science and technology. To extract the cyclical component from the trend, we estimate an unobserved-components time-series model in state-space form, following Durbin and Koopman (2012).[10] We estimate the model for each aggregate field separately by maximum likelihood via the Kalman filter. Appendix Figure A.5 reports the model fit (R^2) and estimated cycle periods for each field. Figure 7 displays the estimated cyclical components (gray) and their deterministic transition paths (blue, excluding the noise innovations) for each field.

The figure reveals pronounced cyclical patterns in innovation centrality at both the science and technology levels. The science fields, shown in the top panel of Figure 7, tend to exhibit longer and smoother cycles with an average estimated period of approximately 40 years. These multi-decade swings are consistent with the slow-moving nature of foundational knowledge production and paradigm shifts in science: a dominant scientific discipline rises to prominence as it opens new research directions, sustains central influence over a period of roughly one to two generations, and then gradually recedes as successors emerge. This pattern echoes historical episodes such as the post-war rise of nuclear physics, the genomics revolution, and the ongoing surge of machine learning.

Technology fields, shown in the bottom panel, follow shorter-term and more volatile cycles with an average estimated period of approximately 28 years—roughly 12 years shorter than the science cycles. This faster rhythm likely reflects the tighter feedback loops between technological innovation and market adoption: a general-purpose technology or platform (e.g., combustion engines, semiconductors, or the internet) rises to centrality as it diffuses broadly across the economy, but is eventually supplanted by the next platform, on a timescale governed more by commercial and economic forces than by the pace of fundamental scientific discovery.

[10]More specifically, the unobserved components model decomposes a time series into trend, cyclical, and noise components (Durbin and Koopman, 2012):

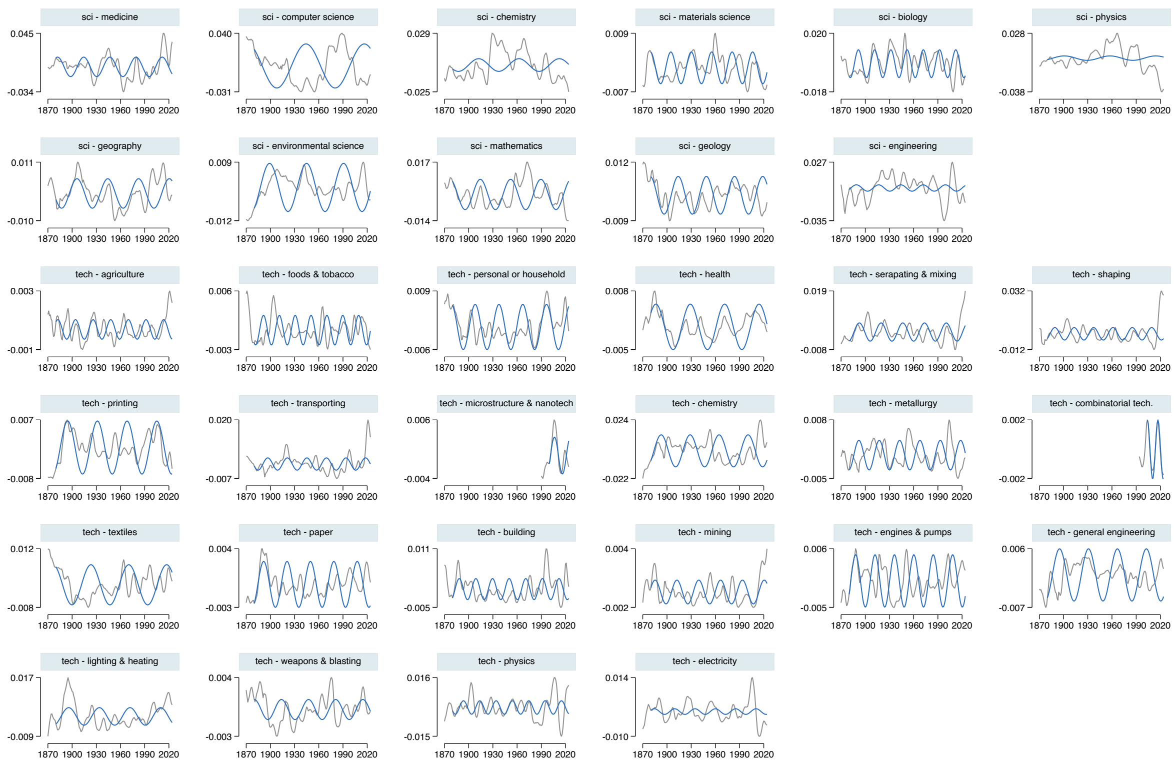
$$y_t = a_t + c_t + \epsilon_t,$$

where y_t is the observed Katz centrality at time t, a_t is the trend component, c_t is the cyclical component, and $\epsilon_t \sim N(0, \sigma_y^2)$ is the noise. The trend follows a local linear model, $a_{t+1} = a_t + b$, where b is a constant drift. The cyclical component is parameterized as a bivariate state-space model:

$$c_{t+1} = c_t \cos(\lambda_c) + c_t^* \sin(\lambda_c) + \epsilon_t^c, \quad \text{and} \quad c_{t+1}^* = -c_t \sin(\lambda_c) + c_t^* \cos(\lambda_c) + \epsilon_t^{c*},$$

where $\epsilon_t^c, \epsilon_t^{c*} \sim N(0, \sigma_c^2)$ are i.i.d. innovations. The parameter $\lambda_c \in (0, \pi)$ governs the frequency of the cycle; we restrict the implied period $2\pi/\lambda_c$ to lie between 8 and 80 years to exclude implausibly short- or long-run oscillations.

Figure 7. Long-Run Cyclicality of Network Centrality



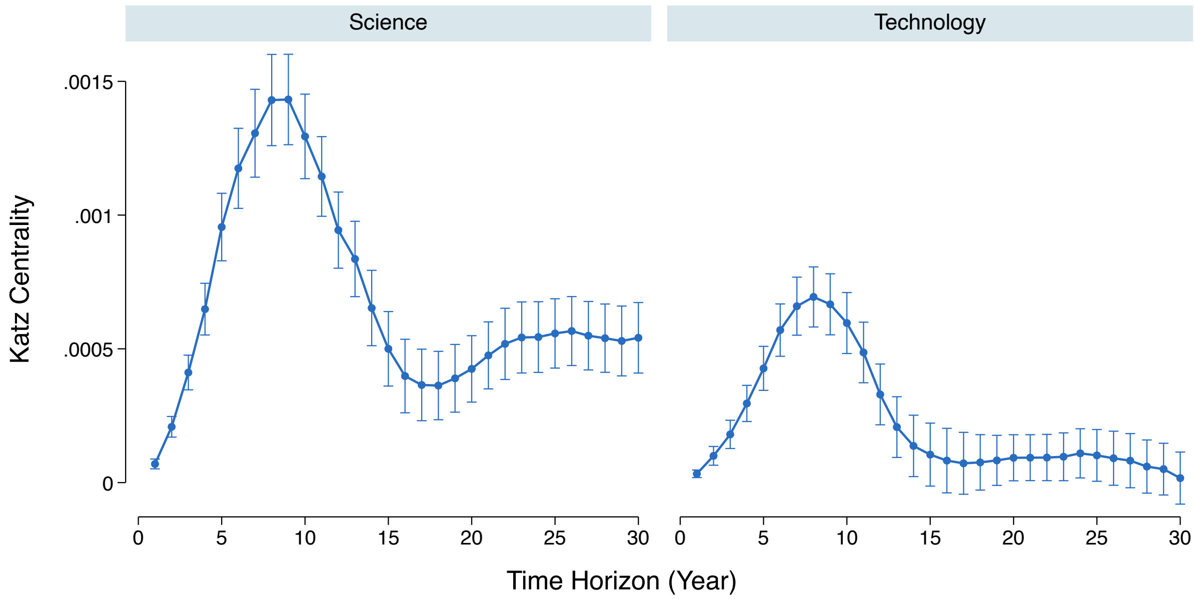
Notes. This figure shows the estimated cyclical components of Katz centrality from an unobserved-components time-series model (see footnote in Section 4), estimated separately for aggregate science fields (top panel) and aggregate technology classes (bottom panel). For each field, the gray lines show the full estimated cyclical component c_t (including noise innovations), and the corresponding blue lines show the deterministic transition component—the path of the cycle excluding noise—which captures the smooth, mean-reverting oscillation. Each model is estimated by maximum likelihood via the Kalman filter using the annual Katz centrality series from 1870 to 2024, with the cycle period restricted to lie between 8 and 80 years. Appendix Figure A.5 reports the R^2 and estimated periods of the models for each field.

4.3. Fact #3: Emerging Field Centrality After Breakthrough Innovation

We examine how breakthrough innovations affect a field's position within the innovation network, as measured by its Katz centrality—which, as noted above, captures a field's direct and indirect influence on the rest of the network, including spillovers several steps removed from the source.

Figure 8 presents local projections of the effect of breakthrough innovations on innovation centrality. We identify breakthrough innovations following Kelly et al. (2021). For each document, we compute a novelty score as the ratio of its textual similarity to future

Figure 8. Innovation Centrality and Breakthrough Innovations



Notes. This figure shows local projection estimates of the dynamic effect of breakthrough innovations on Katz centrality, separately for science fields (Panel a) and technology classes (Panel b). The horizontal axis measures the forecast horizon in years; the vertical axis shows the estimated coefficient on log breakthrough count at each horizon. Breakthrough innovations are identified following Kelly et al. (2021): for each document, we compute a novelty score as the ratio of its textual similarity to documents published in the same field over the subsequent five years to its similarity to documents in the field over the prior five years; after demeaning the novelty score by calendar year to remove secular language drift, documents in the global top 5% of the year-demeaned distribution are classified as breakthroughs. For each field-year, we count breakthrough innovations accumulated over the preceding ten years and take the log of this count. The local projection regression at horizon h regresses Katz centrality at $t + h$ on log breakthrough count at t and log knowledge stock at t, controlling for field and year fixed effects. Shaded bands show 95% confidence intervals based on standard errors clustered at the field level.

documents in the same field (over the next five years) to its similarity to past documents (over the previous five years). Documents with high scores are distinct from prior work but closely predictive of subsequent innovations, indicating that they represent breakthroughs that open new research directions. After demeaning the novelty score by calendar year to control for secular language shift, we rank documents globally across all fields and years and classify those in the top 5% of the year-demeaned distribution as breakthroughs. For each field-year, we count breakthrough innovations accumulated over the preceding ten years and estimate their effect on future centrality using local projection methods.

We find that breakthrough innovations are associated with a subsequent increase in innovation centrality. In both domains, centrality rises sharply in the first 5–10 years following a surge in breakthroughs. The increase is more pronounced and persistent in science fields, suggesting that scientific breakthroughs not only elevate a field's short-term centrality but also restructure the broader diffusion architecture of the network over time. This accords with the transformative nature of fundamental discoveries: a breakthrough in molecular biology or machine learning does not merely improve output in that field—it reorients the entire downstream knowledge diffusion structure, pulling many other fields into its orbit.

Table 3. Breakthrough Innovation Predicts Innovation Centrality

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Katz Centrality$_{i,t+1}$				Katz Centrality$_{i,t+10}$			
ln(Breakthrough, 95%)$_{i,t}$	0.00039***	0.00019***			0.00021***	0.00005**		
	(0.000)	(0.000)			(0.000)	(0.000)		
ln(Breakthrough, 90%)$_{i,t}$			0.00038***	0.00019***			0.00021***	0.00006**
			(0.000)	(0.000)			(0.000)	(0.000)
ln(Knowledge Stock)$_{i,t}$		0.00058***		0.00059***		0.00048***		0.00048***
		(0.000)		(0.000)		(0.000)		(0.000)
Observations	49,800	49,800	49,800	49,800	48,140	48,140	48,140	48,140
R-squared	0.30445	0.36312	0.30199	0.36230	0.24410	0.28404	0.24458	0.28425
Field FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
No. of Fields	332	332	332	332	332	332	332	332

Notes. This table reports OLS estimates of the effect of breakthrough innovations on future Katz centrality at horizons of 1 year (Columns 1–4) and 10 years (Columns 5–8). The dependent variable is field i's Katz centrality in year $t+1$ (Columns 1–4) or $t+10$ (Columns 5–8). The key explanatory variables are the log count of breakthrough innovations in field i accumulated over the preceding ten years, measured using a 95th-percentile novelty threshold (Columns 1, 2, 5, 6) or a 90th-percentile threshold (Columns 3, 4, 7, 8) following Kelly et al. (2021). Even-numbered columns additionally control for the log cumulative knowledge stock of field i at time t. Breakthrough innovations are defined as documents in the global top 5% (or top 10%) of the distribution of future-to-past textual similarity ratios, after demeaning by calendar year to account for secular language evolution and then ranking globally across all fields and years. All specifications include field and year fixed effects; standard errors (in parentheses) are clustered at the field level. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively. The sample spans all 332 fields (209 science fields and 123 technology classes) over the estimation window.

Table 3 provides complementary regression evidence on the relationship between breakthrough innovations and innovation centrality at horizons of 1 year (Columns 1–4) and 10 years (Columns 5–8). Across all specifications, the coefficients on the log count of breakthrough innovations—whether measured at the 95th or 90th percentile of the importance distribution—are positive, statistically significant, and economically meaningful. The results remain robust after controlling for the existing knowledge stock, indicating that breakthrough innovations exert an independent effect on centrality beyond merely reflecting the prior accumulation of documented knowledge. Intuitively, a field's knowledge stock measures the volume of prior work, while breakthrough innovations capture the creation of genuinely novel, future-shaping ideas.

The estimated coefficients are also broadly stable across forecast horizons. The coefficient on log(Breakthrough, 95%) roughly halves from 0.00039 at the 1-year horizon (Column 1) to 0.00021 at the 10-year horizon (Column 5) but remains positive and statistically significant, indicating that the centrality gains from breakthrough innovations are not purely transient but persist a decade later. This persistence is consistent with the long-run cyclicality documented in Figure 7: breakthroughs initiate a protracted upward phase in a field's centrality cycle rather than a brief and quickly reversed spike. These results underscore the economically meaningful and durable role of breakthrough innovation in driving the structural evolution of the knowledge diffusion network.

5. A Framework for Estimating the Value of Science

This section develops a framework for measuring the social value of R&D across sectors connected by knowledge spillovers. We consider a general environment in which the innovation network—the matrix of cross-sector knowledge spillovers—is time-varying and not restricted to any parametric form. Our main results characterize the social value of R&D in each sector as a function of this network. We first study a deterministic environment in which the network's future path is known, and later extend the framework to stochastic regime changes (Section 5.3).

5.1. Setup

Preferences and Production. Consider an economy with N sectors, indexed by $i = 1, \dots, N$. There is a representative consumer with log flow utility and exponential discounting at rate ρ:

$$V_t = \int_t^\infty e^{-\rho(s-t)} \ln y_s \, ds, \qquad y_t = \prod_{i=1}^N y_{it}^{\beta_i}, \quad \sum_{i=1}^N \beta_i = 1. \tag{4}$$

The consumption good at each time t is a Cobb-Douglas aggregator over sectoral goods $\{y_{it}\}_{i=1}^N$. We refer to β_i as the consumption elasticity of sector i, with $\beta_i \geq 0$ and $\sum_i \beta_i = 1$.

Each sectoral good i is produced from production workers ℓ_{it} with production function $y_{it} = q_{it}^{\psi}\ell_{it}$, where q_{it} is the sector's knowledge stock at time t and ψ parametrizes how knowledge translates into productivity. Substituting the production function into (4):

$$V_t = \int_t^{\infty} e^{-\rho(s-t)} \left(\psi \sum_{i=1}^{N} \beta_i \ln q_{is} + \sum_{i=1}^{N} \beta_i \ln \ell_{is} \right) ds. \tag{5}$$

The sectoral knowledge stocks $\{q_{it}\}_{i=1}^{N}$ are the state variables of the economy.

The Innovation Process. R&D expands the knowledge stock. At time t, mass x_{it} of R&D resources employed in sector i generates a flow of new innovation n_{it}:

$$n_{it} = x_{it}\, \mathcal{O}_i\Big(\{\ln q_{jt}\}_{j=1}^{N}\Big)\,. \tag{6}$$

$\mathcal{O}_i(\cdot)$ is a knowledge aggregation operator that captures how the existing stock of knowledge across all sectors feeds into innovation in sector i. We impose two assumptions on the operators $\mathcal{O}_i$. First, $\mathcal{O}_i$ is homogeneous of degree one; that is, knowledge aggregation exhibits constant returns to scale. Second, the cross-partial elasticities are non-negative: more knowledge in any sector weakly increases innovation in every sector. These two assumptions are maintained throughout the analysis.

New innovation n_{it} expands the knowledge stock according to the law of motion

$$\dot{q}_{it}/q_{it} = \lambda \ln(n_{it}/q_{it}). \tag{7}$$

The key distinction between n_{it} and q_{it} is that the former is a flow variable capturing innovation output at time t, whereas the latter is a stock variable reflecting the accumulation of past innovations. The rate at which knowledge expands in sector i is increasing in the flow of innovation and decreasing in the existing knowledge stock q_{it}, capturing the notion that innovation becomes harder as the knowledge frontier in sector i advances. The parameter λ governs the elasticity with which flow R&D expands the knowledge stock. Because past knowledge affects future R&D efficiency, λ also relates to the rate at which knowledge spillovers materialize.

Substituting (6) into (7) yields the combined knowledge dynamics:

$$\frac{d\ln q_{it}}{dt} = \lambda\left[\ln x_{it} + \ln\mathcal{O}_i\Big(\{\ln q_{jt}\}_{j=1}^N\Big) - \ln q_{it}\right], \quad q_{i0} \text{ given.} \tag{8}$$

The Innovation Network. We define the innovation network as the matrix of knowledge spillover elasticities:

Definition 1 (Innovation Network)**.** *The innovation network at time t is the $N \times N$ matrix $\boldsymbol{\Omega}(t)$ with entries*

$$\Omega_{ij}(t) \equiv \frac{\partial\ln\mathcal{O}_i\Big(\{\ln q_{kt}\}_{k=1}^N\Big)}{\partial\ln q_{jt}}. \tag{9}$$

The entry $\Omega_{ij}(t)$ measures the elasticity of sector i's knowledge aggregation with respect to sector j's knowledge stock. By the non-negativity assumption, $\Omega_{ij}(t) \geq 0$ for all i, j, t. By constant returns to scale, rows sum to one: $\sum_{j=1}^N \Omega_{ij}(t) = 1$ for all i and t. The innovation network is allowed to vary over time, reflecting the evolving structure of knowledge flows across sectors.

The matrix $\boldsymbol{\Omega}(t)$, a key object of this study, represents a weighted directed graph in which economic sectors are the nodes. The entry $\Omega_{ij}(t)$ captures the degree to which sector i's innovation benefits from sector j's existing knowledge stock. We refer to sector j as upstream of sector i and, conversely, i as downstream of j; this terminology captures the direction of knowledge flows from source to recipient. Absent cross-sector knowledge spillovers, $\boldsymbol{\Omega}(t) = \boldsymbol{I}$ is the identity matrix. The concept is not tied to any specific sector definition; innovation networks can be constructed across industrial sectors, technology classes, and scientific fields.

Interpreting Science. Among the N sectors, some may represent scientific fields—disciplines such as mathematics, chemistry, or molecular biology—that do not directly produce consumption goods. For these sectors, the consumption elasticity is zero: $\beta_k = 0$. Their contribution to the economy arises entirely through knowledge spillovers, captured by the off-diagonal entries of $\boldsymbol{\Omega}(t)$. As we show below, such sectors can nonetheless have high social value.

5.2. The Social Value of R&D

We now characterize the relative social value of R&D across sectors. The central question is: holding the production resource allocation $\{\ell_{it}\}$ and the R&D allocation in all other sectors and all future periods fixed, how much does welfare rise from a marginal increase in R&D directed to sector i at time t?

Our framework deliberately leaves the resource cost of R&D unspecified at this stage. R&D may draw on the final good, on production workers, or on a separate endowment of scientists—the relevant specification depends on the institutional setting. The implicit assumption is that whatever the resource constraint, an increase in R&D in any sector carries a shadow cost that, in equilibrium or at an optimum, is common across sectors at each point in time. This common shadow cost determines the absolute level of R&D value but cancels when comparing across sectors. We can therefore derive the relative social value of R&D across sectors without taking a stand on how R&D is financed or on the level of the shadow cost. Later, in the quantitative analysis, we close the model by specifying a resource constraint and calibrate the absolute level to match existing estimates in the literature.

To formalize, fix a time t and consider a given R&D allocation $\{x_{is}\}_{s\geq t}$, with the production allocation $\{\ell_{is}\}$ held fixed. We perturb R&D in the direction $\boldsymbol{h} = (h_1, \ldots, h_N) \in \mathbb{R}^N$ with amplitude $\varepsilon \geq 0$, active only on a short time window $[t, t+\delta)$ of width $\delta > 0$:

$$x_{is}^{\varepsilon,\delta,\boldsymbol{h}} \equiv x_{is} + \varepsilon \cdot h_i \cdot \mathbf{1}_{t\leq s<t+\delta}. \tag{10}$$

The perturbed R&D feeds through the knowledge dynamics (8), generating a perturbed knowledge path $\{\boldsymbol{q}_s^{\varepsilon,\delta,\boldsymbol{h}}\}_{s\geq t}$. Let $V_t^{\varepsilon,\delta,\boldsymbol{h}}$ denote the resulting welfare.

Definition 2 (Marginal Value of R&D)**.** *The marginal social value of R&D in sector i at time t is*

$$\xi_{it} \equiv \lim_{\delta\to 0^+} \frac{1}{\delta} \cdot \left. \frac{\partial V_t^{\varepsilon,\delta,\boldsymbol{h}}}{\partial \varepsilon} \right|_{\varepsilon=0,\, \boldsymbol{h}=\boldsymbol{e}_i}, \tag{11}$$

where $\boldsymbol{e}_i \in \mathbb{R}^N$ is the i-th standard basis vector.

The definition decomposes into two steps, which makes clear that ξ_{it} is a genuine marginal object. First, $\partial V_t^{\varepsilon,\delta,\boldsymbol{h}}/\partial\varepsilon$ evaluated at $\varepsilon = 0$ is a Gateaux derivative in the amplitude direction: it linearizes the welfare response along $\boldsymbol{h}$. Because $\ln(x_{is} + \varepsilon h_i \mathbf{1}_{t \le s < t+\delta})$ differentiates in ε at $\varepsilon = 0$ to $h_i \mathbf{1}_{t\le s<t+\delta}/x_{is}$, this linearization produces the $1/x_{it}$ factor that will appear in Proposition 1 below. Second, the limit $\lim_{\delta\to 0^+}(1/\delta)$ localizes the perturbation to an instantaneous spike at time t, yielding welfare sensitivity per unit time of R&D injection at that instant. Separating amplitude ε from window width δ clarifies that ξ_{it} is a first-order derivative in ε—not the welfare value of a discrete lump of R&D divided by its duration. Since the production allocation $\{\ell_{is}\}$ is held fixed, writing welfare (5) under the perturbed path gives

$$V_t^{\varepsilon,\delta,\boldsymbol{h}} = \int_t^\infty e^{-\rho(s-t)} \left(\psi \sum_{j=1}^N \beta_j \ln q_{js}^{\varepsilon,\delta,\boldsymbol{h}} + \sum_{j=1}^N \beta_j \ln \ell_{js} \right) ds.$$

The labor terms do not depend on $(\varepsilon, \boldsymbol{h})$ and vanish upon differentiation, so the marginal value of R&D depends only on how the knowledge stocks respond to the R&D injection:

$$\xi_{it} = \lim_{\delta\to 0^+} \frac{1}{\delta} \int_t^\infty e^{-\rho(s-t)}\, \psi \sum_{j=1}^N \beta_j \left. \frac{\partial \ln q_{js}^{\varepsilon,\delta,\boldsymbol{h}}}{\partial \varepsilon} \right|_{\varepsilon=0,\, \boldsymbol{h}=\boldsymbol{e}_i} ds. \tag{12}$$

Equation (12) makes the mechanics transparent. Boosting R&D in sector i during the window $[t, t+\delta)$ directly raises sector i's knowledge. This knowledge then spills over to other sectors through the innovation network $\boldsymbol{\Omega}$, raising their knowledge stocks in turn. The marginal value ξ_{it} aggregates these effects across all sectors j, weighted by the importance of each technology in the consumption bundle (β_j) and discounted over time.

Our first main result provides a closed-form characterization of the relative marginal value of R&D across sectors. We introduce the knowledge propagation matrix $\boldsymbol{\Phi}(t, s)$, which tracks how a perturbation to knowledge at time t ripples through the innovation network to affect knowledge at a later time $s \ge t$.

Proposition 1 (Marginal and Total Value of R&D)**.** *Under constant returns to scale in the knowledge aggregation operators $\{\mathcal{O}_i\}$, the marginal social value of R&D satisfies*

$$\xi_{it} \propto_t \frac{\gamma_{it}}{x_{it}}, \tag{13}$$

where $\propto_t$ denotes equality up to a positive constant that is specific to time t and common *across sectors, and* $\boldsymbol{\gamma}_t = (\gamma_{1t}, \ldots, \gamma_{Nt})'$ *is defined by*

$$\boldsymbol{\gamma}_t' = \int_t^\infty \boldsymbol{\beta}' \, \boldsymbol{\Phi}(t,s) \, ds, \tag{14}$$

and the propagation matrix $\boldsymbol{\Phi}(t,s)$ *solves the matrix differential equation*

$$\frac{\partial \boldsymbol{\Phi}(t,s)}{\partial s} = -(\rho+\lambda)\left(\boldsymbol{I} - \frac{\boldsymbol{\Omega}(s)}{1+\rho/\lambda}\right)\boldsymbol{\Phi}(t,s), \quad \boldsymbol{\Phi}(t,t) = \boldsymbol{I}. \tag{15}$$

Proof. See Appendix A.2. □

Proposition 1 decomposes the marginal value of R&D in sector i at time t into two forces. The numerator γ_{it} is the present-discounted welfare impact of knowledge in sector i at time t, integrating all the ways that knowledge propagates through the innovation network over time. The denominator x_{it} reflects diminishing returns: sectors that already receive large R&D investment see smaller proportional knowledge gains from additional resources. Since $\xi_{it} \propto_t \gamma_{it}/x_{it}$, we can write $\gamma_{it} \propto_t \xi_{it} \cdot x_{it}$—the marginal value of R&D times the quantity of R&D resources. We can therefore interpret γ_{it} as the total social value of R&D in sector i at time t. This is the object we take to the data: in the empirical analysis, we refer to γ_{it} as the *economic value* of R&D in field i, and we use the two terms interchangeably.

The total social value γ_{it} captures the full causal chain from knowledge to welfare. An increase in q_{it} raises innovation in all sectors j for which $\Omega_{ji} > 0$, which in turn raises q_{js} for future $s > t$, which feeds back into further innovation, and so on. The propagation matrix $\boldsymbol{\Phi}(t,s)$ encodes this entire chain, discounted at the effective rate $\rho + \lambda$, with the innovation network $\boldsymbol{\Omega}(s)$ governing the spillover structure at each point in time. The integral in (14) is guaranteed to converge under our maintained assumptions: because $\boldsymbol{\Omega}(s)$ is row-stochastic for all s, the propagation matrix decays exponentially, $\|\boldsymbol{\Phi}(t,s)\|_\infty \leq e^{-\rho(s-t)}$, so the cumulative effect of knowledge spillovers remains bounded.[11] No additional assumptions—such as diagonalizability or commutativity of $\{\boldsymbol{\Omega}(s)\}$—are required.

