The Economic Value of Science*

Xugan Chen† *Yale*

Ernest Liu‡ *Princeton and NBER*

Song Ma§ *Yale and NBER*

Preliminary & Incomplete — Please Do Not Circulate

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 tracing these flows, using the text of 463 million scientific articles and 81 million patents to measure knowledge diffusion among 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. We quantify how the value of science of different fields has evolved over time, and how they materialize through different industries.

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

* *This version:* March 25, 2026. All errors are our own. We thank seminar participants at Kellogg, NYU, Princeton, Yale, and the financial support from the Alfred P. Sloan Foundation. Ernest Liu also acknowledges the 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

How valuable is scientific research to economic growth? Despite widespread recognition that science drives technological progress, quantifying its economic value has remained elusive. Two interrelated features of scientific knowledge make this difficult. First, scientific fields—mathematics, chemistry, molecular biology, and the like—do not directly produce consumption goods. Their contribution to the economy flows entirely through knowledge spillovers: scientific discoveries feed into novel inventions and downstream technologies, often indirectly and with long lags. Second, these spillover effects propagate through a complex web of cross-field linkages, making the full chain of indirect contributions difficult to observe and measure empirically. An ideal measure of the economic importance of science must therefore account for the entire structure of knowledge flows—from science to technology and onward through the innovation network. Such a measure is needed to assess science's contribution to economic growth and to guide the allocation of R&D resources across scientific fields.

This paper fills this gap, both theoretically and empirically, by proposing a new framework to quantify the economic importance 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 characterizes the social value of R&D across sectors 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 per-

sistent 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. This VAR structure yields an empirical advantage, as 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. Because the diffusion lag is a priori unclear, we estimate the regression separately for lags $\tau = 1, \ldots, 10$ years within a rolling five-year window and average the resulting coefficients, capturing cumulative spillovers while remaining agnostic about precise timing.

We implement this methodology on a comprehensive collection of data from multiple sources: more than 463 million academic articles from 150 thousand journals and 45 thousand universities and research institutions dating back to the 16th century from OpenAlex; 81 million patents from over 40 major patent authorities covering 1782–2025 from the European Patent Office's DOCDB and INPADOC databases; R&D expenditures from 0.11 million global firms in 160 countries from Worldscope; 6.3 million R&D scientific grants from 651 funding agencies across 43 countries in 1952–2021 from Dimensions; and sectoral production data from the World Input-Output Database (WIOD, Timmer et al., 2015), which covers 43 economies over 2000–2014 representing more than 85% of world GDP. The critical 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-patent 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. The innovation network visualization reveals strong cross-domain asymmetry: arrows representing science-to-patent flows are far more numerous and thicker than those representing patent-to-science flows. Scientific fields such as the world wide web, artificial intelligence, and biochemistry serve as dominant knowledge sources for a broad range of technology sectors. This asymmetry is further illustrated in the heatmap of the 332×332 network sorted by Katz centrality: the science-to-patent quadrant shows dense activity, while the patent-to-science quadrant is comparatively sparse. In 2010, the top ten fields by Katz centrality are entirely scientific—led by nanotechnology, data science, chemical engineering, and cell biology—while the highest-ranked technology classes appear well below the top ten. Second, centrality dynamics differ markedly between the two domains. We track how the Katz centrality rankings of the top 20 fields within each domain evolve over subsequent years. Technology fields experience relatively large short-run movements but stabilize over time, likely reflecting product cycles, shifting market demand, or regulatory changes. Science fields, by contrast, exhibit a higher 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. The

long-run dynamics also display pronounced cyclical patterns, with science fields exhibiting longer cycles (approximately 41 years) than technology fields (approximately 29 years).

We further examine how breakthrough innovations reshape the network. We identify breakthrough innovations following Kelly et al. (2021), classifying documents in the top 1% of the ratio of future-to-past similarity as breakthroughs. Using local projection (LP) methods, we find that breakthroughs are associated with a subsequent increase in Katz centrality. In both domains, centrality rises sharply in the 5–10 years following a breakthrough, with the increase more pronounced and persistent in science. Quantitatively, a 10% increase in the count of breakthrough innovations is associated with a 0.0026 increase in centrality at a 5-year horizon, equivalent to moving a field from the median to the top decile of the centrality distribution. These results hold after controlling for the existing knowledge stock and are statistically significant at the 1% level across all specifications.

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 is that scientific fields—mathematics, chemistry, molecular biology, and the like—do not directly produce consumption goods. Their entire contribution to the economy flows through knowledge spillovers: 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 social value of R&D in each sector. A sector's total social 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 ratio of the discount rate to the R&D arrival rate. 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: Katz centrality measures the social value of a sector by aggregating its direct and indirect contributions to downstream production, discounted by the length of each chain. When the network is time-varying, the social value generalizes to a forward-looking analogue in which each round of network effects uses the future network at the appropriate horizon. The marginal social value of investing in a sector is then proportional to the ratio of that sector's total social 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 all sectors produce consumption goods. Because scientific fields receive zero direct weight in consumer preferences, their entire social 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.

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 a small number of stable "knowledge topics" with time-varying importance weights. This reduces the prediction problem from forecasting each of the N^2 network elements at every future date to forecasting only N eigenvalue time series—a reduction from over 100,000 parameters to roughly 300. We further show that only the top 100 eigenmodes are needed to capture the predictable variation in the network, yielding a compact set of time series that can be forecast accurately using knowledge stocks and breakthrough innovations as predictors. The approach captures 98% of the variation in Katz centrality and achieves strong out-of-sample forecasting performance.

Quantitative results. We implement the quantitative framework on our innovation network of 332 fields. A joint diagonalization decomposes the time-varying network into 150 knowledge topics—stable clusters of related fields with time-varying importance weights—capturing 98% of the variation in Katz centrality. We forecast the evolution of these knowledge topics using ridge regression with knowledge stocks and breakthrough innovations as predictors, achieving out-of-sample R^2 of approximately 0.73 at short horizons. Combining the JD decomposition with these forecasts, we compute the forward-looking social value of R&D for each field.