An important feature of Proposition 1 is that the formula involves the observed innovation

[11] Formally, this follows from the Coppel inequality: the maximum row sum of the drift matrix $\boldsymbol{A}(s) \equiv (\rho+\lambda)(\boldsymbol{I} - \boldsymbol{\Omega}(s)/(1+\rho/\lambda))$ equals ρ for all s because $\boldsymbol{\Omega}(s)$ is row-stochastic, which implies the exponential bound on $\boldsymbol{\Phi}$. Moreover, since $-\boldsymbol{A}(s)$ has non-negative off-diagonal entries, $\boldsymbol{\Phi}(t,s) \geq 0$ entry-wise, guaranteeing $\gamma_{it} \geq 0$ whenever $\beta_i \geq 0$.

network $\boldsymbol{\Omega}(s)$ along the unperturbed path, with no adjustment for how the network itself might respond to the R&D perturbation. To see why, note that a change in R&D at time t perturbs the knowledge path $\boldsymbol{q}_s$ for $s > t$, which shifts the spillover elasticities $\boldsymbol{\Omega}(s)$ since they depend on knowledge stocks. However, this network response is itself proportional to the amplitude ε, so its welfare effect is second-order in ε. Since ξ_{it} is a first-order derivative in ε, it does not require us to model how the network evolves endogenously in response to R&D policy.

The Log-Linear Special Case. When the knowledge aggregation operators are Cobb-Douglas, $\mathcal{O}_i(\{\ln q_{jt}\}) = \prod_j q_{jt}^{\omega_{ij}}$, and $\boldsymbol{\Omega}$ is time-invariant, the propagation matrix admits the closed form

$$\boldsymbol{\Phi}(t,s) = \exp\left[-(\rho+\lambda)\left(\boldsymbol{I} - \frac{\boldsymbol{\Omega}}{1+\rho/\lambda}\right)(s-t)\right]. \tag{16}$$

This is the framework developed in Liu and Ma (2021). The total social values become time-invariant and take the Leontief inverse form

$$\boldsymbol{\gamma}' \propto \boldsymbol{\beta}' \left(\boldsymbol{I} - \frac{\boldsymbol{\Omega}}{1+\rho/\lambda}\right)^{-1}, \tag{17}$$

and the optimal allocation satisfies $\boldsymbol{x}^* \propto_t \boldsymbol{\gamma}$. The Leontief inverse decomposes social value into the consumer's direct valuation of each technology β_i and the contributions of successively longer chains of knowledge spillovers, each discounted at rate ρ/λ per round of spillovers. Note that equation (17) is precisely the Katz centrality of the innovation network (with attenuation factor $1/(1+\rho/\lambda)$), weighted by the consumption vector $\boldsymbol{\beta}$. Intuitively, sectors that are more central in the knowledge diffusion process—those that influence many other sectors both directly and through long chains of spillovers—have higher social value. When the innovation network is time-varying, the discrete-time analogue of the Leontief inverse adds a time subscript to each round of the network effect:

$$\boldsymbol{\gamma}_t' \propto_t \boldsymbol{\beta}' \left(\boldsymbol{I} + \frac{\boldsymbol{\Omega}_{t+1}}{1+\rho/\lambda} + \frac{\boldsymbol{\Omega}_{t+2}\,\boldsymbol{\Omega}_{t+1}}{(1+\rho/\lambda)^2} + \frac{\boldsymbol{\Omega}_{t+3}\,\boldsymbol{\Omega}_{t+2}\,\boldsymbol{\Omega}_{t+1}}{(1+\rho/\lambda)^3} + \cdots\right). \tag{18}$$

The first term captures the direct impact of better technology on consumption; the second reflects how improved technology at time t raises the productivity of R&D at $t+1$, which in turn benefits consumption; successive terms capture higher rounds of network effects at $t+2$, $t+3$, and so on, each discounted at rate ρ/λ. Proposition 1 is the continuous-time analogue of (18): the propagation matrix $\mathbf{\Phi}(t,s)$ in (14) plays the role of the discounted product $\mathbf{\Omega}_s \cdots \mathbf{\Omega}_{t+1}/(1+\rho/\lambda)^{s-t}$, and the integral replaces the sum. The value of investing in a sector today depends not only on the current network but on its entire future trajectory. In the empirical analysis (Section 6), we implement the general time-varying framework, extended to stochastic regime changes in Section 5.3; as we document there, the innovation network $\mathbf{\Omega}(t)$ changes substantially over our sample period, making the time-varying formulation essential for capturing the evolving value of science.

Our second result characterizes the optimal allocation of R&D across sectors, under the assumption that R&D resources are fungible: an additional unit of R&D has the same marginal cost regardless of the sector in which it is deployed.

Proposition 2 (Optimal R&D Allocation)**.** *Suppose R&D resources are fungible across sectors. A necessary condition for the R&D allocation $\{x_{it}\}$ to maximize welfare is*

$$x_{it} \propto_t \gamma_{it}. \tag{19}$$

Equivalently, the optimal R&D shares satisfy $x_{it}/\sum_j x_{jt} = \gamma_{it}/\sum_j \gamma_{jt}$ for all i and t.

Proof. This follows immediately from equating the marginal values ξ_{it} across sectors, which requires γ_{it}/x_{it} to be equal for all i. See Appendix A.2 for the full derivation. □

Proposition 2 says that the optimal share of R&D directed to sector i equals the share of sector i's total social value in the economy-wide total. The vector $\boldsymbol{\gamma}_t$ therefore serves as a sufficient statistic for the optimal R&D allocation: it summarizes all the information in the innovation network $\mathbf{\Omega}(\cdot)$ and the preference weights $\boldsymbol{\beta}$ that is relevant for R&D policy.

The Value of Science. A key implication concerns sectors with $\beta_k = 0$—scientific fields that do not directly produce consumption goods. Although such sectors receive zero direct weight in utility, they can still have high total social value γ_{kt} whenever their knowledge spills over to sectors that produce goods. The total social value of a scientific field is governed by three forces: the strength of its knowledge spillovers to downstream sectors, the number of intermediate steps in the spillover chain (with longer chains discounted more heavily), and the consumption importance of the receiving sectors. Proposition 2 implies that the optimal allocation directs positive R&D resources to science, in proportion to these indirect contributions.

More broadly, the ratio ρ/λ governs how the optimal allocation trades off short-run consumption value against long-run growth potential. When ρ/λ is large, the planner heavily discounts future spillovers and the optimal allocation is close to the direct consumption weights $\boldsymbol{\beta}$. When ρ/λ is small, the planner values long-run knowledge growth and places more weight on sectors that are central in the innovation network—even if those sectors have low direct consumption value.

5.3. Uncertainty and Regime Changes

The analysis so far treats the future path of the innovation network as a deterministic process. However, evidence in Section 4 shows that breakthrough discoveries episodically reorganize and redraw the innovation network, elevating new fields and displacing established ones. In this subsection, we extend the framework to recurrent, random structural change.

Formally, suppose the economy is subject to recurrent structural shocks arriving according to a Poisson process with intensity $\eta > 0$. At each arrival, the knowledge aggregation operators (and hence the innovation network) are redrawn: the new profile $\{\hat{\mathcal{O}}_i\}$ is drawn from a transition kernel $\mu(\cdot \mid \mathcal{O})$, which may depend on the profile $\mathcal{O}$ prevailing at the arrival, and whose support contains only profiles satisfying the maintained assumptions of constant returns to scale and non-negative cross-partial elasticities.[12] The arrival times and the shocks are mutually independent and independent of the R&D allocation. We refer to the stretch between consecutive arrivals, during which a fixed operator profile prevails, as a *regime*, and to each arrival as a *regime change*.

[12] Dependence on the prevailing profile accommodates persistent technologies. For instance, the geometric mixture $\hat{\mathcal{O}}_i = \mathcal{O}_i^{1-\theta}\,(\mathcal{O}_i^{\varepsilon})^{\theta}$ of the current aggregator with a random aggregator $\mathcal{O}_i^{\varepsilon}$ preserves both maintained assumptions, with θ measuring how much of the technology is redrawn.

The model extension thus decomposes network change into two parts: an endogenous part, operating within each regime through the dependence of $\mathbf{\Omega}$ on the knowledge stocks, and an exogenous part—the structural shocks—capturing shifts in how knowledge feeds innovation whose timing and content are not governed by the R&D process itself (e.g., Arrow, 1962; Rosenberg, 1998). Knowledge stocks are continuous at each regime change; what changes is the technology through which knowledge generates innovation, and hence the network. For a realization ω of the arrival times and draws, let $V_t^{\varepsilon,\delta,\boldsymbol{h}}(\omega)$ denote realized welfare; the planner's objective is its expectation over the arrival process and the draws, neither of which is known at time t.

The marginal value of R&D is then defined exactly as in Definition 2, applied to this objective. The perturbation changes only the R&D injected on the window $[t, t+\delta)$; the continuation allocation—which may itself be a plan contingent on the history of regimes—is held fixed, and the continuation values below are computed under this plan.[13]

The value of R&D again has two parts: the value accruing under the current regime up to the first change, plus the expected continuation value of the knowledge then in place. Because further regime changes follow the first, the characterization is recursive.

Proposition 3 (The Value of R&D under Recurrent Regime Changes)**.** *Under the maintained assumptions, the marginal social value of R&D satisfies $\xi_{it} \propto_t \gamma_{it}/x_{it}$, where the value vector solves*

$$\boldsymbol{\gamma}'_t \;=\; \underbrace{\int_t^\infty e^{-\eta(s-t)}\,\boldsymbol{\beta}'\,\mathbf{\Phi}(t,s)\,ds}_{\textit{value under the current regime}} \;+\; \underbrace{\eta\int_t^\infty e^{-\eta(T-t)}\,\bar{\boldsymbol{\gamma}}'_T\,\mathbf{\Phi}(t,T)\,dT}_{\textit{expected continuation value}}, \qquad \bar{\boldsymbol{\gamma}}'_T \;\equiv\; \int \boldsymbol{\gamma}'_T(\hat{o})\,\mu(d\hat{o}\mid\mathcal{O}), \tag{20}$$

where $\mathbf{\Phi}(t,s)$ solves the propagation equation (15) with the network induced by the current operator profile along the no-arrival path, and $\boldsymbol{\gamma}_T(\hat{o})$, the value vector at date T under a freshly drawn profile $\hat{o}$, satisfies the same equation from date T onward under $\hat{o}$, with the kernel then conditioning on $\hat{o}$. The recursive system has a unique bounded solution.

Proof. See Appendix A.2. □

[13] For fixed $\delta > 0$ the perturbation window may contain one or more arrivals; such realizations do not matter as we take the $\delta \to 0^+$ limit (see the proof in the appendix).

Intuitively, the first term is the value realized under the current regime, before the first change. It coincides with the baseline formula (14) except that the discount rate rises from ρ to $\rho + \eta$:[14] each round of spillovers is further discounted by the probability $e^{-\eta(s-t)}$ that the regime survives to carry it, making long spillover chains less valuable. The second term is the expected continuation value: knowledge created by today's R&D propagates according to $\boldsymbol{\Phi}(t, T)$ until the regime changes, and is then valued at the average post-change value vector $\bar{\boldsymbol{\gamma}}_T$, which—with further regime changes ahead—solves the same two-part equation. Only the current regime's propagation matrix appears in (20)—later regimes enter through the recursion—and all randomness has been integrated out against the survival probability, the arrival density, and μ, so every object in (20) is deterministic.

A discrete-time analogue makes the structure transparent. Let $\{\tilde{\boldsymbol{\Omega}}_{t+j}\}_{j\geq 1}$ denote the realized network path: the current regime's networks up to the first regime change, the new regime's thereafter, and so on. Then (20) becomes

$$\boldsymbol{\gamma}_t' \;\propto_t\; \boldsymbol{\beta}' + \frac{\mathbb{E}_t\left[\boldsymbol{\gamma}_{t+1}'\,\tilde{\boldsymbol{\Omega}}_{t+1}\right]}{1+\rho/\lambda} \;=\; \mathbb{E}_t\left[\boldsymbol{\beta}'\left(\boldsymbol{I} + \frac{\tilde{\boldsymbol{\Omega}}_{t+1}}{1+\rho/\lambda} + \frac{\tilde{\boldsymbol{\Omega}}_{t+2}\,\tilde{\boldsymbol{\Omega}}_{t+1}}{(1+\rho/\lambda)^2} + \cdots\right)\right]. \tag{21}$$

Equation (21) first expresses $\boldsymbol{\gamma}_t$ recursively: today's value equals the direct consumption dividend plus the discounted expected value of next period's knowledge, valued at next period's—possibly redrawn—network. Iterating the recursion forward yields the second expression: the expectation, over arrival dates and operator draws, of the deterministic power series (18) evaluated along the realized network path. The value of R&D today thus depends on the expectation of the network's future path, providing the formal foundation for treating the empirical valuation as a forecasting problem.

5.4. From Model to Data

Our results characterize the total social value of R&D in each sector. Because the data are in discrete time, we work with the discrete-time version of the formula, equation (21), which expresses the value of R&D as the expectation of a discounted power series that compounds

[14] To see this, note that multiplying $\boldsymbol{\Phi}(t,s)$ by $e^{-\eta(s-t)}$ adds $-\eta\,\boldsymbol{\Phi}(t,s)$ to the right-hand side of equation (15): the product solves (15) with $\rho+\eta$ in place of ρ.

the network matrices at successive future dates. Computing the social value thus involves computing this expectation over the network's future path. With $N = 332$ fields, the network has $N^2 \approx 110{,}000$ entries *per year*—an infeasible forecasting task.

We address this challenge by reducing the dimensionality of the forecasting problem through a joint diagonalization (JD) of the innovation network. Let $\boldsymbol{\Omega}_t$ denote the $N \times N$ matrix of innovation linkages across N fields in year t. JD seeks a shared K-dimensional basis, with $K \leq N$, such that each yearly network can be approximated by

$$\boldsymbol{\Omega}_t \approx \boldsymbol{U}\,\boldsymbol{\Sigma}(t)\,\boldsymbol{M}, \tag{22}$$

where $\boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_K] \in \mathbb{R}^{N\times K}$ collects K time-invariant basis vectors capturing common knowledge structure, $\boldsymbol{\Sigma}(t) = \operatorname{diag}(\sigma_1(t), \ldots, \sigma_K(t)) \in \mathbb{R}^{K\times K}$ is a diagonal matrix of time-varying importance weights, and $\boldsymbol{M} = [\boldsymbol{m}_1, \ldots, \boldsymbol{m}_K]' = (\boldsymbol{U}'\boldsymbol{U})^{-1}\boldsymbol{U}' \in \mathbb{R}^{K\times N}$ is the Moore-Penrose pseudo-inverse of $\boldsymbol{U}$, satisfying $\boldsymbol{M}\boldsymbol{U} = \boldsymbol{I}_K$; the columns $\boldsymbol{u}_k$ and the rows $\boldsymbol{m}_k'$ are, respectively, right and left eigenvectors of every network of the form (22). Note that K may equal N: the main dimensionality reduction comes from the shared-basis structure itself, which collapses the time-varying part of the network from N^2 entries per date to K eigenvalue series, rather than from $K < N$ (Section 6.1).

The kth column $\boldsymbol{u}_k$ summarizes how each field loads on common knowledge dimension k. We refer to each such dimension as a "knowledge topic"—a linear combination of the original 332 fields that captures a coherent cluster of related knowledge. We assume that the aggregators induce networks of this form at every level of the knowledge stocks, so that a regime change replaces the eigenvalue functions while leaving the topics fixed; this is consistent with the stable basis our estimation recovers over a sample spanning many historical episodes of structural change. Section 6.1 details the estimation procedure and the empirical performance of the JD approximation.

A key algebraic consequence of the JD structure is that products of networks at successive dates collapse topic by topic. This implies that the expectation of the matrix power series in equation (21) reduces to K scalar expectations, one per knowledge topic:

$$\boldsymbol{\gamma}_t' \propto_t \sum_{k=1}^{K} (\boldsymbol{\beta}'\boldsymbol{u}_k)\, \mathbb{E}_t\left[1 + \frac{\sigma_k(t+1)}{1+\rho/\lambda} + \frac{\sigma_k(t+1)\,\sigma_k(t+2)}{(1+\rho/\lambda)^2} + \cdots\right] \boldsymbol{m}_k'. \tag{23}$$

This expression captures the three components that determine the social value of knowledge in each field. The term $\boldsymbol{\beta}'\boldsymbol{u}_k$ measures how strongly the consumption bundle depends on knowledge topic k—topics that load on sectors producing important consumer goods receive higher weight. The bracketed expectation summarizes the expected cumulative spillover potential of topic k, aggregating how its eigenvalue evolves over the entire future path: topics with persistently high eigenvalues contribute more to the social value. The term $\boldsymbol{m}_k'$ maps the topic-level value back to individual fields, distributing the social value according to each field's contribution to the topic. Computing the social value thus requires forecasting only the K eigenvalue series rather than the N^2 network entries: the joint diagonalization serves, first and foremost, as a dimensionality reduction for the forecasting problem.

6. Estimates of the Economic Value of Science

We now move to quantifying the economic value of science. Recall that a field's economic value is the object γ_{it} characterized in Section 5. We first apply the joint diagonalization to reduce the network to a small number of forecastable eigenvalue series and assess their predictability (Section 6.1). We then combine the forecasts into forward-looking estimates of each field's economic value (Section 6.2), decompose where this value is realized (Section 6.3), and validate the estimates against fields' real outcomes (Section 6.4).

6.1. Joint Diagonalization and the Structure of the Innovation Network

The first step is to apply the JD decomposition outlined in Section 5.4 (Equation (22)). Beyond simplifying the forecasting problem, the JD representation serves a broader purpose: it allows us to reconstruct the innovation network organically by linearly combining fields, circumventing the problem that scientific and technological fields are delineated by a rigid classification system.

We estimate the time-invariant basis $\boldsymbol{U}$ (whose K columns are the "knowledge topics") and the eigenvalue paths $\{\boldsymbol{\Sigma}(t)\}$ jointly. The design of the objective reflects how the representation is ultimately used. The economic value formula depends not on the individual bilateral flows, but on the network's centrality structure—the ranking of fields by upstream influence cumulated along all indirect paths. We therefore use a two-part objective that targets both local flows and global centrality. Denoting $\hat{\boldsymbol{\Omega}}_t = \boldsymbol{U}\boldsymbol{\Sigma}(t)\boldsymbol{M}$ as the JD-reconstructed network, the estimator solves

$$\min_{\boldsymbol{U},\{\boldsymbol{\Sigma}(t)\}} \frac{w_\Omega}{T}\sum_t \left\|\hat{\boldsymbol{\Omega}}_t - \boldsymbol{\Omega}_t\right\|_F^2 + \frac{(1-w_\Omega)N}{2}\sum_t \left[\left\|\mathrm{Katz}^{\mathbb{1}}(\hat{\boldsymbol{\Omega}}_t) - \mathrm{Katz}_t^{\mathbb{1}}\right\|^2 + \left\|\mathrm{Katz}^{\mathrm{ew}}(\hat{\boldsymbol{\Omega}}_t) - \mathrm{Katz}_t^{\mathrm{ew}}\right\|^2\right].$$

The first term ensures that the low-dimensional representation faithfully reconstructs the observed bilateral knowledge flows, while the second term preserves the network's global centrality structure. The two centrality anchors are computed from the observed network at the economic decay factor $\delta^\star = 0.95$, the per-round spillover discount $1/(1+\rho/\lambda)$ used in the value formula (Section 6.2): a uniform-weighted Katz centrality $\mathrm{Katz}_t^{\mathbb{1}}$, capturing each field's overall upstream influence, and a technology-weighted value $\mathrm{Katz}_t^{\mathrm{ew}}$, which replaces the uniform seed with the equal-weight indicator of technology fields and so isolates the influence flowing toward the patent fields closest to production. We implement simple reparameterizations and rescaling to enforce the $(0,1)$-eigenvalue constraints and normalize the columns of $\boldsymbol{U}$. We solve the resulting program by stochastic gradient descent with the Adam optimizer and backpropagation (Kingma and Ba, 2014; Rumelhart, Hinton and Williams, 1986).

Choosing K and w_Ω Implementing the estimator requires specifying the number of knowledge topics K and the objective weight w_Ω. The main dimensionality reduction comes from the JD structure itself: even with a full-rank basis ($K = N$), representing the network with a time-invariant $\boldsymbol{U}$ and a diagonal $\boldsymbol{\Sigma}(t)$ reduces the time-varying objects from $N \times N$ network entries per year to N eigenvalues. Capturing most of the network's systematic variation, however, does not require a full-rank basis: sweeping over K, we find that the fit rises steeply as K increases toward 150 and plateaus thereafter. For the weight w_Ω, the centrality correlation becomes high (above 0.9) as soon as w_Ω moves below one, so we place most of the weight on reconstructing the bilateral flows. Our preferred specification sets $K = 150$ and $w_\Omega = 0.90$.[15]

Interpreting $\boldsymbol{U}$ As discussed in Section 5.4, the eigenvectors resulting from the joint diagonalization have a clear economic interpretation. Each eigenvector (i.e., a column of $\boldsymbol{U}$) defines a "knowledge topic", characterized by its loadings on different fields. For instance, one topic loads heavily on parallel, distributed, and real-time computing, software engineering, computer hardware, and control engineering, which we label *Computer Science & Engineering*. A second topic is dominated by embedded systems, computer architecture, and computer hardware, and we interpret it as *Embedded Systems & Hardware*. A third loads strongly on internet privacy, the World Wide Web, and computer security, which we describe as *Internet Security & Privacy*. Additional topics pick out clusters such as *Particle Physics*—spanning particle, quantum, nuclear, and atomic physics—and *Combustion Engines*, dominated by internal combustion engines and related machinery.

These examples illustrate that the JD representation reorganizes the dense innovation network into a small number of economically meaningful knowledge topics with time-varying importance, providing a tractable and interpretable summary of the structure of innovation.

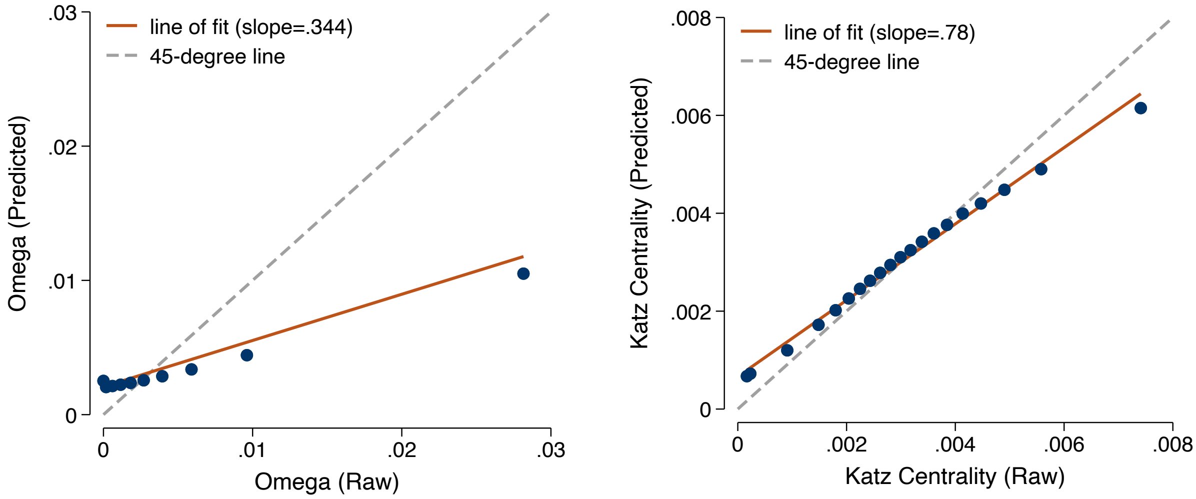
Performance of the Joint Diagonalization We now assess how well the joint diagonalization recovers the original network matrices. In our preferred specification, the correlation between the JD approximation and the observed data is above 0.9 for Katz centrality and approximately 0.52 for the network. Figure 9 illustrates the alignment between the JD approximation and the observed network and Katz centrality: the points lie close to a straight line, with slopes of 0.34 for the network and 0.78 for Katz centrality.

[15] Appendix Figure A.6 summarizes the associated model-selection procedure over the topic count K and the network weight w_Ω. We sweep K over $\{10, 20, \ldots, 332\}$ and w_Ω over $\{0.50, \ldots, 1.00\}$, reporting the correlation between the observed and reconstructed network and centrality objects for each combination. The network correlation rises steeply as K increases toward 150 and then plateaus near 0.524, indicating that 150 topics capture most of the network's systematic variation, while the centrality correlation becomes very high and greater than 0.9 as soon as w_Ω moves off one. Our preferred specification sets $K = 150$ topics and $w_\Omega = 0.90$, which delivers a good trade-off between parsimony and fit.

Figure 9. The Empirical Performance of Joint Diagonalization

Panel (a): Network

Panel (b): Katz Centrality



Notes. This figure compares the JD approximation with the observed data under the preferred specification ($K = 150$, $w_\Omega = 0.90$). Panel (a) plots the JD-reconstructed network entries against the observed entries; Panel (b) plots the JD-implied Katz centrality against the centrality of the observed network. The correlation between the JD approximation and the observed data is approximately 0.52 for the network entries and above 0.9 for Katz centrality, with fitted slopes of 0.34 and 0.78, respectively.

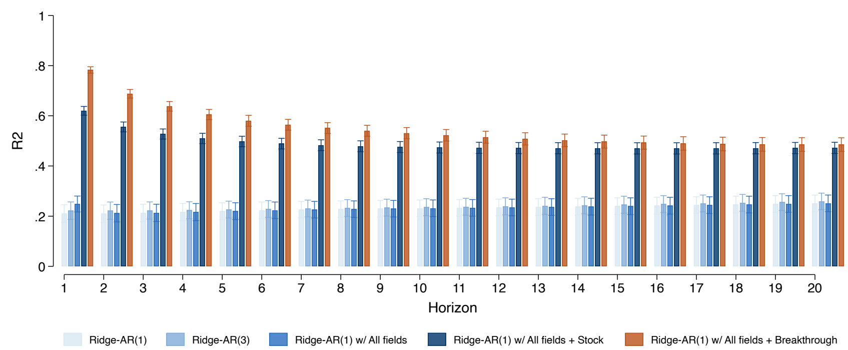
Predictability of Joint-Diagonalized Innovation Network The JD reduces the problem of predicting future innovation networks from forecasting every entry of the network to forecasting the eigenvalues, cutting the dimension from $N \times N = 110{,}224$ ($N = 332$) to $K = 150$. Because the basis is time-invariant, over-time variation in the eigenvalues $\sigma_k(t)$ translates into variation in the spillover matrix: forecasts of $\sigma_k(t+j)$ recover the future spillover structure $\mathbf{\Omega}_{t+j}$.

We assess how well future eigenvalues can be predicted from their own history and from economic fundamentals. For each eigenvalue k, the forecasting model uses two kinds of information. The first is the past values of the eigenvalues, including both eigenvalue k's own history and those of the others. The second is a pair of economic predictors measured across all fields: the knowledge stock, defined as the log of the cumulative count of documents in a field, and breakthrough innovations, defined as the log of the count of a field's most novel documents (the top 5% by novelty) over the past ten years.

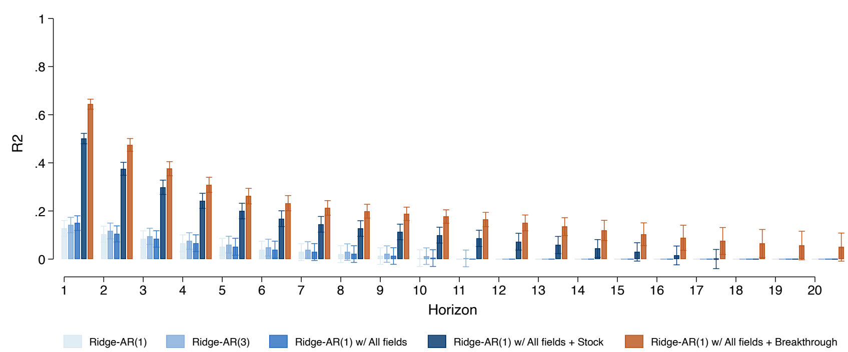
We estimate the model both in-sample and out-of-sample. The in-sample version estimates the parameters once on the full 1870–2024 sample. The out-of-sample version re-estimates them recursively at each origin year t using only data available up to t. In both cases, we forecast each eigenvalue up to 50 years ahead, and discard the first 30 years of the sample as a burn-in period to ensure the models are well-estimated.

Figure 10. Predictability of Eigenvalues

Panel (a): In-Sample R^2



Panel (b): Out-of-Sample R^2



Notes. This figure reports the predictability of the JD eigenvalues at horizons of 1 to 50 years under the preferred Ridge model (penalty 10^3). Panel (a) shows the in-sample R^2, with parameters estimated once on the full 1870–2024 sample; Panel (b) shows the out-of-sample R^2, with parameters re-estimated recursively at each origin year using only data available up to that year. At each horizon, the R^2 pools across all $K = 150$ eigenvalues. The specifications differ in the predictor set: the eigenvalues' own histories (autoregressive benchmark), adding the knowledge stocks of all fields, and further adding breakthrough innovation counts.

Because our value estimates are forward-looking, they are only as credible as our ability to forecast the network's future path, which the joint diagonalization reduces to forecasting the eigenvalues. We therefore study the forecast R^2 horizon by horizon, as the value formula integrates over the entire forecast horizon. At each horizon the R^2 pools across all eigenvalues, so it is variance-weighted: the eigenvalues that move the network most count most. We take the out-of-sample R^2 as our primary criterion and read the in-sample R^2 as an upper bound. Throughout, we benchmark every specification against a purely autoregressive model, so that the object of interest is the predictive power of economic fundamentals beyond the network's own persistence.

Figure 10 reports these R^2 curves for specifications that differ in the set of predictors, using our preferred Ridge model with a penalty parameter of 10^3, which regularizes the high-dimensional predictor set relative to the length of the annual series and attains the best out-of-sample R^2 among the alternatives we examine.[16]

We find that the eigenvalues are substantially predictable, and that economic fundamentals—knowledge stocks and breakthrough innovations—are the primary source of this predictability. A simple autoregressive specification captures about 20% of the in-sample variation ($R^2 \approx 0.20$), indicating modest but nontrivial persistence. Adding the knowledge stocks of all fields as exogenous predictors roughly triples the explanatory power, raising the in-sample R^2 to approximately 0.62 and the out-of-sample R^2 to about 0.50 at short horizons. Intuitively, the cumulative knowledge stock of a field summarizes the state of its research frontier, and changes in this stock predict shifts in the field's position within the innovation network. Adding the count of breakthrough documents as a predictor further improves performance: the in-sample R^2 reaches approximately 0.78 and the out-of-sample R^2 about 0.64 at short horizons, with meaningful gains at long horizons as well ($R^2 \approx 0.50$ in-sample and ≈ 0.10 out-of-sample). This is consistent with the finding in Section 4 that breakthrough innovations reshape the network's centrality structure over extended horizons.

6.2. The Economic Value of Science

We implement the JD economic value formula (23) to quantify the value of field i at any origin year t, replacing the eigenvalues inside the expectation with point forecasts $\hat{\sigma}_k(t + j)$

[16] We further examine robustness to alternative models and regularization choices. Appendix Figure A.7 reports in-sample and out-of-sample R^2 for a range of models, including Ridge, Elastic Net, support vector regression with a sigmoid kernel, kernel ridge with a sigmoid kernel, and Random Forest. We show that a Ridge model provides the best predictive performance in terms of both in-sample and out-of-sample R^2. Appendix Figure A.8 shows how the R^2 varies with the ridge penalty, which we vary from 1 to 10^5 in factors of 10. The choice of 10^3 provides a favorable balance between fit and regularization.

formed with information available at t:

$$\hat{\gamma}_{i,t} \;\propto_t\; \sum_{k=1}^{150} (\boldsymbol{\beta}'\boldsymbol{u}_k) \left[1 + \frac{\hat{\sigma}_k(t+1)}{1+\rho/\lambda} + \frac{\hat{\sigma}_k(t+1)\,\hat{\sigma}_k(t+2)}{(1+\rho/\lambda)^2} + \cdots \right] m_k[i],$$

with an implied per-round spillover discount $1/(1+\rho/\lambda)$ of 0.95 (a ratio $\rho/\lambda \approx 0.053$ of the discount rate to the rate at which knowledge spillovers materialize) and the value power series truncated at 20 spillover rounds.

The estimation proceeds in three steps. First, after the joint diagonalization, we forecast the eigenvalue paths $\{\sigma_k(t{+}1), \sigma_k(t{+}2), \ldots\}$ over the next 50 years with the Ridge prediction model from the previous subsection. Our dynamic estimate uses these out-of-sample forecasts, which draw only on information available up to year t. A static benchmark freezes the network at its base-year eigenvalues, $\hat{\sigma}_k(t{+}j) = \sigma_k(t)$, so that $\hat{\boldsymbol{\gamma}}$ reduces to the current-network Katz centrality. Second, we substitute the forecasted eigenvalues into the discrete-time value formula to compute $\hat{\gamma}_{i,t}$ for each field. Because the forecasts extend 50 years and the value power series runs for 20 rounds, we can compute each field's economic value for origin years spanning the next 30 years. For example, standing in 2010, we can calculate field i's value in 2030, $\hat{\gamma}_{i,2030}$. Third, we construct 90% prediction intervals with a bootstrap that draws each eigenvalue path by resampling from the prediction residuals, pushes it through the formula, and reports the 5th and 95th percentiles of the resulting distribution of values.

The Economic Value of Science vs. Technology. Table 4 reports the science–technology split of economic value alongside R&D shares, where the R&D share is computed using the U.S. aggregate R&D series from the NSF National Patterns report.

Science produced 56.4% of economic value in 2019 while receiving 25.1% of R&D, a value-to-resource ratio of 2.25; technology created 43.6% of value on 74.9% of R&D, a ratio of 0.58. These numbers suggest that science is underfunded relative to its value, by roughly a factor of four compared with technology. Over 1980–2019, the dynamic science value share holds a majority in every snapshot, and the science value-to-resource ratio stays between 1.70 and 2.30 throughout. The static Katz-centrality benchmark puts science's value share even higher, implying a value-to-resource ratio between 2.20 and 2.59. The difference is mechanical: the dynamic estimate uses the consumption weights $\boldsymbol{\beta}$, which place zero direct weight on science fields, whereas the static benchmark weights science and technology equally and so gives science direct weight, raising its measured share.

Table 4. The Economic Value of Science vs Technology

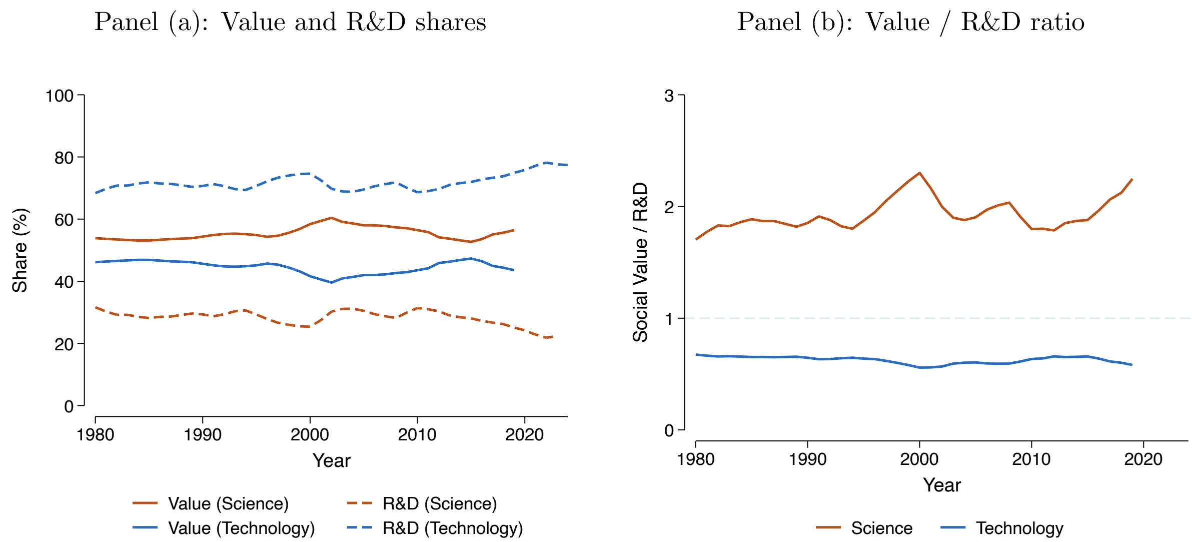
Year	R&D (%)	Economic Value	Economic Value (%)	Economic Value / R&D	Katz Centrality (%)	Katz / R&D
Panel (a): Science						
1980	31.63	9.36	53.87	1.70	72.23	2.28
1990	29.34	9.41	54.37	1.85	71.19	2.43
2000	25.37	11.30	58.38	2.30	65.79	2.59
2010	31.39	10.68	56.44	1.80	69.02	2.20
2019	25.12	10.17	56.44	2.25	60.56	2.41
Panel (b): Technology						
1980	68.37	8.01	46.13	0.67	27.77	0.41
1990	70.66	7.90	45.63	0.65	28.81	0.41
2000	74.63	8.06	41.62	0.56	34.21	0.46
2010	68.61	8.24	43.56	0.63	30.98	0.45
2019	74.88	7.85	43.56	0.58	39.44	0.53

Notes. Science and technology shares in selected years. R&D (%) is each domain's share of total U.S. R&D, calibrated to the NSF National Patterns benchmark. Economic Value is the level of forward-looking, consumption-weighted economic value (from the out-of-sample breakthrough forecasts); Economic Value (%) is the domain's share of economic value; Katz Centrality is its share under the static-network benchmark. The ratio columns divide each value share by the R&D share; a ratio above one means value exceeds resources.

The Value of Science over Time. Figure 11 plots science's and technology's economic value over time in two ways: as shares of total value set against their R&D shares (Panel a), and as the resulting value-to-resource ratio (Panel b). The science value share (Panel a) exceeds its R&D share throughout the sample, and the science value-to-resource ratio (Panel b) stays well above one in every year with R&D data. Two patterns stand out over time. Science's value share is roughly stable near 55%, while its R&D share drifts down from about 31% in 1980 to 22% by the early 2020s and technology's R&D share rises correspondingly, so the value-to-resource gap widens over the sample. The science ratio is broadly rising as a result—from about 1.70 in 1980 to a local peak near 2.30 around 2000, easing to about 1.80 in the early 2010s before climbing back to 2.25 by 2019—whereas technology's ratio stays flat and below one throughout (roughly 0.55–0.67).

Where the Value Sits Across Disciplines. Table 5 breaks the 2019 cross-section down by aggregate discipline, ordered within each panel by the value-to-resource ratio. Within

Figure 11. The Value of Science Over Time



Notes. Panel (a) plots, for each year, the share of total economic value (solid lines) and the share of total R&D (dashed lines) accounted for by science and technology; R&D shares are NSF-calibrated so that the aggregate science share matches the NSF National Patterns U.S. benchmark. Panel (b) plots the ratio of the economic-value share to the R&D share for each domain.

science, most disciplines carry a ratio well above one—mathematics (12.27), engineering (4.63), computer science (4.19), chemistry (3.10), and physics (3.07)—while only medicine (1.00), biology (0.93), and materials science (0.90) sit near or just below parity. Within technology, the largest-R&D disciplines are the most over-resourced: the technology *physics* class (instruments, computing, and measurement) at a ratio of 0.69 on a 20.6% R&D share, electricity at 0.47 on 16.7%, and health at 0.20 on 9.3%.

Field-Level Dynamics. The aggregate shares mask substantial heterogeneity in how individual fields' value evolves. Figure 12 plots value trajectories for 24 selected fields, organized into four panels by major discipline: computer science, medicine, biology, and physics. In each sub-figure, the black line plots the field's Katz centrality over time. As we established in Section 5.2, Katz centrality corresponds to the value obtained if the innovation network at a given date t were held constant forever—it captures the "static" component of the value, abstracting from future changes in the network structure. The colored lines, originating at different dates, show our out-of-sample forecasts of the economic value made from successive origin years. At its starting year t, the gap between a forecast curve and the Katz centrality

Table 5. The Economic Value of Science by Discipline in 2019

Discipline	R&D (%)	Economic Value	Economic Value (%)	Economic Value / R&D	Katz Centrality (%)	Katz / R&D
Panel A: Science						
mathematics	0.25	0.54	3.01	12.27	2.38	9.73
engineering	1.91	1.59	8.83	4.63	10.06	5.27
computer science	3.21	2.43	13.47	4.19	11.35	3.53
environmental science	0.22	0.16	0.91	4.10	1.89	8.51
geography	0.45	0.31	1.72	3.87	2.39	5.37
geology	0.76	0.49	2.74	3.61	3.42	4.51
chemistry	1.95	1.09	6.03	3.10	5.84	3.00
physics	1.89	1.04	5.80	3.07	3.73	1.98
medicine	7.75	1.39	7.72	1.00	11.62	1.50
biology	4.54	0.76	4.24	0.93	6.32	1.39
materials science	2.20	0.36	1.98	0.90	1.56	0.71
Panel B: Technology						
separating & mixing	1.36	0.50	2.77	2.04	3.19	2.35
textiles	0.40	0.13	0.70	1.77	1.98	4.99
foods & tobacco	0.51	0.12	0.68	1.34	0.95	1.86
metallurgy	0.58	0.13	0.72	1.24	1.10	1.92
agriculture	0.87	0.13	0.70	0.81	0.40	0.46
shaping	1.84	0.27	1.48	0.81	5.31	2.89
physics	20.64	2.58	14.30	0.69	5.14	0.25
chemistry	7.94	0.98	5.46	0.69	4.96	0.62
building	1.59	0.14	0.78	0.49	1.33	0.84
electricity	16.70	1.42	7.86	0.47	1.82	0.11
personal or household	0.76	0.06	0.35	0.45	1.50	1.97
transporting	5.85	0.44	2.47	0.42	2.91	0.50
engines & pumps	2.43	0.14	0.79	0.32	1.02	0.42
general engineering	1.63	0.06	0.35	0.22	0.91	0.56
health	9.32	0.34	1.91	0.20	0.76	0.08
lighting & heating	0.94	0.03	0.19	0.20	2.27	2.41

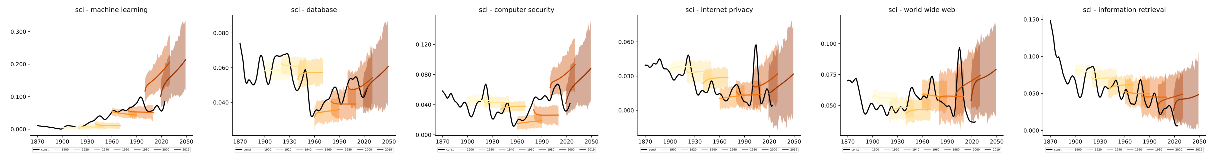
Notes. 2019 cross-section by aggregate discipline, ordered within each panel by Economic Value / R&D. R&D (%), Economic Value (%), and Katz Centrality (%) are each discipline's share of the economy-wide (science + technology) total. The ratio columns divide each value share by the R&D share; a ratio above one means value exceeds resources. We only present the disciplines with greater than 1% of documents within each sector.

line reflects the extent to which anticipated future changes in the network raise or lower the field's value relative to what it would be under a static network; the subsequent trajectory of each colored curve then traces how the forecasted economic value evolves as the predicted network changes unfold.

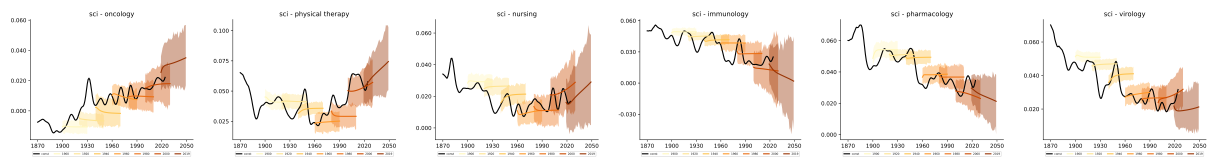
Several patterns emerge from the figure. The computer-science panel reveals sharply divergent dynamics. Machine learning exhibits near-exponential projected growth in economic value: its forecast curves climb far above its static Katz centrality, and the vintage originating in 2019 more than doubles over the following three decades, reflecting its role as a general-purpose scientific tool with broad downstream applications. Database provides the contrast. Its static value is essentially unchanged since 1990, and its forecast curves lie almost exactly on top of the Katz line: a mature infrastructure field whose position in the network the model expects to persist. The fields that built the commercial internet instead

Figure 12. The Economic Value of Science: Field-Level Value Trajectories

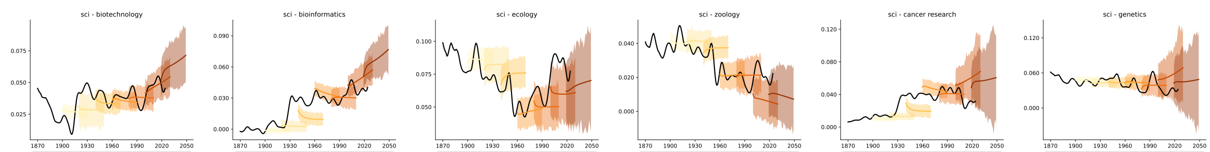
Panel (a): Computer Science



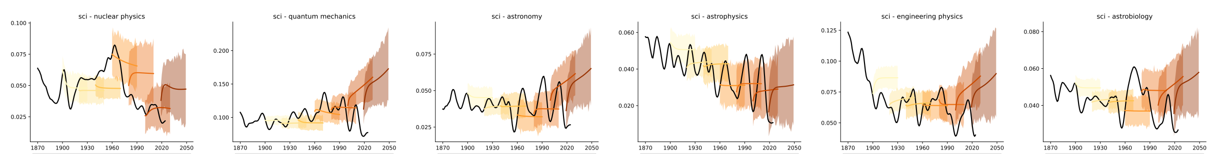
Panel (b): Medicine



Panel (c): Biology



Panel (d): Physics



Notes. Estimated economic value for 24 selected scientific fields, six per major discipline (computer science, medicine, biology, physics). In each sub-figure the black line is the field's Katz centrality and the colored lines are out-of-sample forecasts of the forward-looking economic value made from successive origin years, each using only the data available up to that year. The gap between a forecast curve and the Katz line at its origin reflects anticipated future changes in the network; the subsequent path traces how the forecasted value evolves.

trace hump shapes. The World Wide Web and internet privacy both peak in the early 2000s and then fall back sharply, as their foundational contribution to knowledge diffusion is absorbed by the fields that build on them, and information retrieval declines most steeply of all, displaced by the learning-based methods that now perform the same task. Computer security rose through the 2000s before receding.

A common reallocation from foundational to applied work appears in medicine and biology. In medicine, the clinical and applied fields gain value: oncology and physical therapy both trend upward and are forecast to continue rising. The basic-science areas recede. Immunology, pharmacology, and virology are flat or in decline, and their forecast curves point to no recovery. Biology shows the same tilt. Biotechnology and bioinformatics, the applied and computational areas of the discipline, both rise, and their forecasts are among the most accurate in the figure. Classical descriptive biology, represented here by zoology, fades outright. Genetics illustrates the pattern most starkly: its static value peaked in the early 1990s and has since fallen by more than half, as the foundational contribution of the genomic revolution is absorbed by the applied fields downstream.

Physics is the discipline in which this displacement is most pronounced, and the same tilt toward applied work appears within it. All six subfields lose static value relative to their mid-century peaks, but they do not lose it equally: engineering physics, the most applied of the six, gives up the least, while the fundamental and observational areas of astronomy and astrophysics fall the most steeply. Quantum mechanics remains the most valuable physics subfield throughout. The pattern is consistent with the centrality dynamics documented in Section 4: as newer computer-science and biology fields rise in the rankings, established physics subfields are gradually displaced.

A notable feature across all panels is that the out-of-sample forecast curves generally track the ex post realized Katz centrality closely at short horizons, but diverge more at longer horizons. This is consistent with the predictability results in the previous subsection: the eigenvalue forecasts are most accurate at short horizons and become less precise over time. That the colored curves from different starting dates are broadly consistent with one another—and with the subsequent evolution of the black Katz line—provides reassurance that the forecasting model captures the salient trends in the innovation network.

6.3. Why Is Science So Valuable? A Decomposition

Science carries no direct consumption value ($\beta_{\text{science}} = 0$), so all of its value must be realized indirectly, through the downstream technologies its knowledge feeds. We quantify this mechanism by decomposing a field's forward-looking economic value $\hat{\gamma}$ into the spillover rounds through which it flows. The economic value formula can be rewritten as:

$$\hat{\gamma}_t' \propto_t \sum_{k=1}^{K} (\boldsymbol{\beta}' \boldsymbol{u}_k) \Big[\underbrace{1}_{\text{Round 0}} + \underbrace{\frac{\hat{\sigma}_k(t+1)}{1+\rho/\lambda}}_{\text{Round 1}} + \underbrace{\frac{\hat{\sigma}_k(t+1)\,\hat{\sigma}_k(t+2)}{(1+\rho/\lambda)^2}}_{\text{Round 2}} + \cdots \Big] \boldsymbol{m}_k'.$$

Each spillover round s contributes the degree-s term, $\sum_k(\boldsymbol{\beta}'\boldsymbol{u}_k)\left[\prod_{j=1}^{s}\hat{\sigma}_k(t+j)\right]/(1+\rho/\lambda)^s\,\boldsymbol{m}_k'$: Round 0 is the direct consumption term, Round 1 the one-step spillover, Round 2 the two-step, and Round 3+ collects all higher-order terms.

Table 6. A Decomposition of Economic Value by Spillover Rounds

Year	Round 0 (Direct, %)	Round 1 (1 step, %)	Round 2 (2 steps, %)	Round 3+ (multi, %)
Panel A: Science fields				
1900	0.00	40.06	25.75	34.19
1940	0.00	46.05	26.30	27.64
1980	0.00	44.17	28.18	27.65
2000	0.00	45.70	28.45	25.84
2019	0.00	50.09	28.43	21.49
Panel B: Technology fields				
1900	64.53	17.55	8.07	9.86
1940	68.76	16.49	7.06	7.69
1980	70.82	18.45	5.94	4.80
2000	71.64	19.13	5.71	3.53
2019	73.44	18.56	5.18	2.83

Notes. Each entry is the share (%) of a field type's total forward-looking economic value $\hat{\gamma}$ realized in a given spillover round, averaged across fields of that type in the selected years. Round 0 is the direct consumption term (β), Round 1 the one-step spillover, Round 2 the two-step, and Round 3+ all higher-order terms. Science fields carry $\beta = 0$, so their Round 0 share is exactly zero. Rows sum to 100% up to rounding.

Table 6 contrasts science and technology. Technology's value is overwhelmingly direct: 73.4% is realized in Round 0 in 2019 and 18.6% in Round 1. By contrast, science's value is entirely indirect, with 0% in Round 0 by construction, 50.1% arriving one step downstream, 28.4% two steps, and 21.5% three or more steps. This is the key mechanism behind the economic value of science: a direct-consumption accounting of value misses science entirely, because science is valuable precisely through what technology does with it. Figure A.10 plots the direct/indirect split over time.

Case Studies: Where Science's Value Lands. To see the spillover mechanism concretely, we trace where four canonical science-origin innovations deliver their value. For each origin field, we attribute its decomposed value across the recipient fields its knowledge flows through, year by year on the realized network, averaged over a five-year window around the innovation's peak. Specifically, we follow each chain of spillovers that carries the origin's knowledge outward and split the value it generates across every field on the chain, crediting a field for its position along the path.[17] Because value is ultimately realized in technology ($\beta_{\text{science}} = 0$), a science field earns a share not by absorbing value directly but by relaying it—it appears as an intermediary on the path from the origin to the technology that embodies the knowledge. Summing over both science and technology recipients recovers the source's total value γ_j exactly.

Table 7 lists the ten largest recipients for machine learning (2020–2024), the World Wide Web (1998–2002), molecular biology (1985–1989), and nuclear physics (1958–1962). We choose these four fields because each anchored a defining technology cluster of its era—postwar atomic energy, the biotechnology revolution, the commercial internet, and modern artificial intelligence—so that together they trace six decades of science-driven innovation.

Several regularities stand out. First, a general-purpose computing and calculating technology is the largest recipient in every case, absorbing 32.0% of machine learning's value, 25.7% of the World Wide Web's, 19.5% of nuclear physics's, and 5.7% of molecular biology's. Second, the originating academic field retains only a small slice of the value it creates—5.7% for machine learning, 9.2% for the World Wide Web, 9.8% for molecular biology, and just 2.9% for nuclear physics. Third, and relatedly, value does spread to scientifically adjacent fields—the World Wide Web feeds internet privacy and computer security, and molecular biology feeds genetics and cell biology—but the bulk lands downstream on patented technology. Aggregating over all recipients, industrial technology absorbs 77% of machine learning's value, 73% of the World Wide Web's, 72% of molecular biology's, and 81% of nuclear physics's. The result further supports the idea that science is valuable not for what it delivers to consumers directly but for what technology does with it.

[17] Consider the value decomposition of a given origin field $x_0 = j$. A path of length k is $\pi = (x_0, \ldots, x_k)$, with the value of this path being $V(\pi) = \delta^k \, \beta_{x_k} \prod_{r=1}^{k} \Omega_{x_r x_{r-1}}$. That is, path value equals discounting times terminal-field value times the product of spillover strengths along the path. Therefore, the total value of origin j is $\gamma_j = \sum_{\pi: x_0 = j} V(\pi)$. Because only technology carries weight ($\beta_{\text{science}} = 0$), only paths terminating in technology contribute. For each path π, we then allocate its value across its non-origin nodes $x_1, \ldots, x_k$ with decaying weights $d_r = \delta^{\,r-1} / \sum_{m=1}^{k} \delta^{\,m-1}$ that sum to one. Putting these together, the contribution of a recipient field q is the total value assigned to all occurrences of q across all paths originating at j:

$$C_q = \sum_{\pi: x_0 = j} V(\pi) \sum_{r=1}^{k} \mathbf{1}\{x_r = q\}\, d_r = \sum_{k \geq 1} \delta^k \sum_{r=1}^{k} d_r \, [\boldsymbol{\Omega}^r]_{qj} \, (\boldsymbol{\beta}^\top \boldsymbol{\Omega}^{k-r})_q,$$

where the second equality collects paths by length k, with $[\boldsymbol{\Omega}^r]_{qj}$ the r-step flow from j to q and $(\boldsymbol{\beta}^\top \boldsymbol{\Omega}^{k-r})_q$ the value collected onward from q. Since $\sum_q C_q = \gamma_j$, the contribution share of field q is $s_q = C_q / \gamma_j$.

Table 7. Where Does Science's Value Land? Case Studies

Panel (a): Nuclear physics (1958–1962)

Rank	Recipient field	Share (%)
1	tech - computing & calculating or counting	19.49
2	tech - earth or rock drilling & mining	4.47
3	sci - nuclear physics	2.93
4	tech - measuring & testing	2.58
5	tech - basic electric elements	2.58
6	tech - basic electronic circuitry	2.57
7	tech - physical or chemical apparatus	2.56
8	tech - nuclear physics	2.25
9	tech - electric power	2.24
10	tech - petroleum, gas or coke industries	2.16

Panel (b): Molecular biology (1985–1989)

Rank	Recipient field	Share (%)
1	sci - molecular biology	9.80
2	tech - computing & calculating or counting	5.67
3	tech - basic electric elements	5.08
4	sci - genetics	4.72
5	tech - earth or rock drilling & mining	4.47
6	tech - electric communication technique	3.21
7	sci - cell biology	3.21
8	tech - medical or veterinary science	3.10
9	tech - electric power	2.80
10	tech - vehicles in general	2.70

Panel (c): World Wide Web (1998–2002)

Rank	Recipient field	Share (%)
1	tech - computing & calculating or counting	25.70
2	tech - electric communication technique	11.73
3	sci - internet privacy	9.48
4	sci - world wide web	9.18
5	tech - physical or chemical apparatus	3.61
6	sci - computer security	2.53
7	tech - sports & games	2.48
8	tech - vehicles in general	2.44
9	tech - earth or rock drilling & mining	2.38
10	tech - electric power	2.31

Panel (d): Machine learning (2020–2024)

Rank	Recipient field	Share (%)
1	tech - computing & calculating or counting	31.97
2	tech - earth or rock drilling & mining	10.97
3	sci - machine learning	5.74
4	tech - organic chemistry	3.47
5	tech - controlling & regulating	3.44
6	tech - electric communication technique	3.23
7	tech - physical or chemical apparatus	3.08
8	sci - data mining	2.34
9	tech - sports & games	2.11
10	tech - machine tools	2.01

Notes. Each panel lists the ten fields absorbing the largest shares of a science-origin field's forward-looking value, over the window in the panel header. Share is the percent of the origin's total value landing in that recipient; the `sci -` or `tech -` prefix marks a science or technology recipient, and a field such as nuclear physics can appear as both. Across all recipients, technology absorbs 81%, 72%, 73%, and 77% of the value in panels (a)–(d), respectively.

6.4. Does the Value of Science Predict Real Outcomes?

If our measure of the value of science captures something real, a field's value should forecast its future trajectory. We regress three future outcomes—a field's R&D share, its knowledge output (ln of new documents), and its count of top 5% breakthrough innovations (ln of breakthrough documents)—on the field's current share of total economic value, at horizons of one to twenty years, including field and year fixed effects and clustering standard errors on the 332 fields. Table 8 reports the regressions at the one- and ten-year horizons, and Appendix Figure A.9 plots the estimated response at every horizon and separately for science and technology fields.

Our measure of economic value predicts all three outcomes positively and significantly. A one-point-higher economic-value share predicts a field's future R&D share about 0.61 points higher one year ahead and 0.80 points higher ten years ahead, and it forecasts substantially

Table 8. Does Science Value Predict Real Outcomes?

Panel (a): R&D Share

	(1)	(2)	(3)	(4)
	R&D Share$_{i,t+1}$		R&D Share$_{i,t+10}$	
Economic Value (%)	0.607**	0.597**	0.798**	0.827**
	(0.238)	(0.248)	(0.376)	(0.404)
ln(Knowledge Stock)		0.008		-0.016
		(0.009)		(0.020)
Observations	13,612	13,612	14,731	14,731
R-squared	0.977	0.977	0.969	0.969

Panel (b): # Documents

	(1)	(2)	(3)	(4)
	ln(# Documents)$_{i,t+1}$		ln(# Documents)$_{i,t+10}$	
Economic Value (%)	3.783***	0.298***	3.916***	0.592***
	(0.654)	(0.060)	(0.601)	(0.119)
ln(Knowledge Stock)		0.955***		0.895***
		(0.009)		(0.015)
Observations	40,172	40,172	38,512	38,512
R-squared	0.890	0.989	0.884	0.966

Panel (c): # Breakthrough Documents

	(1)	(2)	(3)	(4)
	ln(# Breakthroughs)$_{i,t+1}$		ln(# Breakthroughs)$_{i,t+10}$	
Economic Value (%)	4.221***	2.359***	3.243***	1.687***
	(0.620)	(0.388)	(0.423)	(0.324)
ln(Knowledge Stock)		0.509***		0.416***
		(0.032)		(0.038)
Observations	39,840	39,840	36,852	36,852
R-squared	0.555	0.600	0.519	0.544
Field	Yes	Yes	Yes	Yes
Year	Yes	Yes	Yes	Yes
No. of Fields	332	332	332	332

Notes. Field-year panel regressions of future outcomes on the field's current economic-value share (its share of total economic value, $\times 100$). The dependent variable is the field's NSF-calibrated R&D share (%) in Panel (a), $\ln(1 + \text{new documents})$ in Panel (b), and $\ln(1 + \text{breakthrough documents})$ (top 5% novelty) in Panel (c), each measured 1 and 10 years ahead. Odd columns include no controls; even columns add $\ln(1 + \text{knowledge stock})$. All specifications include field and year fixed effects. Standard errors are clustered by field in parentheses. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

higher future knowledge output and breakthrough counts, with the breakthrough response the most robust of the three. Adding the field's log knowledge stock as a control leaves the R&D response essentially unchanged; it attenuates the knowledge-output and breakthrough responses—the latter roughly halving—but both remain significant at the 1% level, so economic value carries predictive content beyond a field's accumulated scale. The R&D response, however, is concentrated among technology fields; for science fields it is close to zero, indicating that resources reallocate toward high-value fields only slowly and unevenly.

7. Conclusion

We develop a framework for measuring the economic value of science. Scientific fields produce no consumption goods, so their value flows entirely through the technologies they enable. We trace these flows in a text-based innovation network built from 463 million scientific articles and 81 million patents, covering 209 scientific fields and 123 technology classes in every year from 1870 to 2024. An endogenous growth model turns the network into a valuation. The economic value of R&D in a field equals its discounted influence on downstream consumption. Our framework allows the network to change over time, including through random regime changes, and the value becomes a forward-looking measure that tracks the shifting structure of knowledge flows.

The estimates show that science accounts for the majority of the economic value of R&D. In 2019, scientific fields generated 56% of the forward-looking economic value of R&D in the United States, while receiving about a quarter of R&D resources. All of this value is indirect: half arrives one spillover step downstream, the rest through longer chains. We also trace the value of four canonical breakthroughs: nuclear physics, molecular biology, the internet, and machine learning. In each case, between 72% and 81% of the value ends up in patented technologies.

The value of science also shifts over time. Machine learning has risen sharply, as have applied fields such as biotechnology and oncology. Classical physics subfields have declined as their foundational contributions were absorbed downstream. R&D resources follow these shifts only gradually. A field's value predicts future reallocation among technology fields, while science funding shares adjust more slowly. An up-to-date map of scientific value is therefore useful in itself. Periodic reassessment can help both public and private research investments capture the largest returns.

R&D is among the largest investments a modern economy makes. In the United States alone, it amounts to hundreds of billions of dollars each year, and basic science underpins the technologies that sustain long-run growth. The value of science is nonetheless hard to see, since it cannot be priced by markets. Our framework makes this value visible at the level of individual fields, grounded in welfare. Science's share of economic value is roughly twice its share of R&D resources, and this holds in every snapshot from 1980 through 2019. This finding gives quantitative content to the case for basic research made since Bush (1945) and formalized in the spillover literature (Griliches, 1958; Jaffe, 1986; Jones and Williams, 1998). Policymakers, research funders, universities, and the public can use this accounting to understand where the value of science comes from and how it evolves.

Several caveats apply, and they share a common thread: our estimates are more reliable about rankings and directions than about precise magnitudes. The value formula is marginal. It ranks fields and diagnoses misallocation at the margin, but it cannot evaluate large reallocations, which would reshape the network itself. The network is measured from text. It therefore captures the diffusion of terminology rather than knowledge itself, and its coverage of early periods is thinner. The level of measured value also depends on the ratio of the discount rate to the R&D arrival rate, and this dependence is stronger for science, whose value arrives through long chains. Finally, the model omits creative destruction and directed technical change, so it is silent on how markets would respond to the gap we document. Each caveat accordingly weighs on the magnitudes we report more than on the central pattern: science contributes far more to economic value than the resources it receives.

Two extensions are natural. The first is to further endogenize the innovation network, so that funding choices reshape knowledge flows and large policy changes can be evaluated. The second is to connect the field-level estimates to the scientists, teams, universities, and funders that produce knowledge. Our framework offers a welfare-grounded accounting of science's contribution to growth. It is not a substitute for expert judgment in science policy. We view our result as a quantitative input that has been largely absent from these decisions.

References

Acemoglu, Daron, Ufuk Akcigit, and William R Kerr. 2016. "Innovation network." *Proceedings of the National Academy of Sciences*, 113(41): 11483–11488.

Acemoglu, Daron, Ufuk Akcigit, Harun Alp, Nicholas Bloom, and William Kerr. 2018. "Innovation, reallocation, and growth." *American Economic Review*, 108(11): 3450–3491.

Agrawal, Ajay, and Rebecca Henderson. 2002. "Putting patents in context: Exploring knowledge transfer from MIT." *Management science*, 48(1): 44–60.

Ahmadpoor, Mohammad, and Benjamin F Jones. 2017. "The dual frontier: Patented inventions and prior scientific advance." *Science*, 357(6351): 583–587.

Ahmadpoor, Mohammad, and Benjamin F Jones. 2019. "Decoding team and individual impact in science and invention." *Proceedings of the National Academy of Sciences*, 116(28): 13885–13890.

Akcigit, Ufuk, and William R Kerr. 2018. "Growth through heterogeneous innovations." *Journal of Political Economy*, 126(4): 1374–1443.

Akcigit, Ufuk, Douglas Hanley, and Nicolas Serrano-Velarde. 2021. "Back to basics: Basic research spillovers, innovation policy, and growth." *The Review of Economic Studies*, 88(1): 1–43.

Andrews, Michael J. 2023. "How do institutions of higher education affect local invention? Evidence from the establishment of US colleges." *American Economic Journal: Economic Policy*, 15(2): 1–41.

Arora, Ashish, Sharon Belenzon, and Lia Sheer. 2021. "Knowledge spillovers and corporate investment in scientific research." *American Economic Review*, 111(3): 871–898.

Arora, Ashish, Sharon Belenzon, Konstantin Kosenko, Jungkyu Suh, and Yishay Yafeh. 2024. "The Rise of Scientific Research in Corporate America." *Organization Science*, 36(4): 1466–1488.

Arqué-Castells, Pere, and Daniel F Spulber. 2022. "Measuring the private and social returns to R&D: Unintended spillovers versus technology markets." *Journal of Political Economy*, 130(7): 1860–1918.

Arrow, Kenneth. 1962. "Economic welfare and the allocation of resources for invention." In *The rate and direction of inventive activity: Economic and social factors*. 609–626. Princeton University Press: Princeton, NJ.

Atkeson, Andrew, and Ariel Burstein. 2019. "Aggregate implications of innovation policy." *Journal of Political Economy*, 127(6): 2625–2683.

Azoulay, Pierre, and Danielle Li. 2022. "Scientific Grant Funding." *Innovation and Public Policy*, 117.

Azoulay, Pierre, Joshua Graff-Zivin, Brian Uzzi, Dashun Wang, Heidi Williams, James A Evans, Ginger Zhe Jin, Susan Feng Lu, Benjamin F Jones, Katy Börner, et al. 2018. "Toward a more scientific science." *Science*, 361(6408): 1194–1197.

Azoulay, Pierre, Joshua S Graff Zivin, and Gustavo Manso. 2011. "Incentives and creativity: evidence from the academic life sciences." *The RAND Journal of Economics*, 42(3): 527–554.

Azoulay, Pierre, Joshua S Graff Zivin, Danielle Li, and Bhaven N Sampat. 2019. "Public R&D investments and private-sector patenting: evidence from NIH funding rules." *The Review of economic studies*, 86(1): 117–152.

Belenzon, Sharon, and Mark Schankerman. 2013. "Spreading the word: Geography, policy, and knowledge spillovers." *Review of Economics and Statistics*, 95(3): 884–903.

Biasi, Barbara, and Petra Moser. 2021. "Effects of copyrights on science: Evidence from the wwii book republication program." *American Economic Journal: Microeconomics*, 13(4): 218–260.

Bikard, Michaël, and Matt Marx. 2020. "Bridging academia and industry: How geographic hubs connect university science and corporate technology." *Management Science*, 66(8): 3425–3443.

Bloom, Nicholas, Charles I Jones, John Van Reenen, and Michael Webb. 2020. "Are ideas getting harder to find?" *American Economic Review*, 110(4): 1104–44.

Bloom, Nicholas, Mark Schankerman, and John Van Reenen. 2013. "Identifying technology spillovers and product market rivalry." *Econometrica*, 81(4): 1347–1393.

Branstetter, Lee. 2005. "Exploring the link between academic science and industrial innovation." *Annales d'Economie et de Statistique*, 119–142.

Buera, Francisco J, and Ezra Oberfield. 2020. "The global diffusion of ideas." *Econometrica*, 88(1): 83–114.

Bush, Vannevar. 1945. *Science, the Endless Frontier.* Washington, D.C.:United States Government Printing Office.

Caballero, Ricardo J, and Adam B Jaffe. 1993. "How high are the giants' shoulders: An empirical assessment of knowledge spillovers and creative destruction in a model of economic growth." *NBER macroeconomics annual*, 8: 15–74.

Cai, Jie, and Can Tian. 2021. "Innovation and Endogenous Knowledge Network Dynamics." Working Paper.

Cai, Jie, and Nan Li. 2019. "Growth through inter-sectoral knowledge linkages." *The Review of Economic Studies*, 86(5): 1827–1866.

Cai, Jie, Nan Li, and Ana Maria Santacreu. 2022. "Knowledge diffusion, trade, and innovation across countries and sectors." *American Economic Journal: Macroeconomics*, 14(1): 104–145.

Cai, Sheng, Lorenzo Caliendo, Fernando Parro, and Wei Xiang. 2022. "Mechanics of spatial growth." National Bureau of Economic Research.

Clauset, Aaron, Samuel Arbesman, and Daniel B Larremore. 2015. "Systematic inequality and hierarchy in faculty hiring networks." *Science advances*, 1(1): e1400005.

Coe, David T, and Elhanan Helpman. 1995. "International r&d spillovers." *European economic review*, 39(5): 859–887.

Coe, David T, Elhanan Helpman, and Alexander W Hoffmaister. 2009. "International R&D spillovers and institutions." *European Economic Review*, 53(7): 723–741.

Coppel, W. A. 1965. *Stability and Asymptotic Behavior of Differential Equations.* Boston:D. C. Heath.

Durbin, James, and Siem Jan Koopman. 2012. *Time series analysis by state space methods.* Oxford university press.

Eaton, Jonathan, and Samuel Kortum. 1999. "International technology diffusion: Theory and measurement." *International Economic Review*, 40(3): 537–570.

Eaton, Jonathan, and Samuel S Kortum. 2006. "Innovation, diffusion, and trade."

Eberhardt, Markus, Christian Helmers, and Hubert Strauss. 2013. "Do spillovers matter when estimating private returns to R&D?" *Review of Economics and Statistics*, 95(2): 436–448.

Fleming, Lee, and Olav Sorenson. 2004. "Science as a map in technological search." *Strategic Management Journal*, 25(8-9): 909–928.

Fleming, Lee, Hillary Greene, G Li, Matt Marx, and Dennis Yao. 2019. "Government-funded research increasingly fuels innovation." *Science*, 364(6446): 1139–1141.

Fortunato, Santo, Carl T Bergstrom, Katy Börner, James A Evans, Dirk Helbing, Staša Milojević, Alexander M Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, et al. 2018. "Science of science." *Science*, 359(6379): eaao0185.

Garcia-Macia, Daniel, Chang-Tai Hsieh, and Peter J Klenow. 2019. "How destructive is innovation?" *Econometrica*, 87(5): 1507–1541.

Griffith, Rachel, Rupert Harrison, and John Van Reenen. 2006. "How special is the special relationship? Using the impact of US R&D spillovers on UK firms as a test of technology sourcing." *American Economic Review*, 96(5): 1859–1875.

Griliches, Zvi. 1958. "Research costs and social returns: Hybrid corn and related innovations." *Journal of political economy*, 66(5): 419–431.

Guillard, Charlotte, Ralf Martin, Catherine Thomas, and Dennis Verhoeven. 2021. “Efficient industrial policy for innovation: standing on the shoulders of hidden giants.”

Hale, Jack K. 1969. *Ordinary Differential Equations.* New York:Wiley-Interscience.

Hall, Bronwyn H, Jacques Mairesse, and Pierre Mohnen. 2010. “Measuring the Returns to R&D.” In *Handbook of the Economics of Innovation.* Vol. 2, 1033–1082. Elsevier.

Hastie, T., R. Tibshirani, and J.H. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer series in statistics*, Springer.

Hausman, Naomi. 2022. “University innovation and local economic growth.” *Review of Economics and Statistics*, 104(4): 718–735.

Hoerl, Arthur E, and Robert W Kennard. 1970. “Ridge regression: Biased estimation for nonorthogonal problems.” *Technometrics*, 12(1): 55–67.

Hofstra, Bas, Vivek V Kulkarni, Sebastian Munoz-Najar Galvez, Bryan He, Dan Jurafsky, and Daniel A McFarland. 2020. “The diversity–innovation paradox in science.” *Proceedings of the National Academy of Sciences*, 117(17): 9284–9291.

Howell, Sabrina T. 2017. “Financing innovation: Evidence from R&D grants.” *American economic review*, 107(4): 1136–1164.

Huang, Jingong, and Yves Zenou. 2020. “Key Sectors in Endogenous Growth.”

Huang, Junming, Alexander J Gates, Roberta Sinatra, and Albert-László Barabási. 2020. “Historical comparison of gender inequality in scientific careers across countries and disciplines.” *Proceedings of the National Academy of Sciences*, 117(9): 4609–4616.

Iaria, Alessandro, Carlo Schwarz, and Fabian Waldinger. 2018. “Frontier knowledge and scientific production: Evidence from the collapse of international science.” *The Quarterly Journal of Economics*, 133(2): 927–991.

Jaffe, Adam B. 1986. “Technological Opportunity and Spillovers of R&D: Evidence from Firms’ Patents, Profits, and Market Value.” *American Economic Review*, 984–1001.

Jaffe, Adam B. 1989. “Real effects of academic research.” *The American economic review*, 957–970.

Jaffe, Adam B., Manuel Trajtenberg, and Rebecca Henderson. 1993. “Geographic localization of knowledge spillovers as evidenced by patent citations.” *Quarterly Journal of Economics*, 108(3): 577–598.

Jones, Benjamin F, Lawrence H Summers, et al. 2020. “A calculation of the social returns to innovation.” 27863.

Jones, Charles I, and John C Williams. 1998. “Measuring the social return to R&D.” *The Quarterly Journal of Economics*, 113(4): 1119–1135.

Kelly, Bryan, Dimitris Papanikolaou, Amit Seru, and Matt Taddy. 2021. "Measuring technological innovation over the long run." *American Economic Review: Insights*, 3(3): 303–320.

Kingma, Diederik P, and Jimmy Ba. 2014. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980*.

Kogan, Leonid, Dimitris Papanikolaou, Amit Seru, and Noah Stoffman. 2017. "Technological innovation, resource allocation, and growth." *Quarterly Journal of Economics*, 132(2): 665–712.

König, Michael, Kjetil Storesletten, Zheng Song, and Fabrizio Zilibotti. 2022. "From imitation to innovation: Where is all that Chinese R&D going?" *Econometrica*, 90(4): 1615–1654.

Krieger, Joshua Lev, Monika Schnitzer, and Martin Watzinger. 2024. "Standing on the Shoulders of Science." *Strategic Management Journal*, 45(9): 1670–1695.

Lin, Zihang, Yian Yin, Lu Liu, and Dashun Wang. 2023. "SciSciNet: A large-scale open data lake for the science of science research." *Scientific Data*, 10(1): 315.

Liu, Ernest, and Song Ma. 2021. "Innovation networks and R&D allocation." National Bureau of Economic Research.

Liu, Lu, Benjamin F Jones, Brian Uzzi, and Dashun Wang. 2023. "Data, measurement and empirical methods in the science of science." *Nature human behaviour*, 1–13.

Liu, Lu, Yang Wang, Roberta Sinatra, C Lee Giles, Chaoming Song, and Dashun Wang. 2018. "Hot streaks in artistic, cultural, and scientific careers." *Nature*, 559(7714): 396–399.

Lucking, Brian, Nicholas Bloom, and John Van Reenen. 2019. "Have R&D Spillovers Declined in the 21st Century?" *Fiscal Studies*, 40(4): 561–590.

Lybbert, Travis J, and Nikolas J Zolas. 2014. "Getting patents and economic data to speak to each other: An 'algorithmic links with probabilities' approach for joint analyses of patenting and economic activity." *Research Policy*, 43(3): 530–542.

Ma, Athen, Raúl J Mondragón, and Vito Latora. 2015. "Anatomy of funded research in science." *Proceedings of the National Academy of Sciences*, 112(48): 14760–14765.

Mansfield, Edwin. 1980. "Basic research and productivity increase in manufacturing." *The American economic review*, 70(5): 863–873.

Mansfield, Edwin, John Rapoport, Anthony Romeo, Samuel Wagner, and George Beardsley. 1977. "Social and private rates of return from industrial innovations." *The quarterly Journal of economics*, 91(2): 221–240.

Marx, Matt, and Aaron Fuegi. 2020. "Reliance on science: Worldwide front-page patent citations to scientific articles." *Strategic Management Journal*, 41(9): 1572–1594.

Marx, Matt, and Aaron Fuegi. 2022. "Reliance on science by inventors: Hybrid extraction of in-text patent-to-article citations." *Journal of Economics & Management Strategy*, 31(2): 369–392.

Ma, Song. 2026. "Technological Obsolescence." *Review of Financial Studies.*

Ma, Yifang, and Brian Uzzi. 2018. "Scientific prize network predicts who pushes the boundaries of science." *Proceedings of the National Academy of Sciences*, 115(50): 12608–12615.

McCabe, Mark J, and Christopher M Snyder. 2005. "Open access and academic journal quality." *American Economic Review*, 95(2): 453–458.

Moretti, Enrico, Claudia Steinwender, and John Van Reenen. 2019. "The Intellectual Spoils of War? Defense R&D, Productivity, and International Spillovers." *The Review of Economics and Statistics*, 1–46.

Murray, Fiona, Philippe Aghion, Mathias Dewatripont, Julian Kolev, and Scott Stern. 2016. "Of mice and academics: Examining the effect of openness on innovation." *American Economic Journal: Economic Policy*, 8(1): 212–252.

Myers, Kyle. 2020. "The elasticity of science." *American Economic Journal: Applied Economics*, 12(4): 103–134.

Myers, Kyle R, and Lauren Lanahan. 2022. "Estimating spillovers from publicly funded R&D: Evidence from the US Department of Energy." *American Economic Review*, 112(7): 2393–2423.

Nelson, Richard R. 1959. "The Simple Economics of Basic Scientific Research." *Journal of Political Economy*, 67(3): 297–306.

Poege, Felix, Dietmar Harhoff, Fabian Gaessler, and Stefano Baruffaldi. 2019. "Science quality and the value of inventions." *Science advances*, 5(12): eaay7323.

Rosenberg, Nathan. 1998. "Uncertainty and technological change." *Dale Neef, Anthony Siesfeld y Jacquelyn Cefola, The Economic Impact of Knowledge, Boston, Butterworth Heinemann*, 17–34.

Rumelhart, David E, Geoffrey E Hinton, and Ronald J Williams. 1986. "Learning representations by back-propagating errors." *nature*, 323(6088): 533–536.

Santacreu, Ana Maria. 2015. "Innovation, diffusion, and trade: Theory and measurement." *Journal of Monetary Economics*, 75: 1–20.

Sinatra, Roberta, Dashun Wang, Pierre Deville, Chaoming Song, and Albert-László Barabási. 2016. "Quantifying the evolution of individual scientific impact." *Science*, 354(6312): aaf5239.

Tewksbury, JG, MS Crandall, and WE Crane. 1980. "Measuring the societal benefits of innovation." *Science*, 209(4457): 658–662.

Wang, Dashun, and Albert-Lászlo Barabási. 2021. *The science of science.* Cambridge University Press.

Way, Samuel F, Allison C Morgan, Aaron Clauset, and Daniel B Larremore. 2017. "The misleading narrative of the canonical faculty productivity trajectory." *Proceedings of the National Academy of Sciences*, 114(44): E9216–E9223.

Williams, Heidi L. 2013. "Intellectual property rights and innovation: Evidence from the human genome." *Journal of Political Economy*, 121(1): 1–27.

Wuchty, Stefan, Benjamin F Jones, and Brian Uzzi. 2007. "The increasing dominance of teams in production of knowledge." *Science*, 316(5827): 1036–1039.

Wu, Lingfei, Dashun Wang, and James A Evans. 2019. "Large teams develop and small teams disrupt science and technology." *Nature*, 566(7744): 378–382.

Yin, Yian, Jian Gao, Benjamin F Jones, and Dashun Wang. 2021. "Coevolution of policy and science during the pandemic." *Science*, 371(6525): 128–130.

Yin, Yian, Yuxiao Dong, Kuansan Wang, Dashun Wang, and Benjamin F Jones. 2022. "Public use and public funding of science." *Nature human behaviour*, 6(10): 1344–1350.

Zacchia, Paolo. 2020. "Knowledge spillovers through networks of scientists." *The Review of economic studies*, 87(4): 1989–2018.

Appendix

(For Online Publication Only)

A.1. Details on Network Construction

In this appendix, we provide details on our network construction.

A.1.1. Text Representation and TFIDF

Each document (a patent or a scientific publication) is represented as a bag of words drawn from its title and abstract, using n-grams of length up to three. Pooling the patent and science corpora yields a very large raw vocabulary, which we prune in two stages. First, we drop terms that appear in fewer than 500 documents and terms that appear in more than 5% of documents, retaining approximately 2.17 million terms; the lower cutoff removes idiosyncratic noise words and the upper cutoff removes generic boilerplate that carries no field-specific signal. Second, for our baseline vocabulary we keep the subsample of terms whose document share is greater than 0.005% in at least one of the two corpora, leaving a working vocabulary of approximately 250,000 terms.

We measure the word-frequency vector $\boldsymbol{v}_{it} \in \mathbb{R}^{W\times 1}$ for field i in year t with TFIDF weighting. For a document d, the term frequency of word w is its count divided by the document length,

$$\mathrm{tf}_{w,d} = \frac{\mathrm{count}_{w,d}}{\sum_{w'} \mathrm{count}_{w',d}}.$$

The inverse-document-frequency weight is a *domain frequency mixture* that treats the patent and science corpora symmetrically. Let p_w^{pat} and p_w^{sci} denote the share of patent documents and of science documents, respectively, that contain word w. With a mixing weight $\alpha = 0.5$, the idf weight is

$$\mathrm{idf}_w = \log\left(\frac{1}{\alpha\, p_w^{\mathrm{pat}} + (1-\alpha)\, p_w^{\mathrm{sci}}}\right), \qquad \alpha = 0.5. \tag{A1}$$

Anchoring the idf to a fixed 50/50 blend of patent and science document frequencies, rather than to the pooled corpus, prevents the far larger science corpus from dominating the term weights and keeps a term's weight stable as the relative volume of patents and publications shifts over time. The document TFIDF vector is $\text{tfidf}_{w,d} = \text{tf}_{w,d} \cdot \text{idf}_w$.

We aggregate documents to the field-year level as a document-weighted average, where each document carries a fractional field-membership weight (a document assigned to several fields splits its weight equally across them):

$$\boldsymbol{v}_{it}[w] = \sum_{d \in (i,t)} \frac{\text{weight}_d}{\sum_{d' \in (i,t)} \text{weight}_{d'}} \, \text{tfidf}_{w,d}. \tag{A2}$$

We then rescale each field-year vector to unit Euclidean norm, $\boldsymbol{v}_{it} \leftarrow \boldsymbol{v}_{it}/\|\boldsymbol{v}_{it}\|_2$, so that the measure reflects the relative composition of a field's vocabulary rather than the raw volume of documents behind it. Finally, field-years supported by fewer than 20 documents are treated as unobserved and set to zero, since a handful of documents cannot reliably pin down a 250,000-dimensional frequency vector.

A.1.2. The Diffusion Estimator and Its Interpretation

Knowledge diffusion is inferred from the lead–lag structure of text similarity: field j diffuses knowledge to field i when field i's current vocabulary moves toward field j's past vocabulary. Collecting the field-year vectors into the frequency matrix $\boldsymbol{V}_t \equiv [\boldsymbol{v}_{1t}, \ldots, \boldsymbol{v}_{Nt}]'$, define the lagged and contemporaneous similarity matrices $\boldsymbol{P}_{t,t-1} = \boldsymbol{V}_t \boldsymbol{V}_{t-1}'$ and $\boldsymbol{P}_t = \boldsymbol{V}_t \boldsymbol{V}_t'$, so that $[\boldsymbol{P}_{t,t-1}]_{ij} = \boldsymbol{v}_{it}' \boldsymbol{v}_{jt-1}$ is the cross-temporal similarity of field i today with field j a year earlier. Raw lagged similarity conflates genuine diffusion with the static overlap that historically related fields always share, so we partial the contemporaneous correlation structure out of the upstream fields. The diffusion matrix is

$$\boldsymbol{\Omega}_t = \boldsymbol{P}_{t,t-1} \, \boldsymbol{P}_{t-1}^{-1}. \tag{A3}$$

The right-multiplication by $\boldsymbol{P}_{t-1}^{-1}$ decorrelates the upstream fields, so $\omega_{ij,t}$ captures the *net* predictive content of field j for field i's future trajectory, conditional on all other fields. Equation (A3) is exactly the coefficient matrix of the first-order vector autoregression $\boldsymbol{V}_t = \boldsymbol{\Omega}_t \boldsymbol{V}_{t-1} + \boldsymbol{\Upsilon}_t$ that regresses each field's current vector on the lagged vectors of all fields, and coincides with the Yule–Walker moment estimator for VAR coefficients; the right-multiplication by $\boldsymbol{P}_{t-1}^{-1}$ implements the Frisch–Waugh–Lovell partialling that makes $\omega_{ij,t}$ measure what field j contributes that no other field can substitute for.

We estimate this VAR as a penalized, fixed-effects word-level regression. Because the timing of diffusion is not known a priori, we take a mean-lag approach: for each upstream field j we first average its lagged word vectors over $\tau = 1, \ldots, 10$, $\bar{v}_{jw,t} = \frac{1}{10}\sum_{\tau=1}^{10} v_{jw,t-\tau}$, and then run a single regression on these averaged predictors. For each downstream field i and year t, the estimating equation is

$$v_{iw,t} = \sum_{j} \omega_{ij,t}\, \bar{v}_{jw,t} + \alpha_{iw} + \alpha_{it} + \varepsilon_{iw,t}, \qquad \text{(A4)}$$

where the observation is a (word, year) pair, the field-by-word fixed effect α_{iw} absorbs time-invariant vocabulary differences, and the field-by-year fixed effect α_{it} normalizes the TFIDF vector within each field-year. Averaging the predictors before estimation, rather than estimating each lag separately and averaging coefficients afterward, yields one coefficient $\omega_{ij,t}$ per field pair from a single fixed-effects regression. The baseline model uses a ten-year estimation window ending in year t, with the coefficients smoothed on a five-year calendar rolling window. Because each regression estimates N coefficients, we regularize with a fixed ℓ_2 (ridge) penalty: after partialling out the fixed effects (represented by the annihilator $\boldsymbol{M}_\alpha$), the penalized normal equations for the coefficient vector are

$$\big(\boldsymbol{X}'\boldsymbol{M}_\alpha\boldsymbol{X} + c\,\boldsymbol{I}\big)\,\boldsymbol{\omega}_{i,t} = \boldsymbol{X}'\boldsymbol{M}_\alpha\boldsymbol{y}, \qquad c = 0.2, \qquad \text{(A5)}$$

where $\boldsymbol{X}$ collects the lagged field vectors and $\boldsymbol{y}$ the current vector of field i. The penalty $c = 0.2$ is held fixed, not cross-validated; it is calibrated to roughly match the median diagonal element of $\boldsymbol{X}'\boldsymbol{M}_\alpha\boldsymbol{X}$, so the shrinkage is of the same order as the signal in the design matrix. The ridge penalty is the Lagrangian dual of a norm constraint on the coefficient row, so the estimator inherits a shrinkage interpretation while remaining a single, transparent linear system; it is not a hard-norm constraint.

The smoothed raw coefficients are then mapped to a non-negative, row-stochastic network: entries below zero are winsorized to zero and each downstream field's row is rescaled to sum to one,

$$\omega_{ij,t} = \frac{\max(\hat{\omega}_{ij,t},\, 0)}{\sum_{j'} \max(\hat{\omega}_{ij',t},\, 0)}.$$

This regularized fixed-effects VAR projection, lag-averaged, winsorized, and row-normalized, is the network object $\boldsymbol{\Omega}_t$ used throughout the paper.

A.1.3. Monte Carlo Simulation

We conduct a simulation exercise to confirm that our measure can recover the underlying innovation network from text representations. The simulation is designed to closely follow the knowledge diffusion model described in Section 3.1.

Setup In the simulation, we consider a setting with N fields and a vocabulary of $W = 1000N$ words, such that each field is associated with $1,000$ unique field-specific words. At each time period t, every field i generates $D_{it} = 10,000$ documents, and the number of words per document is randomly drawn from a uniform distribution between 5 and 30.

Let $\boldsymbol{\Omega} \in \mathbb{R}^{N\times N}$ denote the underlying true network, where each element ω_{ij} represents the knowledge flows from field j to field i. We assume a one-period lag structure in knowledge diffusion, meaning that the knowledge at time t directly relies on the knowledge at time $t-1$.

Let $\boldsymbol{v}_{it} \in \mathbb{R}^{W\times 1}$ represent the word-frequency vector for field i in period t, which is the empirical probability distribution over words with $\mathbf{1}'\boldsymbol{v}_{it} = 1$. Let $\boldsymbol{V}_t \equiv [\boldsymbol{v}_{1t}, \ldots, \boldsymbol{v}_{Nt}]'$ denote the frequency matrix across fields at time t.

Simulation Timeline The simulation proceeds in two stages. In the initialization period ($t = 0$), for each field i, we randomly generate 10,000 documents using a uniform distribution over the 1,000 field-specific words. This results in an initial term-frequency matrix $\boldsymbol{V}_0$, where the word distribution $v_{i,0}$ for each field i is concentrated only on its own vocabulary.

Starting from $t = 1$, the simulation allows for knowledge diffusion across fields. In each subsequent period $t > 0$, documents for field i are generated using a multinomial process with the distribution updated from the knowledge diffusion process $\boldsymbol{\Omega}_{i,\cdot}\boldsymbol{V}_{t-1} + u'_{it}$. Here, $\boldsymbol{\Omega}_{i,\cdot}$ is the i-th row of the true innovation network $\boldsymbol{\Omega}$, and u_{it} represents random noise orthogonal to $\boldsymbol{V}_{t-1}$, capturing idiosyncratic innovations uncorrelated with past knowledge. This implies that the underlying word distribution in each field is a weighted combination of the word distributions across all fields in the previous period, with weights determined by the true innovation network Ω. We then obtain the term-frequency matrix $\boldsymbol{V}_t$ using the generated documents, and represent it using the recursive structure $\boldsymbol{V}_t = \boldsymbol{\Omega}_t \boldsymbol{V}_{t-1} + \boldsymbol{\Upsilon}_t$, where $\boldsymbol{\Upsilon}_t$ is the matrix of noise terms.

We simulate the data over 1,000 periods, and use the simulated data in the last 100 periods to estimate the innovation network.

Constructing the Innovation Network To recover the innovation network, we compute the cosine similarity between term-frequency matrices. We compute the cosine similarity matrix with one-period time lag, $\boldsymbol{P}_{t,t-1} = \boldsymbol{V}_t \boldsymbol{V}'_{t-1}$, and the contemporaneous cosine similarity matrix $\boldsymbol{P}_t = \boldsymbol{V}_t \boldsymbol{V}'_t$. The estimated innovation network is computed as $\hat{\boldsymbol{\Omega}}_t = \boldsymbol{P}_{t,t-1} \boldsymbol{P}^{-1}_{t-1}$.

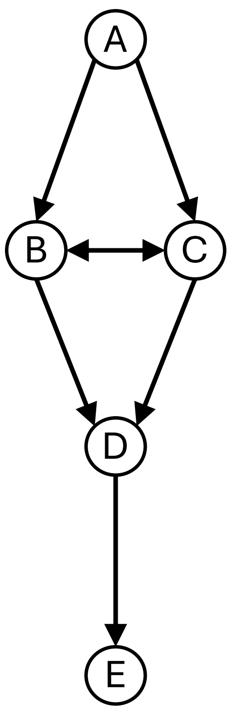
Simulation Results Figure A.1 illustrates the outcome of our simulation exercise when considering $N = 5$ fields. The left panel presents the stylized example of the true innovation network among 5 fields (A–E). Field A serves as a dominant upstream source, diffusing knowledge to fields B and C, which in turn influence D and E. Field E is positioned at the downstream end of the network. The true innovation network is presented in the Panel (b.1), with a (block) lower-triangular structure.

Panel (b.2) shows the estimated innovation network recovered using our methodology. The estimated matrix accurately approximates the structure of the true network (i.e., upstream vs. downstream fields), with a correlation coefficient of 0.93. For example, field A maintains a high self-weight and substantial outgoing links to B and C, while field D shows strong inbound weights from B and C and moderate self-dependence. Notably, the estimation closely recovers each value of the knowledge flows in the network. The centrality measures derived from the estimated network also closely track those from the true network, preserving the rank order and resulting in a correlation coefficient of 0.99.

Panel (b.3) shows the cosine similarity matrix, calculated directly from the term-frequency distributions across fields. It shows that the similarity matrix is nearly uniform and uninformative. As discussed in Section 3.1, the cosine similarity matrix mixes different sources of information, including persistent similarities between fields rather than genuine knowledge spillovers. Without modeling the dynamic diffusion process, simple pairwise similarity measures fail to uncover the knowledge diffusion structure embedded in the innovation network.

Figure A.1. Simulation Results

Panel (a): Simulation Example



Panel (b): Estimated Network

	A	B	C	D	E	Katz Centrality
Panel (b.1): True Innovation Network						
A	1.000	0.000	0.000	0.000	0.000	0.512
B	0.364	0.455	0.182	0.000	0.000	0.143
C	0.364	0.182	0.455	0.000	0.000	0.143
D	0.000	0.308	0.308	0.385	0.000	0.107
E	0.000	0.000	0.000	0.375	0.625	0.094
Panel (b.2): Estimated Innovation Network						
A	0.755	0.103	0.109	0.019	0.014	0.371
B	0.330	0.277	0.282	0.091	0.020	0.206
C	0.331	0.289	0.278	0.084	0.018	0.207
D	0.080	0.333	0.339	0.236	0.011	0.132
E	0.013	0.103	0.101	0.417	0.367	0.084
Panel (b.3): Similarity Matrix						
A	0.201	0.200	0.200	0.200	0.199	0.200
B	0.200	0.200	0.200	0.200	0.199	0.200
C	0.200	0.200	0.200	0.200	0.199	0.200
D	0.200	0.200	0.200	0.200	0.200	0.200
E	0.199	0.200	0.200	0.200	0.200	0.200

Notes. This figure shows the simulation results. Panel (a) shows the simulation example and Panel (b) shows the true network, the estimated network from our methodology, and the cosine similarity matrix.

A.1.4. Remarks on TF Vectors and Generative Text Models

Drawing each document as $x_d \sim \text{Multinomial}(v_{it}, n_d)$ ensures $\mathbb{E}[x_d/n_d] = v_{it}$, so the empirical average of document-level word shares is an unbiased estimator of the latent distribution. This is the same data-generating logic that underpins topic models such as LDA; here, however, the objective is to back out cross-field spillovers rather than latent topics.

A Side Note on Empirical TF Vector and Generative Models of Text We now link the empirical TF vector to the true distribution v_{it}: Given field i at time t, draw $D_{i,t}$

documents from the list of words following the probability distribution $v_{i,t}$. That is equivalent to assuming that each document $d = 1, \ldots, D_{i,t}$ follows a multinomial count distribution with parameter $v_{i,t}$:

$$x_d \sim \text{Multinomial}(v_{i,t}, n_d), \tag{A6}$$

where x_d is the word-count vector of document d, and n_d is the number of words in document d. Then the expected frequency vector of document d is $\mathbb{E}\left[\frac{x_d}{n_d}\right] = v_{i,t}$. The empirical estimate of this expectation is exactly the average word-frequency vector of each document in field i at time t.

In generative models of text such as LDA topic models, the multinomial distribution is used to model the generation of words in a document, assuming that each word is drawn independently from a distribution. It models the data-generating process for text with a set of probabilistic rules.

A.1.5. Event Studies: Tracing Knowledge Diffusion Using Words

A direct test of whether the text-based network measures genuine knowledge flows, rather than static similarity, is whether the distinctive vocabulary of a rising field subsequently spreads to exactly the fields the network marks as downstream. We implement this test as a keyword-diffusion event study anchored on discrete innovation episodes. An innovation episode is a sustained interval during which a field is unusually central in the innovation network—a wave in which it becomes a major upstream source of knowledge flowing to others. The procedure has three steps: identify episodes, extract each episode's distinctive keywords, and trace those keywords' diffusion across fields.

Detecting episodes. Because the diffusion matrix $\mathbf{\Omega}_t$ is row-normalized, the cross-field mean of Katz centrality is mechanically pinned at $1/N$ in every year, so a field's importance is inherently a cross-sectional question. We therefore convert each field's centrality c_{it} to a within-year z-score,

$$z_{it} = \frac{c_{it} - \mu_t}{\sigma_t}, \qquad \mu_t = \frac{1}{N}\sum_j c_{jt}, \quad \sigma_t = \mathrm{sd}_j\big(c_{jt}\big), \tag{A7}$$

which carries the same units across all years. We smooth z_{it} with a centered five-year moving average, re-standardize, and fit a Markov regime-switching model (a constant mean per regime, with $k \in \{2, 3\}$ regimes selected by BIC over 30 random restarts) to each field's smoothed series. A candidate episode is a maximal run of years, of length between 3 and 25, in which the smoothed probability of the highest-mean regime exceeds 0.5. An episode is retained as "major" when its peak clears a within-year height threshold *and* the field rises to that peak from a low base:

$$z^{\text{peak}} \geq \underline{z}_{\text{peak}}, \quad z^{\text{peak}} - z^{\text{pre}} \geq \underline{z}_{\text{rise}}, \quad z^{\text{pre}} \leq \bar{z}_{\text{pre}}, \tag{A8}$$

where z^{pre} is the field's mean z over the five years before emergence. The pre-emergence gate encodes an acceleration logic: selection rewards fields that genuinely rise from obscurity, not fields that are already central and merely remain so. The baseline thresholds are $\underline{z}_{\text{peak}} = 3.0$, $\underline{z}_{\text{rise}} = 1.0$, and $\bar{z}_{\text{pre}} = 1.75$, which yield 95 selected episodes (71 science, 24 technology). The detected sequence is substantively coherent, running from seismology (peak 1887), radiochemistry (1908), and biochemistry (1941) through quantum electrodynamics (1954), nuclear physics (1960), computer architecture (mid-1980s), the World Wide Web (2003), nanotechnology (2007), and machine learning (1995 and again 2024).

Distinctive keywords. For each episode we identify the words that most characterize the focal field while the episode is active. The innovation-network regression, which explains each field's word-frequency vector from the lagged vectors of all fields, yields, per field-year and word w, the share R^2_{iw} of the field's word-frequency variation that w accounts for. We average R^2_{iw} over the episode's years, up-weighting the emergence-to-peak window, and drop "common" words, defined as those whose document share already exceeded 0.5% in a window centered 25 years before emergence. Ranking the remaining words by pre-peak R^2 and keeping the top 50 yields, for each episode, a list of terms that were distinctive to the focal field precisely when it became central.

Tracing diffusion. For each episode we take its top 50 keywords and compute, for every field and every year in the window [peak − 50, peak + 30], the share of that field's documents containing at least one of the keywords. Grouping recipient fields by their network dependence on the focal field and plotting adoption against years-to-peak produces the event study. Figure A.4 shows six representative episodes spanning a century of innovation, from combustion engines (1917) and nuclear physics (1960) to the World Wide Web (2003), electric communication (2005), nanotechnology (2007), and data science (2020). In every panel the same pattern recurs: the distinctive keywords appear first in the focal field, then in network-proximate fields, and only later—if at all—in distant ones. This temporal ordering aligns with the directional structure of $\mathbf{\Omega}_t$, providing direct evidence that the text-based network captures genuine, time-ordered knowledge flows rather than static co-occurrence.

A.2. Proofs

This appendix proves Propositions 1–3. The argument for Proposition 1 proceeds in four steps, each stated as a lemma; the remaining proofs build on these lemmas. Lemma 1 establishes the key properties of the propagation matrix: it is entrywise non-negative and decays exponentially at rate ρ, which guarantees that all integrals below converge. Lemma 2 verifies that the perturbed knowledge paths are well defined and that welfare is finite. Lemma 3 shows that the knowledge path is differentiable in the perturbation amplitude ε and derives the variational equation it satisfies. Lemma 4 justifies differentiating under the welfare integral. The proofs of the propositions then assemble these ingredients. Importantly, no commutativity or diagonalizability of the network path $\{\boldsymbol{\Omega}(s)\}$ is required at any step: all bounds follow from the row-stochasticity of $\boldsymbol{\Omega}(s)$, which is itself implied by constant returns to scale in the knowledge aggregators.

Notation and regularity conditions. Fix the base time t and write $\boldsymbol{z}_s \equiv (\ln q_{1s}, \ldots, \ln q_{Ns})'$ for the log knowledge stocks. For each sector i, define $G_i(\boldsymbol{z}) \equiv \ln \mathcal{O}_i(\{z_j\}_{j=1}^N)$ and $\boldsymbol{G}(\boldsymbol{z}) \equiv (G_1(\boldsymbol{z}), \ldots, G_N(\boldsymbol{z}))'$, so that the knowledge dynamics (8) read

$$\dot{\boldsymbol{z}}_s = \lambda \left[\ln \boldsymbol{x}_s + \boldsymbol{G}(\boldsymbol{z}_s) - \boldsymbol{z}_s\right], \tag{A9}$$

where $\ln \boldsymbol{x}_s$ is taken componentwise. By Definition 1, the innovation network is the Jacobian $\boldsymbol{\Omega}(\boldsymbol{z}) = D\boldsymbol{G}(\boldsymbol{z})$; we write $\boldsymbol{\Omega}(s) \equiv \boldsymbol{\Omega}(\boldsymbol{z}_s)$ for its value along the unperturbed path. The two maintained assumptions from Section 5.1 translate as follows. Non-negative cross-partial elasticities mean $\boldsymbol{\Omega}(\boldsymbol{z}) \geq 0$ entrywise at every $\boldsymbol{z}$. Constant returns to scale means $\mathcal{O}_i$ is homogeneous of degree one in the knowledge stocks, so $G_i(\boldsymbol{z} + a\mathbf{1}) = G_i(\boldsymbol{z}) + a$ for every scalar a; differentiating in a yields the Euler identity

$$\sum_{j=1}^{N} \Omega_{ij}(\boldsymbol{z}) = 1 \qquad \text{for every } \boldsymbol{z} \text{ and } i, \tag{A10}$$

that is, $\boldsymbol{\Omega}(\boldsymbol{z})$ is row-stochastic at every state, not merely along the equilibrium path. Throughout we impose two mild regularity conditions: (R1) each $\mathcal{O}_i$ is positive and continuously differentiable; (R2) the baseline R&D path $s \mapsto \boldsymbol{x}_s$ is measurable, takes values in $[\underline{x}, \bar{x}]^N$ for some $0 < \underline{x} \leq \bar{x} < \infty$, and is right-continuous at $s = t$. Given a bounded direction $\boldsymbol{h}$, we fix $\varepsilon_0 > 0$ small enough that $\underline{x}/2 \leq x_{is} + \varepsilon h_i \leq 2\bar{x}$ for all $|\varepsilon| \leq \varepsilon_0$, all i, and all s in the perturbation window.

A first consequence of (A10) is that $\boldsymbol{G}$ is globally Lipschitz: by the mean value theorem applied componentwise, for any $\boldsymbol{z}, \boldsymbol{w}$ there exist intermediate points $\boldsymbol{\xi}_i$ such that $G_i(\boldsymbol{z}) - G_i(\boldsymbol{w}) = \nabla G_i(\boldsymbol{\xi}_i)'(\boldsymbol{z} - \boldsymbol{w})$, and since $\nabla G_i \geq 0$ sums to one,

$$\|\boldsymbol{G}(\boldsymbol{z}) - \boldsymbol{G}(\boldsymbol{w})\|_\infty \leq \|\boldsymbol{z} - \boldsymbol{w}\|_\infty. \tag{A11}$$

Lemma 1 (Properties of the propagation matrix)**.** *Let $s \mapsto \boldsymbol{\Omega}(s)$ be any measurable path of non-negative, row-stochastic matrices, let $\boldsymbol{A}(s) \equiv \lambda(\boldsymbol{\Omega}(s) - \boldsymbol{I})$, and let $\boldsymbol{\Psi}(s, \tau)$ denote the transition matrix of the linear system $\dot{\boldsymbol{v}} = \boldsymbol{A}(s)\boldsymbol{v}$—the matrix that maps a vector at time τ into the solution at time s, so that $\partial_s \boldsymbol{\Psi}(s, \tau) = \boldsymbol{A}(s)\boldsymbol{\Psi}(s, \tau)$ with $\boldsymbol{\Psi}(\tau, \tau) = \boldsymbol{I}$. Then: (i) $\boldsymbol{\Psi}(s, \tau) \geq 0$ entrywise and $\|\boldsymbol{\Psi}(s, \tau)\|_\infty \leq 1$ for all $s \geq \tau$; (ii) $\boldsymbol{\Phi}(t, s) \equiv e^{-\rho(s-t)}\boldsymbol{\Psi}(s, t)$ is the unique solution of the matrix differential equation (15), and satisfies $\boldsymbol{\Phi}(t, s) \geq 0$ and $\|\boldsymbol{\Phi}(t, s)\|_\infty \leq e^{-\rho(s-t)}$.*

Proof. Since $(\rho + \lambda)/(1 + \rho/\lambda) = \lambda$, the drift matrix in (15) satisfies the identity

$$-(\rho + \lambda)\left(\boldsymbol{I} - \frac{\boldsymbol{\Omega}(s)}{1 + \rho/\lambda}\right) = -\rho\boldsymbol{I} + \lambda(\boldsymbol{\Omega}(s) - \boldsymbol{I}) = -\rho\boldsymbol{I} + \boldsymbol{A}(s),$$

so $e^{-\rho(s-t)}\boldsymbol{\Psi}(s, t)$ solves (15); uniqueness follows from linearity with bounded measurable coefficients. For (i), note that $\boldsymbol{A}(s)$ has non-negative off-diagonal entries (it is a Metzler matrix), because $\boldsymbol{\Omega}(s) \geq 0$; hence the linear flow it generates preserves the non-negative orthant, and $\boldsymbol{\Psi}(s, \tau) \geq 0$ entrywise (Coppel, 1965). For the norm bound, we use the logarithmic norm of $\boldsymbol{A}(s)$: the quantity $\mu_\infty(\boldsymbol{A}) \equiv \max_i [A_{ii} + \sum_{j \neq i} |A_{ij}|]$, which bounds the instantaneous growth rate of the flow measured in the ℓ_∞ norm. Here

$$\mu_\infty\big(\boldsymbol{A}(s)\big) = \lambda \max_i \Big[\sum_{j=1}^{N} \Omega_{ij}(s) - 1\Big] = 0,$$

where the last equality uses row-stochasticity (A10). Coppel's inequality, $\|\boldsymbol{\Psi}(s,\tau)\|_\infty \leq \exp\big(\int_\tau^s \mu_\infty(\boldsymbol{A}(u))\,du\big)$, then delivers $\|\boldsymbol{\Psi}(s,\tau)\|_\infty \leq 1$ (Coppel, 1965). Part (ii) follows immediately. □

Lemma 2 (Well-posedness and finite welfare)**.** *Under (R1)–(R2), for every $|\varepsilon| \leq \varepsilon_0$ the perturbed path $\boldsymbol{z}_s^{\varepsilon,\delta,\boldsymbol{h}}$ solving (A9) with R&D input $\boldsymbol{x}_s^{\varepsilon,\delta,\boldsymbol{h}}$ from (10) exists, is unique and locally absolutely continuous on $[t,\infty)$, and grows at most linearly: there is a constant C_z, independent of ε, with $\|\boldsymbol{z}_s^{\varepsilon,\delta,\boldsymbol{h}}\|_\infty \leq \|\boldsymbol{z}_t\|_\infty + C_z(s-t)$. Consequently the welfare integral $V_t^{\varepsilon,\delta,\boldsymbol{h}}$ converges absolutely.*

Proof. The right-hand side of (A9) is measurable in s (by (R2)) and globally Lipschitz in $\boldsymbol{z}$ with constant 2λ, by (A11); existence and uniqueness on $[t,\infty)$ follow from the standard existence theorem for differential equations whose time dependence is merely measurable (Carathéodory's theorem; see Hale, 1969). For the growth bound, let $m(s)$ denote the largest component of $\boldsymbol{z}_s$; it is absolutely continuous, and almost everywhere its derivative equals that of a component attaining the maximum. At such a component i, monotonicity of G_i (non-negative cross-partials) and the translation property give $G_i(\boldsymbol{z}_s) \leq G_i(m(s)\mathbf{1}) = G_i(\mathbf{0}) + m(s)$, so $\dot{z}_{is} = \lambda[\ln x_{is} + G_i(\boldsymbol{z}_s) - m(s)] \leq \lambda[\ln(2\bar{x}) + \max_j G_j(\mathbf{0})]$. The smallest component satisfies the symmetric lower bound (with $\ln(\underline{x}/2)$ and $\min_j G_j(\mathbf{0})$), so $\|\boldsymbol{z}_s\|_\infty$ grows at most linearly, at a rate C_z that depends only on λ, $\bar{x}$, $\underline{x}$, and $\boldsymbol{G}(\mathbf{0})$, uniformly in ε. Finally, since the fixed labor allocation is bounded, the welfare integrand in (5) is bounded by $e^{-\rho(s-t)}$ times a linear function of $(s-t)$, which is integrable on $[t,\infty)$. □

Lemma 3 (Differentiability and the variational equation)**.** *Under (R1)–(R2), for every $s \geq t$ the map $\varepsilon \mapsto \boldsymbol{z}_s^{\varepsilon,\delta,\boldsymbol{h}}$ is differentiable at $\varepsilon = 0$, and $\boldsymbol{v}_s \equiv \partial_\varepsilon \boldsymbol{z}_s^{\varepsilon,\delta,\boldsymbol{h}}\big|_{\varepsilon=0}$ solves the linear variational equation*

$$\dot{\boldsymbol{v}}_s = \lambda\big(\boldsymbol{\Omega}(s) - \boldsymbol{I}\big)\boldsymbol{v}_s + \lambda\left(\boldsymbol{h} \oslash \boldsymbol{x}_s\right)\mathbf{1}_{t\leq s<t+\delta}, \qquad \boldsymbol{v}_t = \mathbf{0}, \tag{A12}$$

where $\oslash$ denotes componentwise division and $\boldsymbol{\Omega}(s)$ is evaluated along the unperturbed path. By the variation-of-constants formula (the solution formula for a linear equation with a forcing term),

$$\boldsymbol{v}_s = \lambda \int_t^{\min\{s,\,t+\delta\}} \boldsymbol{\Psi}(s,\tau)\left(\boldsymbol{h} \oslash \boldsymbol{x}_\tau\right)d\tau. \tag{A13}$$

Proof. Write the perturbed dynamics as $\dot{\boldsymbol{z}} = F(s, \boldsymbol{z}, \varepsilon)$ with $F(s, \boldsymbol{z}, \varepsilon) = \lambda[\ln(\boldsymbol{x}_s + \varepsilon \boldsymbol{h}\, \mathbf{1}_{t \le s < t+\delta}) + \boldsymbol{G}(\boldsymbol{z}) - \boldsymbol{z}]$. For each fixed $(\boldsymbol{z}, \varepsilon)$, F is measurable in s; for each fixed s, F is continuously differentiable in $(\boldsymbol{z}, \varepsilon)$ on $\{|\varepsilon| \le \varepsilon_0\}$: differentiability in $\boldsymbol{z}$ holds by (R1) with $\partial_{\boldsymbol{z}} F = \lambda(\boldsymbol{\Omega}(\boldsymbol{z}) - \boldsymbol{I})$, and differentiability in ε holds because $x_{is} + \varepsilon h_i \ge \underline{x}/2 > 0$ keeps the logarithm smooth, with $\partial_\varepsilon F = \lambda \big(\boldsymbol{h} \oslash (\boldsymbol{x}_s + \varepsilon \boldsymbol{h})\big)\, \mathbf{1}_{t \le s < t+\delta}$; both partial derivatives are bounded on compact time intervals. Under these hypotheses the solution depends differentiably on the parameter: writing the equations for $\boldsymbol{z}^{\varepsilon,\delta,\boldsymbol{h}}$ and for the solution of (A12) in integral form, subtracting, and applying Grönwall's inequality shows that the difference quotient $(\boldsymbol{z}_s^{\varepsilon,\delta,\boldsymbol{h}} - \boldsymbol{z}_s)/\varepsilon$ converges, as $\varepsilon \to 0$, to the solution of the variational equation obtained by evaluating $\partial_{\boldsymbol{z}} F$ and $\partial_\varepsilon F$ along the unperturbed path, which is precisely (A12) (see, e.g., Hale, 1969, Ch. I). Equation (A13) is the variation-of-constants representation of the unique solution of this linear equation with zero initial condition. $\square$

Lemma 4 (Differentiating under the welfare integral)**.** *Under (R1)–(R2),*

$$\left.\frac{\partial V_t^{\varepsilon,\delta,\boldsymbol{h}}}{\partial \varepsilon}\right|_{\varepsilon=0} = \psi \int_t^\infty e^{-\rho(s-t)}\, \boldsymbol{\beta}' \boldsymbol{v}_s \, ds. \tag{A14}$$

Proof. Since the labor terms in (5) do not depend on ε, it suffices to interchange ∂_ε and the integral in $\int_t^\infty e^{-\rho(s-t)} \psi\, \boldsymbol{\beta}' \boldsymbol{z}_s^{\varepsilon,\delta,\boldsymbol{h}}\, ds$, which we do by dominated convergence applied to the difference quotients $\boldsymbol{d}_s^\varepsilon \equiv (\boldsymbol{z}_s^{\varepsilon,\delta,\boldsymbol{h}} - \boldsymbol{z}_s)/\varepsilon$. Subtracting the state equations for $\boldsymbol{z}^{\varepsilon,\delta,\boldsymbol{h}}$ and $\boldsymbol{z}$ and dividing by ε,

$$\dot{\boldsymbol{d}}_s^\varepsilon = \lambda \left[\frac{\boldsymbol{G}(\boldsymbol{z}_s^{\varepsilon,\delta,\boldsymbol{h}}) - \boldsymbol{G}(\boldsymbol{z}_s)}{\varepsilon} - \boldsymbol{d}_s^\varepsilon \right] + \lambda\, \boldsymbol{q}^\varepsilon(s), \qquad \boldsymbol{d}_t^\varepsilon = \mathbf{0},$$

where $q_i^\varepsilon(s) = \varepsilon^{-1}\big[\ln(x_{is} + \varepsilon h_i) - \ln x_{is}\big]\, \mathbf{1}_{t \le s < t+\delta}$. By the mean value theorem applied to the logarithm on an interval bounded below by $\underline{x}/2$, $|q_i^\varepsilon(s)| \le 2|h_i|/\underline{x}$ uniformly in $|\varepsilon| \le \varepsilon_0$. To bound $\boldsymbol{d}^\varepsilon$, let $m(s)$ denote its largest component; almost everywhere, $\dot{m}$ equals the derivative of a component i attaining the maximum. By the mean value theorem, $[\boldsymbol{G}(\boldsymbol{z}_s^{\varepsilon,\delta,\boldsymbol{h}}) - \boldsymbol{G}(\boldsymbol{z}_s)]_i/\varepsilon = \nabla G_i(\boldsymbol{\xi}_i)' \boldsymbol{d}_s^\varepsilon$ for an intermediate point $\boldsymbol{\xi}_i$, and since ∇G_i is nonnegative and sums to one, this is at most $m(s)$; hence $\dot{m} \le \lambda[m - m] + \lambda \|\boldsymbol{q}^\varepsilon\|_\infty = \lambda \|\boldsymbol{q}^\varepsilon\|_\infty$.

The symmetric argument bounds the smallest component from below, so

$$\|\boldsymbol{d}_s^\varepsilon\|_\infty \;\le\; \lambda \int_t^{t+\delta} \|\boldsymbol{q}^\varepsilon(\tau)\|_\infty \, d\tau \;\le\; \frac{2\lambda\delta\|\boldsymbol{h}\|_\infty}{\underline{x}} \qquad \text{for all } s \ge t \text{ and } |\varepsilon| \le \varepsilon_0.$$

Hence $|e^{-\rho(s-t)}\boldsymbol{\beta}'\boldsymbol{d}_s^\varepsilon| \le Ce^{-\rho(s-t)}$ with C independent of ε, an integrable majorant on $[t,\infty)$. Since $\boldsymbol{d}_s^\varepsilon \to \boldsymbol{v}_s$ pointwise as $\varepsilon \to 0$ by Lemma 3, the dominated convergence theorem delivers (A14). □

Proof of Proposition 1

Proof. Combining Lemmas 3 and 4 and substituting the representation (A13),

$$\left.\frac{\partial V_t^{\varepsilon,\delta,\boldsymbol{h}}}{\partial \varepsilon}\right|_{\varepsilon=0} = \lambda\psi \int_t^\infty \int_t^{\min\{s,\,t+\delta\}} e^{-\rho(s-t)}\, \boldsymbol{\beta}'\, \boldsymbol{\Psi}(s,\tau)\, \big(\boldsymbol{h} \oslash \boldsymbol{x}_\tau\big)\, d\tau\, ds.$$

By Lemma 1(i) and (R2), the integrand is bounded in absolute value by $2e^{-\rho(s-t)}\|\boldsymbol{h}\|_\infty/\underline{x}$ (using $\sum_i \beta_i = 1$), so the double integral converges absolutely and Fubini's theorem permits exchanging the order of integration:

$$\left.\frac{\partial V_t^{\varepsilon,\delta,\boldsymbol{h}}}{\partial \varepsilon}\right|_{\varepsilon=0} = \lambda\psi \int_t^{t+\delta} g(\tau)\, d\tau, \qquad g(\tau) \equiv \left[\int_\tau^\infty e^{-\rho(s-t)}\, \boldsymbol{\beta}'\, \boldsymbol{\Psi}(s,\tau)\, ds\right] \big(\boldsymbol{h} \oslash \boldsymbol{x}_\tau\big).$$

The map $\tau \mapsto g(\tau)$ is right-continuous at $\tau = t$: the inner integral is continuous in τ because the transition matrix is continuous in τ uniformly on compact s-intervals (write $\boldsymbol{\Psi}(s,\tau) = \boldsymbol{\Psi}(s,t)\, \boldsymbol{\Psi}(\tau,t)^{-1}$, where $\tau \mapsto \boldsymbol{\Psi}(\tau,t)$ is continuous with invertible values) while the tail is dominated by $e^{-\rho(s-t)}$, and $\boldsymbol{x}_\tau$ is right-continuous at t by (R2). Hence, dividing by δ and letting $\delta \to 0^+$,

$$\lim_{\delta \to 0^+} \frac{1}{\delta}\, \left.\frac{\partial V_t^{\varepsilon,\delta,\boldsymbol{h}}}{\partial \varepsilon}\right|_{\varepsilon=0} = \lambda\psi \left[\int_t^\infty e^{-\rho(s-t)}\, \boldsymbol{\beta}'\, \boldsymbol{\Psi}(s,t)\, ds\right] \big(\boldsymbol{h} \oslash \boldsymbol{x}_t\big).$$

By Lemma 1(ii), $e^{-\rho(s-t)}\boldsymbol{\Psi}(s,t) = \boldsymbol{\Phi}(t,s)$ is exactly the propagation matrix of (15), so the bracketed row vector equals $\boldsymbol{\gamma}_t'$ as defined in (14). Setting $\boldsymbol{h} = \boldsymbol{e}_i$ as in Definition 2 yields

$$\xi_{it} = \lambda\psi \, \frac{\gamma_{it}}{x_{it}}.$$

Since $\lambda\psi > 0$ is common across sectors, this establishes $\xi_{it} \propto_t \gamma_{it}/x_{it}$, which is (13). The convergence of the integral defining $\boldsymbol{\gamma}_t$ and the non-negativity $\gamma_{it} \geq 0$ follow from the bounds in Lemma 1(ii). Note that no commutativity or diagonalizability of $\{\boldsymbol{\Omega}(s)\}$ was used at any point in the argument. $\square$

Proof of Proposition 2

Proof. Fungibility means that at each instant the planner can reallocate R&D across sectors while keeping the total fixed, so the feasible perturbation directions at time t are exactly those $\boldsymbol{h} \in \mathbb{R}^N$ with $\mathbf{1}'\boldsymbol{h} = 0$. Suppose $\{x_{it}\}$ maximizes welfare. For any feasible direction $\boldsymbol{h}$, the localized directional derivative computed in the proof of Proposition 1 equals

$$\xi_t^{\boldsymbol{h}} \equiv \lim_{\delta \to 0^+} \frac{1}{\delta} \left. \frac{\partial V_t^{\varepsilon,\delta,\boldsymbol{h}}}{\partial \varepsilon} \right|_{\varepsilon=0} = \lambda\psi \sum_{i=1}^{N} \frac{\gamma_{it}}{x_{it}} \, h_i.$$

Optimality requires $\xi_t^{\boldsymbol{h}} \leq 0$ for every feasible $\boldsymbol{h}$; applying the same inequality to $-\boldsymbol{h}$, which is also feasible, gives $\xi_t^{\boldsymbol{h}} = 0$. Hence the linear functional $\boldsymbol{h} \mapsto \sum_i (\gamma_{it}/x_{it})\, h_i$ vanishes on the entire hyperplane $\{\boldsymbol{h} : \mathbf{1}'\boldsymbol{h} = 0\}$, which forces the coefficient vector to be proportional to $\mathbf{1}$: the ratio γ_{it}/x_{it}, and therefore the marginal social value ξ_{it}, is equalized across all sectors i at each t. Equivalently, $x_{it} \propto_t \gamma_{it}$, which is (19); dividing by the cross-sector sums yields the share form. Intuitively, if γ_{it}/x_{it} were not equalized, reallocating R&D from a sector with a low ratio toward one with a high ratio would raise welfare to first order, contradicting optimality. $\square$

Proof of Proposition 3

Proof. Let $\omega = (\tau_1, \hat{o}_1, \tau_2, \hat{o}_2, \ldots)$ denote a realization of the arrival times and operator draws. Conditional on ω, the economy is deterministic: the log knowledge stocks follow (A9) with the aggregator profile prevailing in each inter-arrival interval, and the state is continuous at every arrival. We assume in addition that the continuation plan satisfies (R2) along every regime history, and that the aggregator levels are bounded uniformly over all realizable profiles, i.e., $\sup_{\hat{o}} \max_i |\ln \hat{\mathcal{O}}_i(\mathbf{1})| < \infty$; the latter guarantees, by the growth bound of Lemma 2 applied regime by regime, that realized welfare is integrable and the objective is well defined. Since every realizable profile satisfies the maintained assumptions, the Euler identity (A10) holds regime by regime, so the piecewise system is Carathéodory with a globally Lipschitz right-hand side, and Lemmas 1–4 apply pathwise, with the piecewise network $\boldsymbol{\Omega}^{\omega}(s)$ and its propagation matrix $\boldsymbol{\Phi}^{\omega}(t,s)$ satisfying $0 \le \boldsymbol{\Phi}^{\omega}(t,s)$ and $\|\boldsymbol{\Phi}^{\omega}(t,s)\|_{\infty} \le e^{-\rho(s-t)}$ uniformly in ω, because the bounds in Lemma 1 use only row-stochasticity.

Conditional on ω, the proof of Proposition 1 applies without change and yields

$$\lim_{\delta \to 0^+} \frac{1}{\delta} \left. \frac{\partial V_t^{\varepsilon,\delta,\boldsymbol{h}}(\omega)}{\partial \varepsilon} \right|_{\varepsilon=0} = \lambda\psi \left[\int_t^{\infty} \boldsymbol{\beta}' \, \boldsymbol{\Phi}^{\omega}(t,s)\, ds \right] \big(\boldsymbol{h} \oslash \boldsymbol{x}_t \big).$$

The majorants in Lemma 4 and in the proof of Proposition 1 are uniform in ω, so dominated convergence permits interchanging the expectation with the ε-derivative. For the $\delta \to 0^+$ limit, note that $\tau_1 > t$ almost surely, and for each realization the perturbation window $[t, t+\delta)$ lies entirely within the initial regime once $\delta < \tau_1 - t$, so the conditional scaled derivative converges pointwise; realizations in which the window contains one or more arrivals, so that part of the perturbed R&D operates under later regimes, are controlled by the same bounds: the variational argument applies with the piecewise network, the scaled conditional derivative is bounded by $2\lambda\psi\|\boldsymbol{\beta}\|_1\|\boldsymbol{h}\|_{\infty}/(\underline{x}\rho)$ uniformly in (δ,ω), and the event has probability $1 - e^{-\eta\delta} = O(\delta)$. Dominated convergence over ω therefore permits interchanging the expectation with the $\delta \to 0^+$ limit as well. Setting $\boldsymbol{h} = \boldsymbol{e}_i$,

$$\xi_{it} = \lambda\psi \, \frac{\gamma_{it}}{x_{it}}, \qquad \boldsymbol{\gamma}_t' = \mathbb{E}_t \left[\int_t^{\infty} \boldsymbol{\beta}' \, \boldsymbol{\Phi}^{\omega}(t,s)\, ds \right].$$

To obtain the recursive representation (20), condition on the first arrival and draw, $(\tau_1, \hat{o}_1) = (T, \hat{o})$. On $[t, T)$ the path is deterministic under the current profile, so $\boldsymbol{\Phi}^{\omega}(t,s) = \boldsymbol{\Phi}(t,s)$ for $s < T$; for $s \ge T$, the transition property of the piecewise linear system gives

$\mathbf{\Phi}^{\omega}(t,s) = \mathbf{\Phi}^{\theta_T\omega}(T,s)\,\mathbf{\Phi}(t,T)$, where $\theta_T\omega$ denotes the continuation realization after T. By the memorylessness of the Poisson process and the Markov structure of the draws (the shocks are i.i.d., so given the profile $\hat{o}$ prevailing after the arrival, the law of all later draws depends on the history only through $\hat{o}$), conditional on $(T,\hat{o})$ the continuation realization has the law of a fresh realization started at date T from profile $\hat{o}$; the law of iterated expectations therefore gives

$$\mathbb{E}\left[\int_T^{\infty} \boldsymbol{\beta}'\,\mathbf{\Phi}^{\theta_T\omega}(T,s)\,ds \,\middle|\, \tau_1 = T,\ \hat{o}_1 = \hat{o}\right] = \boldsymbol{\gamma}_T'(\hat{o}),$$

the value vector at date T under $\hat{o}$, given the knowledge stocks in place at T (which, conditional on $\tau_1 = T$, are deterministic); by the same argument, this value vector satisfies (20) from T onward. Splitting the integral at T and taking expectations, Tonelli's theorem applied to the non-negative first piece gives

$$\mathbb{E}_t\left[\int_t^{\tau_1} \boldsymbol{\beta}'\,\mathbf{\Phi}(t,s)\,ds\right] = \int_t^{\infty} \Pr(\tau_1 > s)\,\boldsymbol{\beta}'\,\mathbf{\Phi}(t,s)\,ds = \int_t^{\infty} e^{-\eta(s-t)}\,\boldsymbol{\beta}'\,\mathbf{\Phi}(t,s)\,ds,$$

while integrating the continuation term against the arrival density $\eta e^{-\eta(T-t)}$ and averaging over the draw $\hat{o} \sim \mu(\cdot \mid \mathcal{O})$, which is independent of the arrival time, yields the second term of (20); this integral converges absolutely because $\|\boldsymbol{\gamma}_T'(\hat{o})\,\mathbf{\Phi}(t,T)\|_1 \le (\|\boldsymbol{\beta}\|_1/\rho)\,e^{-\rho(T-t)}$, combining the bound on $\mathbf{\Phi}$ from Lemma 1 with $\|\boldsymbol{\gamma}_T(\hat{o})\|_1 \le \|\boldsymbol{\beta}\|_1/\rho$, which follows from the conditional-expectation representation.

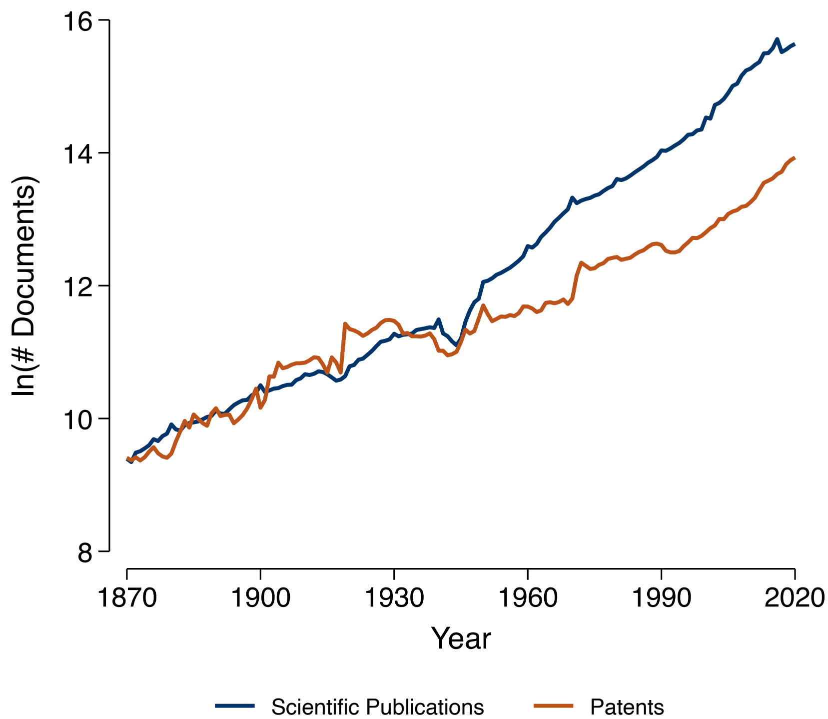
Finally, uniqueness. With general aggregators, the continuation value at an arrival depends on the date, the drawn profile, and the knowledge stocks in place at that moment. Accordingly, consider the space of measurable candidate value functions $\boldsymbol{g}(T,\hat{o},\boldsymbol{z})$, defined on dates, profiles in the support of μ, and states, taking values in row vectors bounded by $\|\boldsymbol{\beta}\|_1/\rho$ in the ℓ_1 norm, and let F map a candidate $\boldsymbol{g}$ into the right-hand side of (20) with $\bar{\boldsymbol{g}}$ in place of $\bar{\boldsymbol{\gamma}}$, evaluated at the corresponding date, profile, and state. Note the bound $\|\boldsymbol{g}'\mathbf{\Phi}(t,T)\|_1 \le \|\boldsymbol{g}\|_1\|\mathbf{\Phi}(t,T)\|_\infty \le \|\boldsymbol{g}\|_1 e^{-\rho(T-t)}$, which holds because the ℓ_1 norm of a row vector times a matrix is at most the ℓ_1 norm of the vector times the largest row sum of the matrix. For two candidates $\boldsymbol{g}^1, \boldsymbol{g}^2$,

$$\big\|(F\boldsymbol{g}^1 - F\boldsymbol{g}^2)_t\big\|_1 \;\le\; \eta\int_t^{\infty} e^{-\eta(T-t)}\,\sup\|\boldsymbol{g}^1 - \boldsymbol{g}^2\|_1\,e^{-\rho(T-t)}\,dT \;=\; \frac{\eta}{\rho+\eta}\,\sup\|\boldsymbol{g}^1 - \boldsymbol{g}^2\|_1,$$

where the supremum is over all dates, profiles, and states. Moreover, F maps this space into itself: the same bounds give $\|(F\boldsymbol{g})_t\|_1 \leq \|\boldsymbol{\beta}\|_1/(\rho+\eta) + \big(\eta/(\rho+\eta)\big)\|\boldsymbol{\beta}\|_1/\rho = \|\boldsymbol{\beta}\|_1/\rho$. Thus F is a contraction with modulus $\eta/(\rho+\eta) < 1$, and the solution of (20) in this space is unique by the contraction mapping theorem; since the same computation makes F a self-map of any larger ball, no bounded solution exists outside this space either. The conditional-expectation representation constructed above is bounded by $\|\boldsymbol{\beta}\|_1/\rho$ and satisfies (20), so it is that unique solution. $\square$

A.3. Additional Results

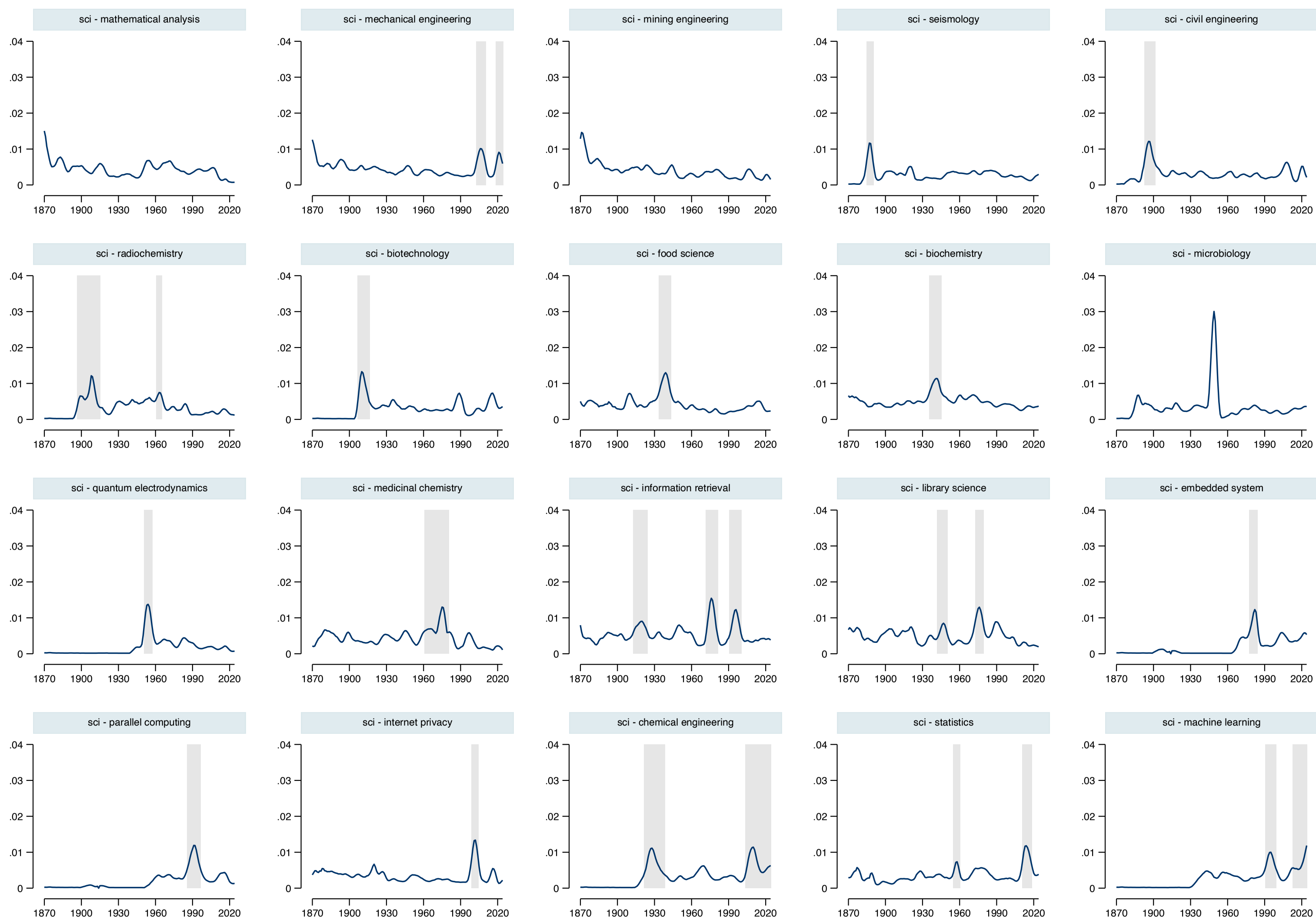
Figure A.2. Number of Patents and Publications Over Time



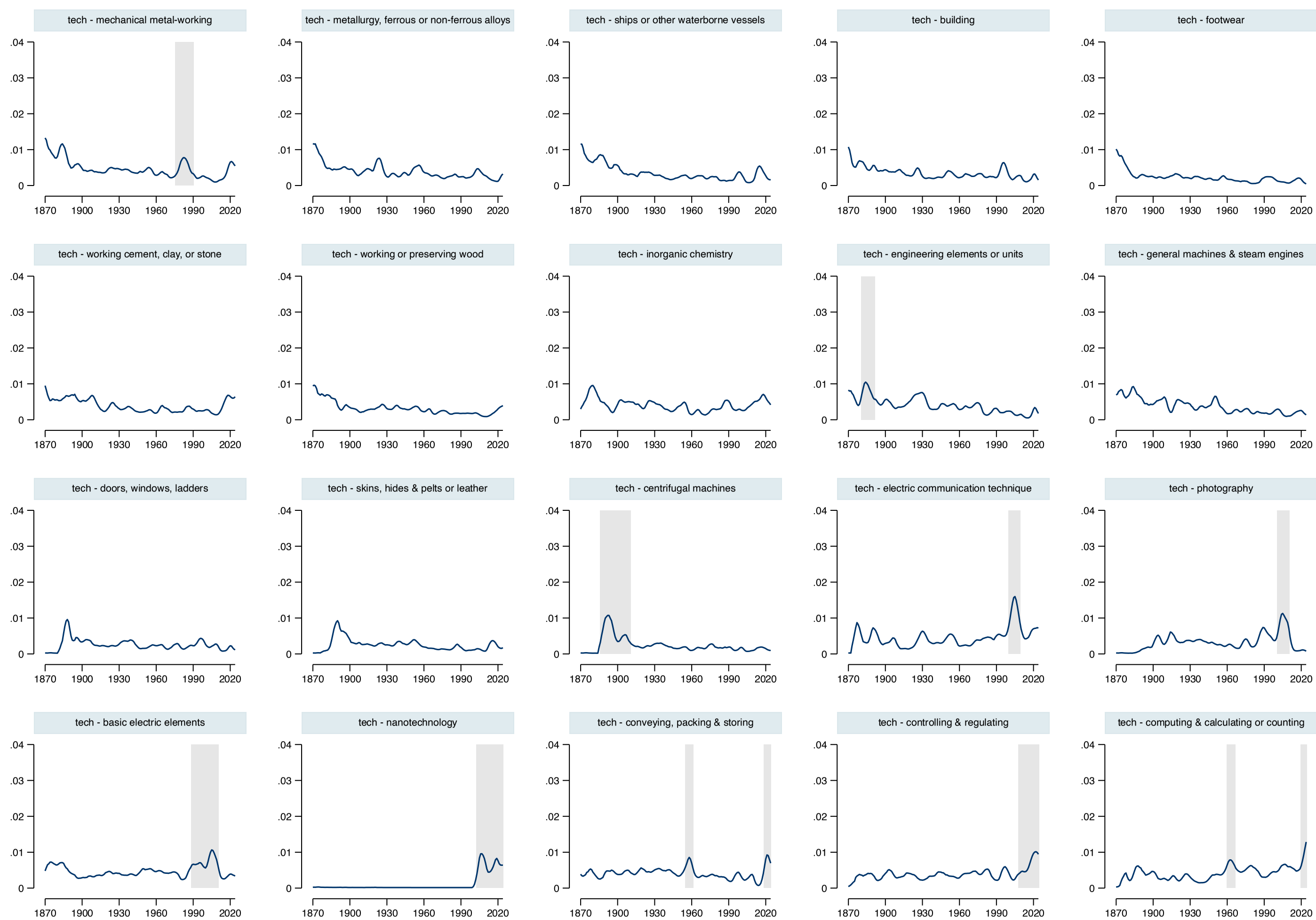
Notes. This figure shows the number of patents (solid line) and articles (dashed line) over time.

Figure A.3. Katz Centrality of Science and Technology Fields Over Time

Panel (a): Science Fields



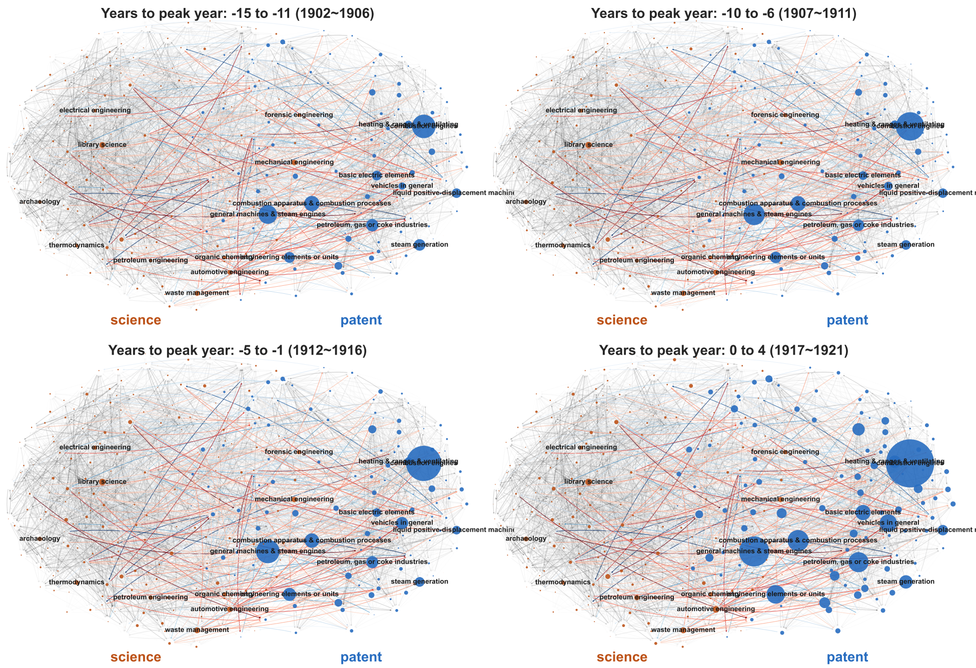
Panel (b): Technology Fields



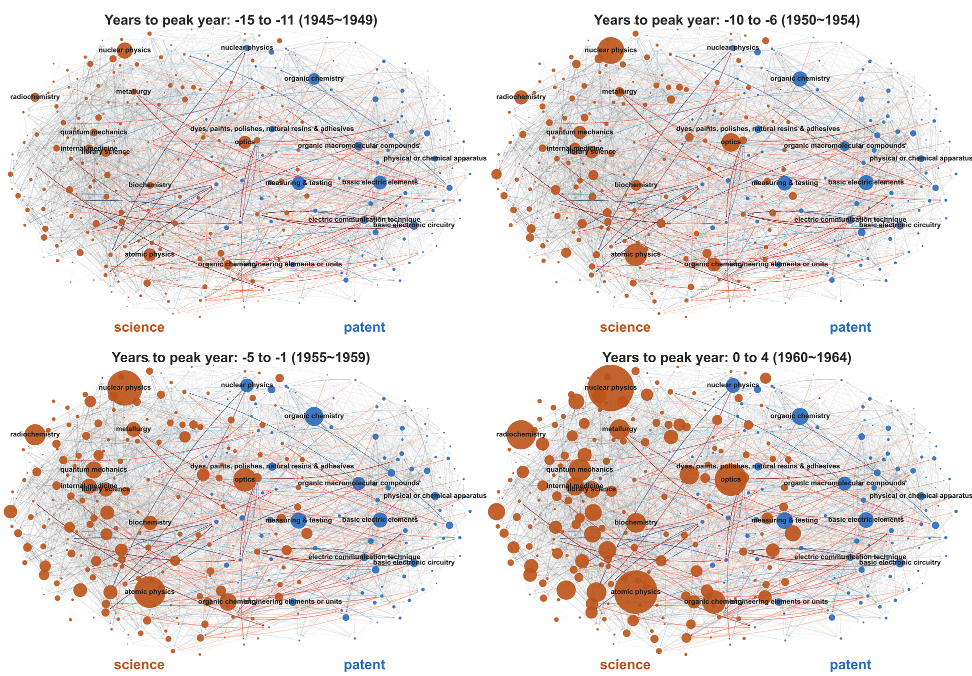
Notes. This figure shows the Katz centrality of science and technology fields over time. We show the top 20 fields sorted by their maximum Katz centrality across all years. The gray area represents the major technological innovation waves.

Figure A.4. Example of the Diffusion of Important Keywords

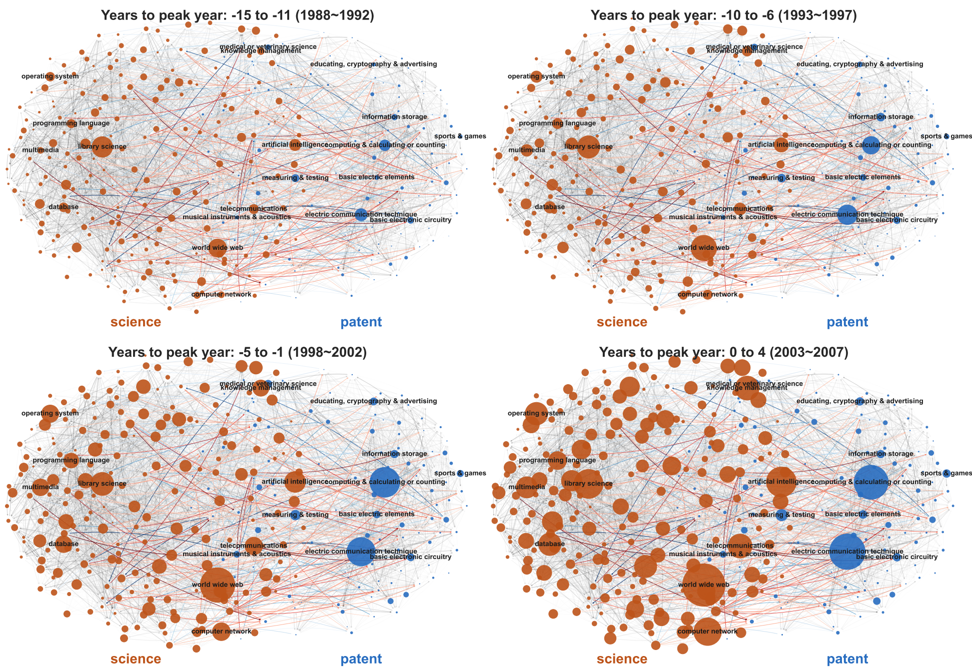
Panel (a): Tech - Combustion Engines (1917)



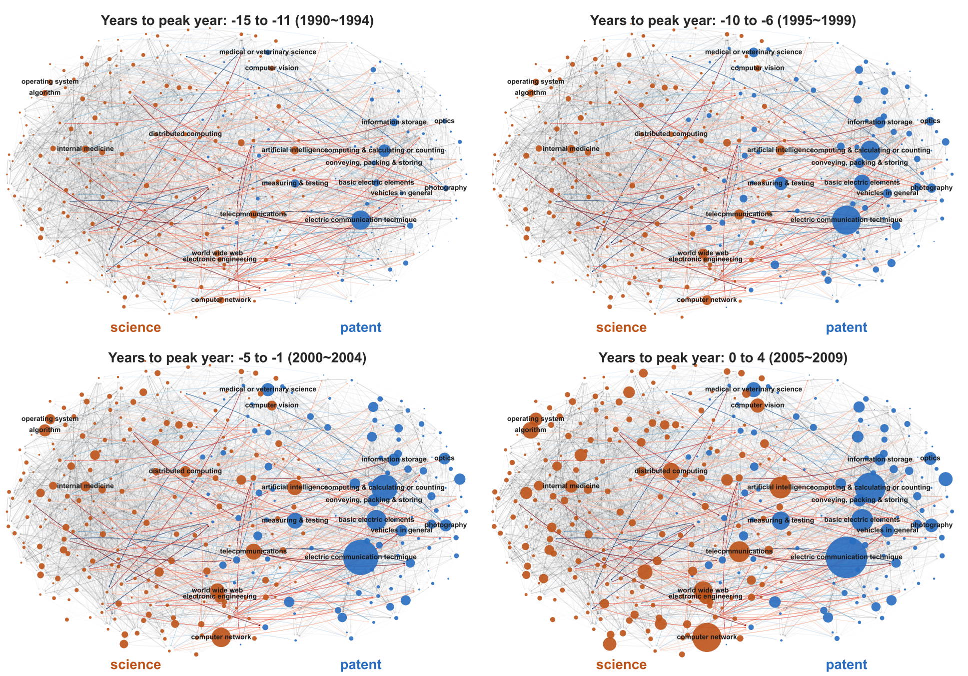
Panel (b): Sci - Nuclear Physics (1960)



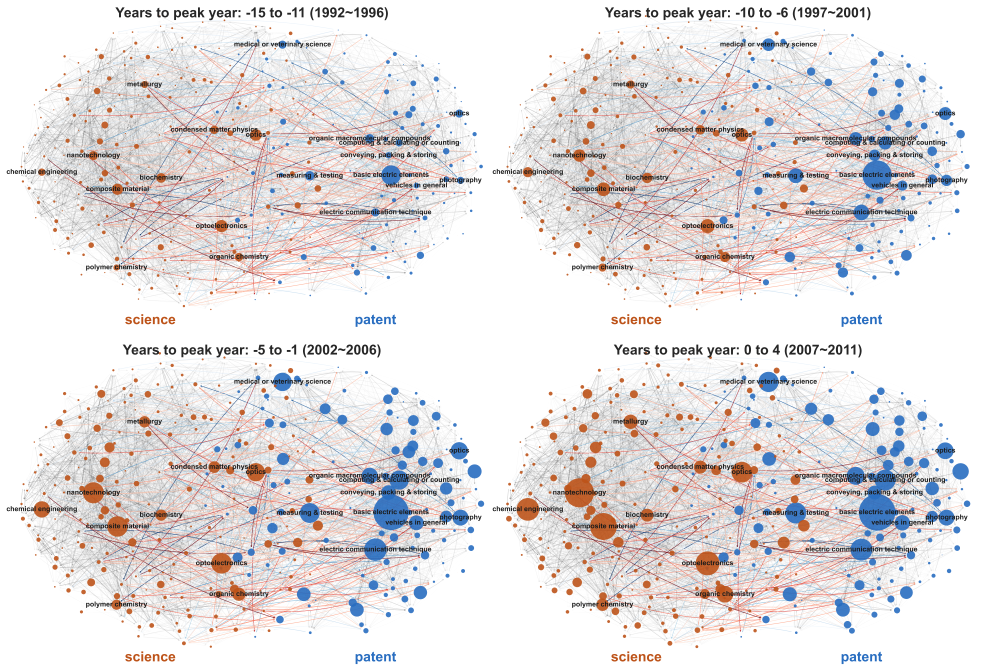
Panel (c): Sci - World Wide Web (2003)



Panel (d): Tech - Electric Communication Technique (2005)



Panel (e): Tech - Nanotechnology (2007)



Panel (f): Sci - Data Science (2020)

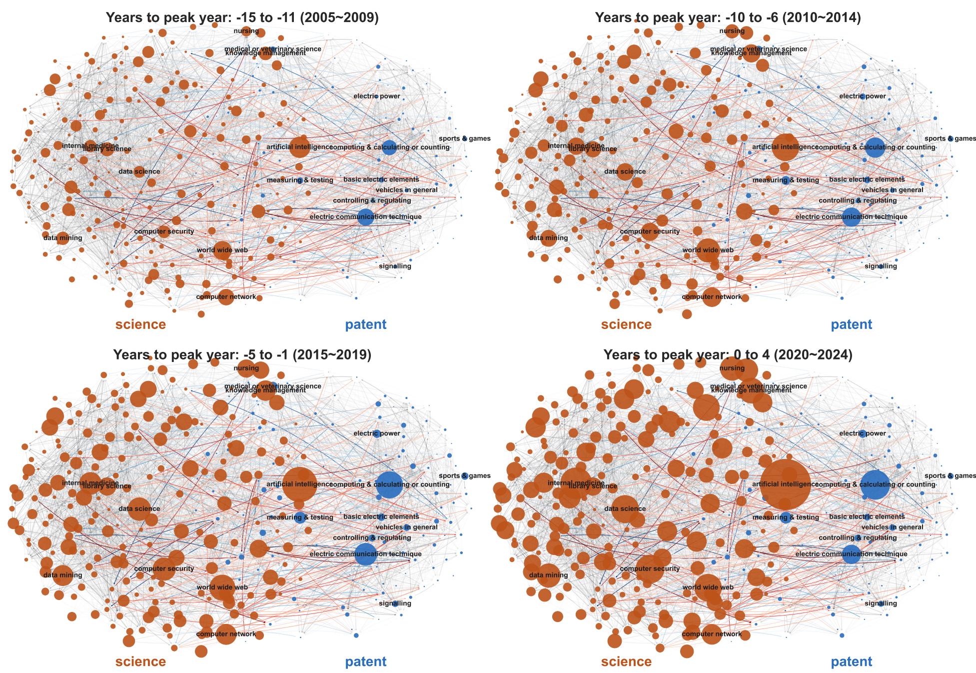
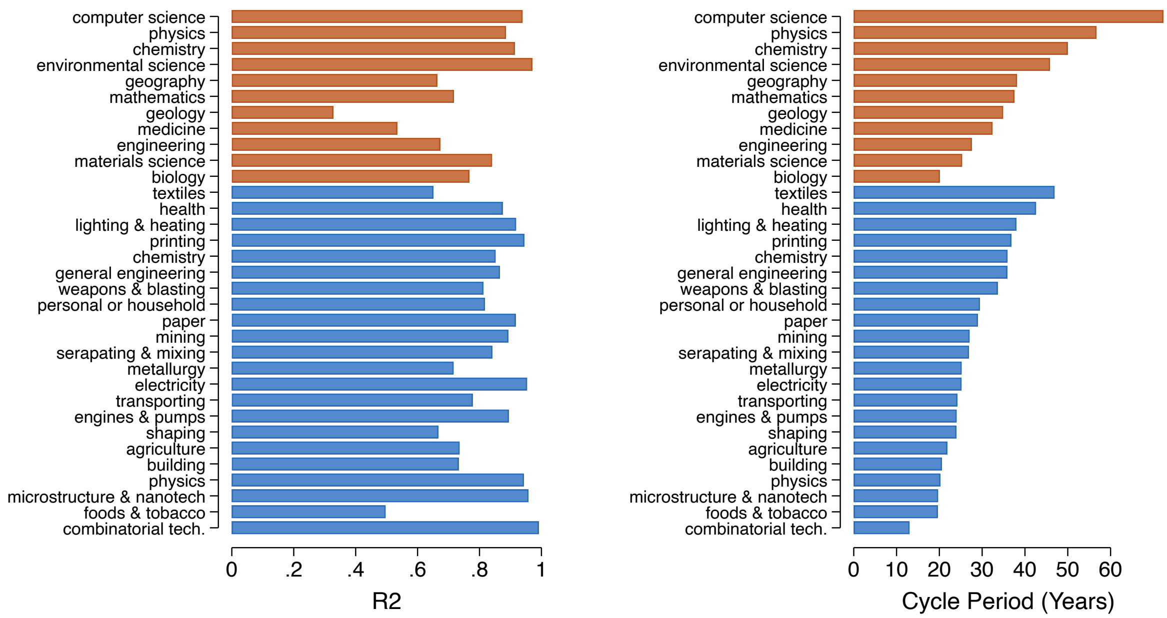


Figure A.5. Model Performance and Estimated Periods of Long-Run Dynamics

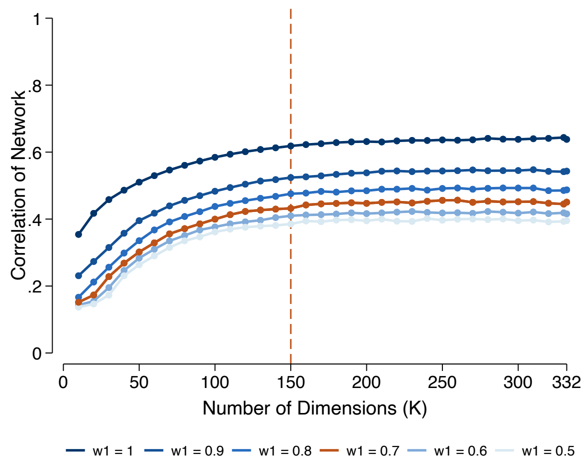
Panel (a): R^2 Panel (b): Estimated Periods



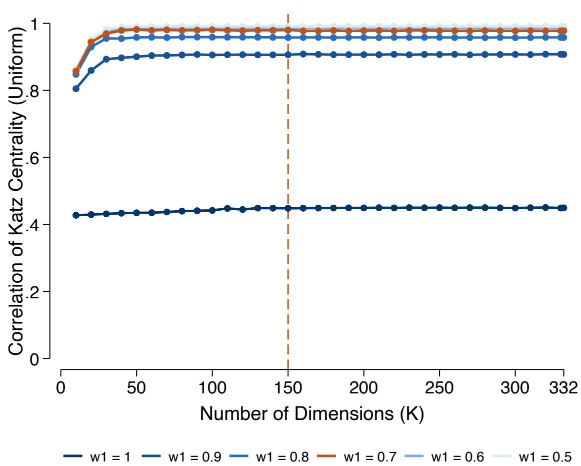
Notes. This figure shows the model performance and estimated periods of long-run dynamics in the unobserved-components time-series model, estimated separately for each aggregate field by maximum likelihood via the Kalman filter (cycle period restricted to $[8, 80]$ years). Panel (a) shows the R^2 of the model, and Panel (b) shows the estimated cycle periods.

Figure A.6. Parameter Selection of Joint Diagonalization

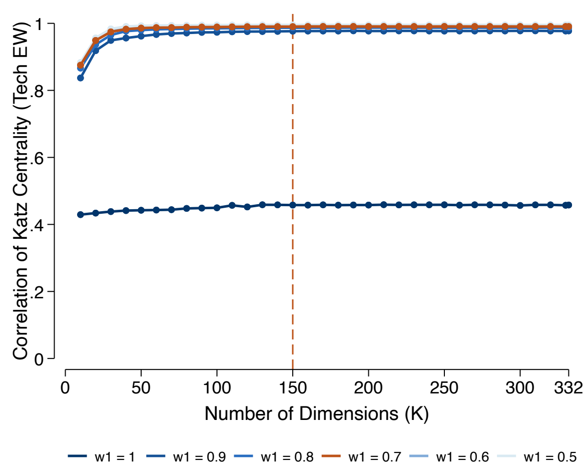
Panel (a): Network



Panel (b): Katz Centrality (Uniform Weight)



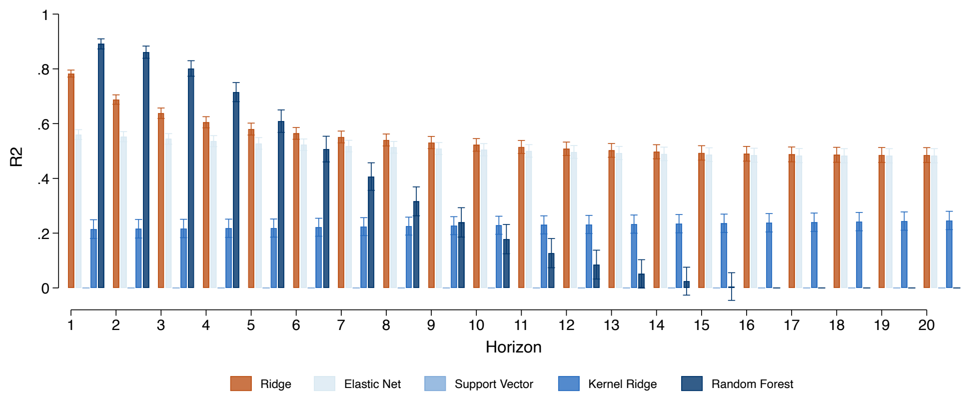
Panel (c): Katz Centrality (Technology Weight)



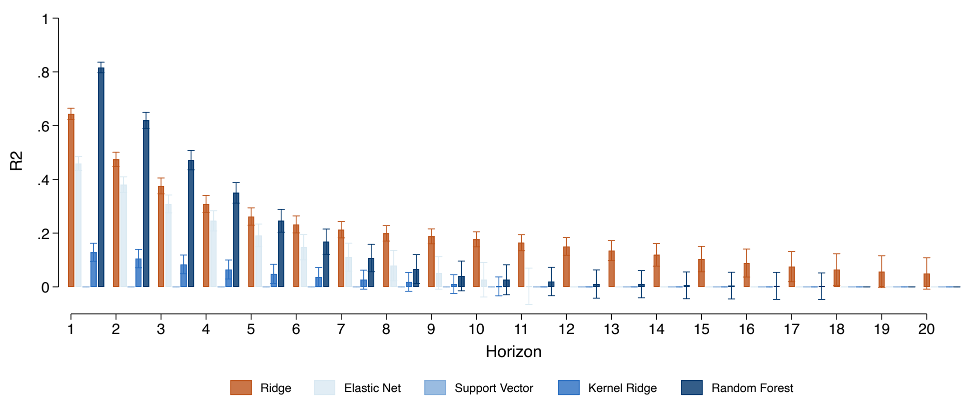
Notes. This figure shows the within-field correlation between the observed target and its rank-K joint-diagonalization reconstruction, plotted against the topic count K with one line per network weight w_Ω ($\delta^\star = 0.95$). Panel (a) is the network $\boldsymbol{\Omega}_t$; Panel (b) is the uniform-Katz centrality; Panel (c) is the technology-weighted Katz centrality. The network correlation plateaus near $K = 150$; the preferred specification is $K = 150$, $w_\Omega = 0.90$.

Figure A.7. Predictability of Eigenvalues: Model Selection

Panel (a): In-Sample R^2



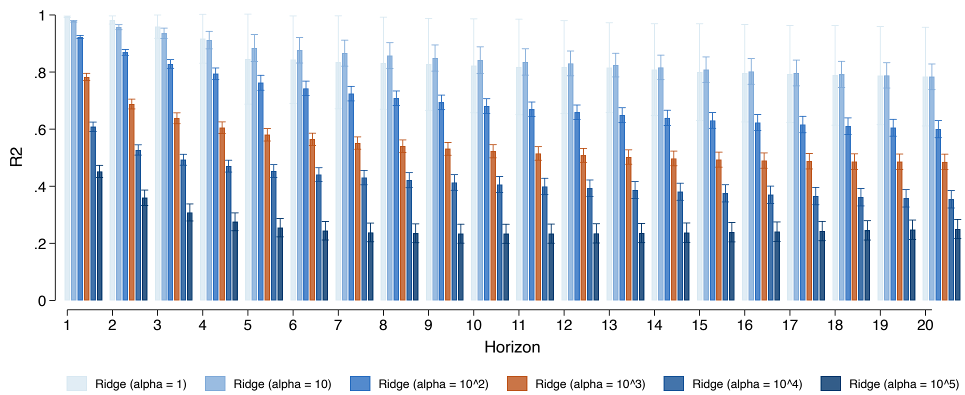
Panel (b): Out-of-Sample R^2



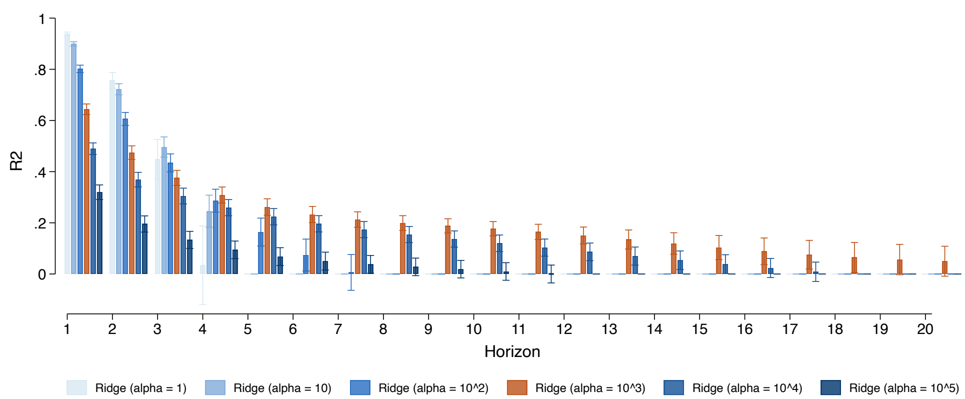
Notes. This figure shows the in-sample and out-of-sample cross-eigenvalue R^2 by horizon across the estimator zoo (Ridge with a penalty of 10^3, Elastic Net, sigmoid-kernel SVR, sigmoid-kernel Ridge, Random Forest).

Figure A.8. Predictability of Eigenvalues: Penalty Selection of Ridge Model

Panel (a): In-Sample R^2



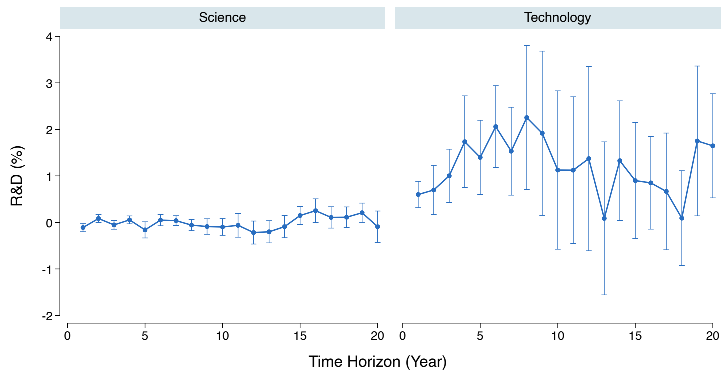
Panel (b): Out-of-Sample R^2



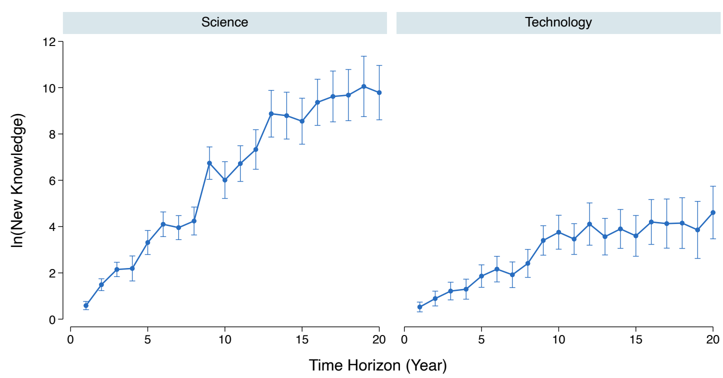
Notes. This figure shows the in-sample and out-of-sample cross-eigenvalue R^2 by horizon along the Ridge penalty ladder: $10^0, \ldots, 10^5$.

Figure A.9. Does Economic Value Predict Real Outcomes? Local Projections

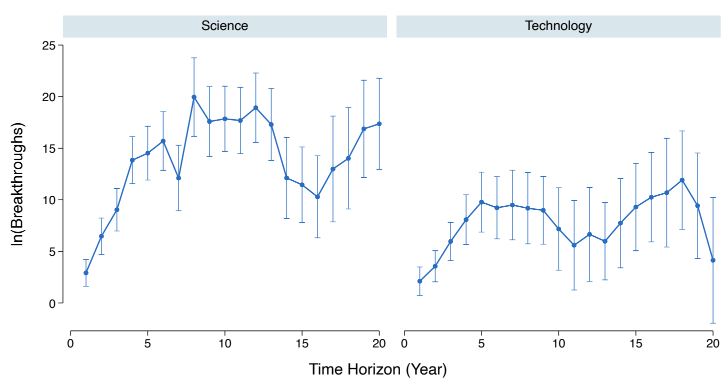
Panel (a): Future R&D Reallocation



Panel (b): Future Knowledge Growth

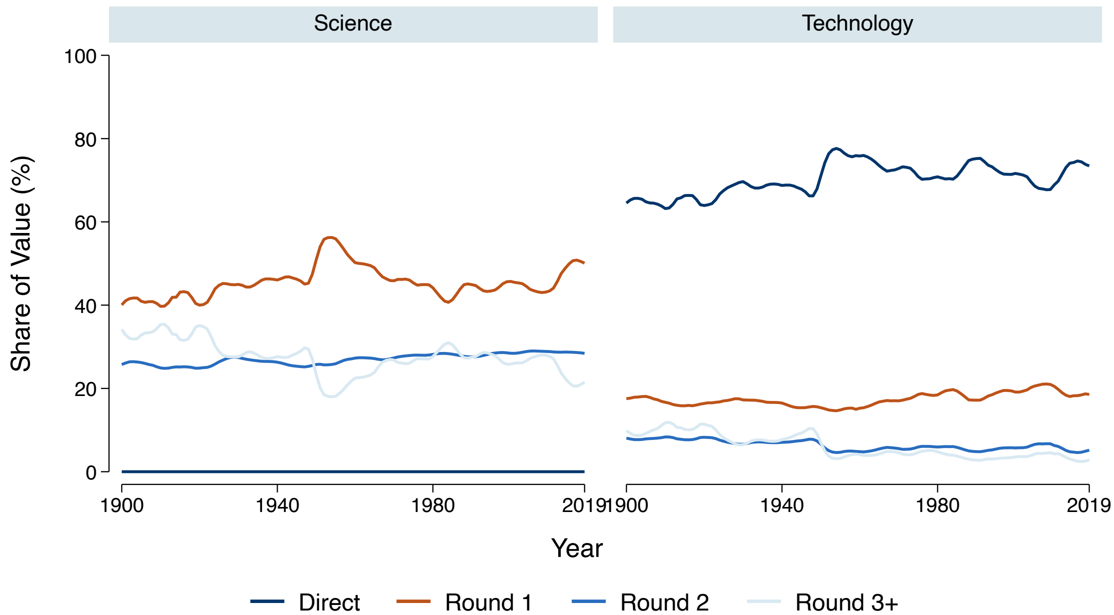


Panel (c): Future Breakthrough Innovations



Notes. Local-projection coefficients on the field's economic-value share (%) at horizons $h = 1, \ldots, 20$ years, estimated separately for science (left) and technology (right) fields. The outcome is the NSF-calibrated R&D share (%) in Panel (a), $\ln(1 + \text{new documents})$ in Panel (b), and $\ln(1 + \text{breakthrough documents})$ in Panel (c), each h years ahead. Regressions control for the contemporaneous outcome and log knowledge stock (with two lags of each) and include field and year fixed effects; standard errors clustered by field.

Figure A.10. Direct and Indirect Rounds of Economic Value Over Time



Notes. Share of each type's forward-looking economic value γ realized in the direct round (Round $0 = \beta$) versus the indirect rounds (one or more spillover steps), by year.

Table A.1. Correlation Between the Text-Based and Citation-Based Networks

	Scientific Field		Technology Class	
Year	Pearson	Spearman	Pearson	Spearman
1950	0.21	0.19	0.22	0.23
1960	0.36	0.21	0.26	0.24
1970	0.50	0.25	0.17	0.30
1980	0.55	0.26	0.42	0.25
1990	0.56	0.29	0.19	0.24
2000	0.58	0.31	0.66	0.39
2010	0.64	0.34	0.39	0.28
2020	0.58	0.29	0.54	0.32

Notes. This table shows the pairwise correlations between the elements from the text-based and citation-based innovation networks. In Columns (1) and (2), we use the science-to-science network, and in Columns (3) and (4), we use the patent-to-patent network. We show the Pearson correlation coefficients and Spearman's rank correlations.

Table A.2. List of Technological Innovation Episodes

Episode Label	IPC/OpenAlex Field Name	Emergence Year	Peak Year	Decline Year	Max Katz Centrality	Keywords
Machine elements and fittings	tech - engineering elements or units	1881	1884	1891	0.0105	clevis, tube, plow, bolt, swivel
Earthquakes and volcanic eruptions	sci - seismology	1885	1887	1890	0.0117	earthquake, volcano, eruption, seismology, fault
Centrifugal cream and milk separators	tech - centrifugal machines	1886	1892	1910	0.0108	centrifuge, separator, milk, cream, dust collector
Railways, bridges, and public works	sci - civil engineering	1893	1896	1901	0.0121	railway, bridge, brick, excavation, tunnel
Synthetic dyes and organic compounds	tech - organic chemistry	1899	1901	1905	0.0089	dye, aromatic amine, ester, fatty acid, formaldehyde
Tuberculosis, plague, and bacteriology	sci - pathology	1900	1903	1915	0.0084	tuberculosis, plague, bacteriology, cholera, tuberculin
Epidemic infectious diseases and vaccines	sci - virology	1899	1904	1906	0.0088	plague, cholera, vaccine, diphtheria, influenza
Radioactivity: radium, thorium, uranium	sci - radiochemistry	1897	1908	1915	0.0121	radium, thorium, radioactivity, uranium, radiation
Crop and livestock breeding	sci - biotechnology	1907	1910	1916	0.0133	breeding, cattle, potato, tobacco, seed
Internal combustion engines and carburetors	tech - combustion engines	1914	1917	1927	0.0087	carburetor, gas engine, explosion engine, internal combustion engine, starter
Automobile wheels, brakes, and tires	tech - vehicles in general	1914	1919	1921	0.0087	automobile, wheel, brake, tire, shock absorber
Scientific journal notices and book reviews	sci - information retrieval	1913	1920	1924	0.0090	book received, patent, book review, bibliography, planet
Warfare, fraud, and secrecy	sci - computer security	1917	1920	1923	0.0087	soldier, enemy, warfare, fraud, telegram
Tractors, bicycles, and horse-drawn wagons	tech - land vehicles for travelling	1918	1921	1924	0.0089	tractor, bicycle, wagon, motorcycle, brake
Medical practice, midwifery, and public health	sci - data science	1917	1921	1924	0.0086	plague, midwives, medical practice, medical service, eugenics
Gauges and measuring instruments	tech - measuring & testing	1923	1927	1929	0.0088	gauge, telemeter, caliper, water meter, tachometer
Registers, catalogs, and directories	sci - database	1920	1927	1930	0.0082	register, warehouse, catalog, directory, invoice
Colloid chemistry and adsorption	sci - chemical engineering	1922	1928	1938	0.0111	colloid, adsorption, gelatin, cellulose, emulsion
Atomic structure and chemical reactions	sci - computational chemistry	1926	1929	1934	0.0109	reaction, atomic structure, dissociation, isomer, valence
Photochemistry and luminescence	sci - photochemistry	1926	1929	1932	0.0103	photochemistry, luminescence, absorption spectra, phosphorescence, fluorescence
Cell physiology and membrane permeability	sci - biophysics	1924	1930	1943	0.0082	cell, protein, hydrogen ion, bioluminescence, permeability
Analytical chemistry and toxicology	sci - environmental chemistry	1929	1932	1936	0.0079	toxicology, analytical chemistry, carbon monoxide, poison, mercury
Steam turbines, boilers, and combustion	sci - nuclear engineering	1930	1934	1937	0.0076	steam, boiler, turbine, combustion, furnace
Reproductive hormones and gynecology	sci - gynecology	1931	1936	1940	0.0080	gynecology, hormone, pregnancy, pituitary, ovary
Vitamins and nutrition	sci - food science	1934	1939	1943	0.0130	vitamin, milk, diet, carotene, nutrition
Hormones, insulin, and the thyroid	sci - endocrinology	1935	1939	1943	0.0093	hormone, diabetes, insulin, pituitary, thyroid
Vitamins and metabolic biochemistry	sci - biochemistry	1936	1941	1945	0.0114	vitamin, metabolism, ascorbic acid, carbohydrate, glycogen
Public health and epidemiology	sci - data science	1938	1944	1949	0.0077	epidemiology, eugenics, venereal disease, tuberculosis, maternal mortality
Infectious disease pathology and epidemiology	sci - pathology	1940	1945	1948	0.0076	epidemiology, tuberculosis, dysentery, bacteriology, sarcoma
Military aviation and aircraft	sci - aeronautics	1943	1946	1949	0.0097	airship, aircraft, aeronautics, aeroplane, aviation
Bibliographies and journal announcements	sci - library science	1942	1947	1950	0.0085	bibliography, masthead, announcement, yearbook, index
Prosthetics, grafts, and bioelectric instrumentation	sci - biomedical engineering	1950	1952	1960	0.0094	tissue, electrode, prosthesis, tomography, electrocardiograph
Mesons, neutrinos, and particle decay	sci - particle physics	1951	1953	1956	0.0083	meson, neutrino, nucleon, decay, parity
Servomechanisms and automatic control	sci - control engineering	1950	1953	1961	0.0080	servomechanism, automatic control, control system, feedback, frequency response
Quantum electrodynamics and field theory	sci - quantum electrodynamics	1951	1954	1957	0.0138	meson, electrodynamics, quantum, electromagnetic field, field theory
Radiology and radiotherapy	sci - medical physics	1944	1954	1963	0.0081	radiology, radiography, radium, radiotherapy, radiation
Biological assay and microbiology methods	sci - biological system	1952	1954	1961	0.0080	biological material, cell, bacteria, colorimetric, titration
Microwave spectroscopy and cosmic rays	sci - computational physics	1952	1955	1968	0.0088	scattering, microwave, spectrum, ionosphere, cosmic ray
Conveyors and packaging machinery	tech - conveying, packing & storing	1955	1958	1961	0.0085	conveyor, package, carton, pallet, conveyor belt
Vital statistics and statistical methods	sci - statistics	1955	1958	1960	0.0073	statistics, vital statistics, probit, regression, binomial
Transistor radio and pulse circuits	tech - basic electronic circuitry	1950	1959	1963	0.0078	pulse, radio receiver, transistor, amplifier, oscillator
Nuclear reactions, mesons, and fission	sci - nuclear physics	1951	1960	1967	0.0078	meson, nucleon, cyclotron, neutron, fission
Nuclear reactors and atomic power	sci - nuclear engineering	1958	1961	1965	0.0077	reactor, nuclear reactor, atomic energy, nuclear power, uranium
Digital and analog computing machines	tech - computing & calculating or counting	1960	1963	1966	0.0078	digital, data processing, analog, binary, calculating machine
Radioisotopes and gamma radiation	sci - radiochemistry	1961	1963	1965	0.0075	radium, radiation, gamma, radioisotope, isotope
Digital computer hardware and magnetic tape	sci - computer hardware	1960	1963	1972	0.0073	digital, computer, magnetic tape, transistor, hardware
Magnetic and phonograph sound recording	tech - information storage	1951	1967	1970	0.0076	phonograph, magnetic recording, sound recording, magnetic tape, magnetic memory
Linear and dynamic programming	sci - mathematical optimization	1958	1971	1977	0.0071	linear programming, optimization, optimal control, minimax, dynamic programming
Cell ultrastructure and electron microscopy	sci - biophysics	1961	1972	1984	0.0075	membrane, DNA, electron microscope, ribosome, mitochondria
Organic synthesis and stereochemistry	sci - medicinal chemistry	1961	1975	1980	0.0130	synthesis, substitution, aromatic, stereochemistry, heterocyclic
Chemical literature retrieval and indexing	sci - information retrieval	1972	1976	1981	0.0155	retrieval, information retrieval, author index, book received, current literature
Bibliographies and journal indexes	sci - library science	1973	1976	1979	0.0130	subject index, bibliography, book received, new book, editorial board
Graph theory and combinatorics	sci - combinatorics	1957	1976	1981	0.0080	graph, permutation, combinatorics, polynomial, Fibonacci
Microprocessor systems and VLSI	sci - embedded system	1978	1982	1984	0.0123	microprocessor, hardware, microcomputer, chip, VLSI
Microprocessors and digital signal processing	sci - computer hardware	1978	1982	1986	0.0096	microprocessor, signal processing, microcomputer, chip, LSI
Metal rolling, stamping, and forming	tech - mechanical metal-working	1976	1982	1990	0.0078	die, rolling mill, sheet metal, extrusion, stamping
VLSI and microprocessor architecture	sci - computer architecture	1980	1985	1988	0.0105	microprocessor, VLSI, processor, multiprocessor, microcomputer
Computer graphics and image processing	sci - computer graphics (images)	1982	1986	1990	0.0082	graphics, image, image processing, raster, microcomputer
VLSI and CMOS circuit design	sci - electronic engineering	1983	1987	1991	0.0088	VLSI, CMOS, transistor, MOSFET, switched capacitor
VLSI design and fault-tolerant logic	sci - computer engineering	1984	1987	1994	0.0082	VLSI, algorithm, logic, fault tolerance, integrated circuit

Episode Label	IPC/OpenAlex Field Name	Emergence Year	Peak Year	Decline Year	Max Katz Centrality	Keywords
Parallel computing and multiprocessors	sci - parallel computing	1986	1991	1996	0.0119	supercomputer, multiprocessor, hypercube, parallel computing, systolic array
Scientific computing and numerical simulation	sci - computational science	1983	1991	1994	0.0080	supercomputer, simulation, solver, finite element, parallel computing
Technology transfer and information management	sci - knowledge management	1972	1991	1994	0.0080	technology transfer, expert system, information system, knowledge base, innovation
Gene cloning and molecular genetics	sci - genetics	1985	1992	2005	0.0094	gene, plasmid, clone, chromosome, DNA
Chaos, fractals, and statistical mechanics	sci - statistical physics	1986	1992	1999	0.0082	chaos, fractal, Monte Carlo, statistical mechanics, percolation
Neural networks and expert systems	sci - machine learning	1991	1995	1999	0.0100	neural network, expert system, fuzzy logic, machine learning, artificial intelligence
Polymers and conducting polymers	sci - polymer chemistry	1991	1995	2000	0.0096	polymer, copolymer, methacrylate, polyaniline, conducting polymer
Chemical databases and information retrieval	sci - information retrieval	1991	1996	2000	0.0123	database, retrieval, bibliography, molecular structure, patent
Data mining and knowledge-based systems	sci - data mining	1992	2001	2004	0.0083	fuzzy logic, neural network, database, expert system, data mining
Internet privacy and e-commerce	sci - internet privacy	1999	2002	2004	0.0134	internet, cyberspace, privacy, hacker, electronic commerce
Molecular biology and genomics	sci - computational biology	1984	2002	2006	0.0097	gene, molecular, clone, sequence, genome
World Wide Web and the internet	sci - world wide web	2000	2003	2006	0.0110	internet, world wide web, web site, hypertext, cyberspace
Mobile and wireless telecommunications	tech - electric communication technique	2000	2005	2009	0.0160	mobile, packet, wireless, base station, CDMA
Electrophotography and silver-halide imaging	tech - photography	2001	2005	2010	0.0113	image formation, toner, silver halide, photoreceptor, electrophotography
Semiconductor devices and components	tech - basic electric elements	1989	2005	2010	0.0107	semiconductor, silicon, wafer, transistor, capacitor
Optics: LCD displays and lenses	tech - optics	2002	2005	2010	0.0089	lens, liquid crystal display, pixel, optical fiber, laser
Mechanical simulation and rapid prototyping	sci - mechanical engineering	2003	2006	2010	0.0101	simulation, optimization, prototype, compressor, CFD
Inkjet and thermal printing	tech - printing	1988	2006	2012	0.0091	ink, printer, inkjet, thermal transfer, printhead
Carbon nanotubes and nanostructures	tech - nanotechnology	2003	2007	2024	0.0096	nanotube, carbon nanotube, nanostructure, nanoparticle, nanowire
Organic electroluminescent (OLED) displays	tech - other electric techniques	2004	2008	2011	0.0085	light emitting, electroluminescence, organic electroluminescence, OLED, LED
Nanomaterials and fuel cells	sci - chemical engineering	2004	2010	2024	0.0114	nanoparticle, carbon nanotube, mesoporous, fuel cell, self-assembly
Nanotubes, nanoparticles, and nanostructures	sci - nanotechnology	2007	2011	2024	0.0082	nanotube, nanoparticle, carbon nanotube, nanostructure, nanowire
LED lighting and luminaires	tech - lighting	2007	2011	2015	0.0077	LED, luminaire, light emitting diode, backlight, bulb
Regression, forecasting, and Bayesian statistics	sci - statistics	2011	2014	2018	0.0118	regression, uncertainty, forecast, Bayesian, Monte Carlo
Obstetrics, pregnancy, and gynecology	sci - gynecology	2011	2014	2017	0.0082	pregnancy, gynecology, obstetrics, endometrial, contraception
Organic synthesis of herbicides and pharmaceuticals	tech - organic chemistry	2010	2017	2024	0.0079	synthesis, herbicide, pharmaceutical composition, ionic liquid, inhibitor
Big data and cloud analytics	sci - data science	2017	2020	2024	0.0084	big data, cloud computing, data mining, social network, data analytics
Automated conveying and palletizing	tech - conveying, packing & storing	2019	2021	2024	0.0092	rack, conveying device, pallet, conveyor belt, robot
CAD, turbines, and rapid prototyping	sci - mechanical engineering	2019	2021	2024	0.0091	CAD, wind turbine, prototype, gas turbine, finite element
Big data mining and machine learning	sci - data mining	2018	2021	2024	0.0077	big data, data mining, neural network, machine learning, association rule
Automated control: robots and UAVs	tech - controlling & regulating	2008	2022	2024	0.0101	control method, robot, unmanned aerial vehicle, process control, PLC
Cloud computing and neural networks	tech - computing & calculating or counting	2020	2024	2024	0.0129	cloud, neural network, machine learning, RFID, deep learning
Deep learning and neural networks	sci - machine learning	2013	2024	2024	0.0119	machine learning, neural network, deep learning, support vector machine, convolutional neural network
Quantum computing and deep-learning hardware	sci - computer engineering	2022	2024	2024	0.0093	quantum computing, neural network, deep learning, GPU, VLSI
COVID, HIV, and viral pandemics	sci - virology	2022	2024	2024	0.0088	COVID, HIV, hepatitis, influenza, coronavirus

Notes. This table lists each innovation episode with its emergence, peak, and decline years, its maximum Katz centrality, and its characteristic keywords. The Episode Label is a concise descriptive label assigned to each episode based on the field's characteristic keywords during the episode and its peak year. The IPC/OpenAlex Field Name is the major field name from the underlying OpenAlex and IPC taxonomies.

Table A.3. Top 10 Science and Technology Fields by Katz Centrality Over Time

Panel (a): Top 10 Fields in 1880

Scientific Fields			Technological Classes		
Rank	Katz Centrality	Field Name	Rank	Katz Centrality	Field Name
2	0.0089	polymer science	1	0.0096	inorganic chemistry
3	0.0086	environmental chemistry	6	0.0081	mechanical metal-working
4	0.0085	pure mathematics	8	0.0076	natural or man-made threads or fibres
5	0.0081	atmospheric sciences	13	0.0074	explosives & matches
7	0.0080	theoretical physics	18	0.0069	heat exchange in general
9	0.0076	archaeology	19	0.0068	haberdashery & jewellery
10	0.0075	telecommunications	20	0.0068	building
11	0.0075	operations research	22	0.0067	general machines & steam engines
12	0.0075	nursing	24	0.0066	medical or veterinary science
14	0.0073	medical emergency	25	0.0066	working or preserving wood

Panel (b): Top 10 Fields in 1900

Scientific Fields			Technological Classes		
Rank	Katz Centrality	Field Name	Rank	Katz Centrality	Field Name
2	0.0073	applied mathematics	1	0.0082	organic chemistry
3	0.0073	civil engineering	7	0.0068	combustion apparatus & combustion processes
4	0.0072	regional science	8	0.0067	furnaces & kilns, ovens or retorts
5	0.0071	medical physics	9	0.0066	steam generation
6	0.0069	nuclear physics	10	0.0065	heating & ranges & ventilating
12	0.0065	radiochemistry	11	0.0065	liquid positive-displacement machines
14	0.0063	inorganic chemistry	13	0.0064	foods or foodstuffs
15	0.0060	library science	18	0.0058	metallurgy of iron
16	0.0059	gynecology	20	0.0057	land vehicles for travelling
17	0.0059	automotive engineering	21	0.0057	lighting

Panel (c): Top 10 Fields in 1920

Scientific Fields			Technological Classes		
Rank	Katz Centrality	Field Name	Rank	Katz Centrality	Field Name
1	0.0090	information retrieval	3	0.0085	land vehicles for travelling
2	0.0087	computer security	4	0.0080	vehicles in general
5	0.0078	data science	14	0.0060	petroleum, gas or coke industries
6	0.0074	library science	20	0.0057	combustion apparatus & combustion processes
7	0.0068	environmental health	23	0.0056	general machines & steam engines
8	0.0067	internet privacy	25	0.0055	combustion engines
9	0.0064	engineering ethics	26	0.0055	working of plastics
10	0.0063	aeronautics	27	0.0054	conveying, packing & storing
11	0.0061	process engineering	29	0.0053	machine tools
12	0.0060	agricultural engineering	32	0.0052	engineering elements or units

Panel (d): Top 10 Fields in 1940

Scientific Fields			Technological Classes		
Rank	Katz Centrality	Field Name	Rank	Katz Centrality	Field Name
1	0.0125	food science	7	0.0060	aircraft, aviation & cosmonautics
2	0.0111	biochemistry	13	0.0053	organic chemistry
3	0.0087	endocrinology	17	0.0050	medical or veterinary science
4	0.0067	internal medicine	18	0.0050	conveying, packing & storing
5	0.0062	optics	21	0.0050	heat exchange in general
6	0.0062	physiology	24	0.0049	organic macromolecular compounds
8	0.0058	gynecology	27	0.0049	working of plastics
9	0.0057	chemical physics	32	0.0047	petroleum, gas or coke industries
10	0.0056	atomic physics	40	0.0044	metallurgy of iron
11	0.0056	electronic engineering	41	0.0044	musical instruments & acoustics

Panel (e): Top 10 Fields in 1960

Scientific Fields			Technological Classes		
Rank	Katz Centrality	Field Name	Rank	Katz Centrality	Field Name
1	0.0078	nuclear physics	3	0.0076	basic electronic circuitry
2	0.0077	nuclear engineering	5	0.0070	conveying, packing & storing
4	0.0076	computational physics	10	0.0065	computing & calculating or counting
6	0.0067	biochemistry	29	0.0051	nuclear physics
7	0.0067	composite material	32	0.0050	working of plastics
8	0.0066	metallurgy	33	0.0049	crystal growth
9	0.0065	chemical physics	37	0.0047	information storage
11	0.0063	molecular physics	40	0.0046	measuring & testing
12	0.0062	systems engineering	42	0.0046	basic electric elements
13	0.0060	medicinal chemistry	44	0.0045	aircraft, aviation & cosmonautics

Panel (f): Top 10 Fields in 1980

Scientific Fields			Technological Classes		
Rank	Katz Centrality	Field Name	Rank	Katz Centrality	Field Name
1	0.0095	embedded system	4	0.0067	mechanical metal-working
2	0.0073	computer hardware	8	0.0064	foods or foodstuffs
3	0.0072	information retrieval	9	0.0062	basic electronic circuitry
5	0.0067	library science	10	0.0062	computing & calculating or counting
6	0.0065	regional science	28	0.0050	electric power
7	0.0065	optics	36	0.0048	controlling & regulating
11	0.0062	geotechnical engineering	39	0.0046	treatment of water
12	0.0060	transport engineering	45	0.0044	electric communication technique
13	0.0059	medicinal chemistry	54	0.0043	fertilisers
14	0.0057	molecular biology	59	0.0042	earth or rock drilling & mining

Panel (g): Top 10 Fields in 2000

Scientific Fields			Technological Classes		
Rank	Katz Centrality	Field Name	Rank	Katz Centrality	Field Name
1	0.0111	internet privacy	4	0.0078	electric communication technique
2	0.0083	computational biology	8	0.0061	hand tools
3	0.0081	data mining	10	0.0060	nuclear physics
5	0.0063	molecular biology	13	0.0056	basic electric elements
6	0.0063	computer network	17	0.0052	biochemistry, beer, microbiology & enzymology
7	0.0062	cell biology	23	0.0050	medical or veterinary science
9	0.0061	human-computer interaction	25	0.0050	coating metallic material
11	0.0058	world wide web	28	0.0047	vehicles in general
12	0.0057	information retrieval	32	0.0046	specific information & communication tech.
14	0.0056	gerontology	33	0.0045	computing & calculating or counting

Panel (h): Top 10 Fields in 2020

Scientific Fields			Technological Classes		
Rank	Katz Centrality	Field Name	Rank	Katz Centrality	Field Name
2	0.0084	data science	1	0.0097	controlling & regulating
3	0.0081	mechanical engineering	4	0.0081	conveying, packing & storing
6	0.0076	data mining	5	0.0077	nanotechnology
7	0.0072	nanotechnology	8	0.0070	electric communication technique
11	0.0067	process engineering	9	0.0070	machine tools
14	0.0064	machine learning	10	0.0068	hand tools
20	0.0058	meteorology	12	0.0065	mechanical metal-working
23	0.0056	petroleum engineering	13	0.0065	working cement, clay, or stone
28	0.0053	human-computer interaction	15	0.0063	computing & calculating or counting
29	0.0052	civil engineering	16	0.0062	inorganic chemistry

Notes. This table shows the top 10 science and patent fields by centrality in selected years.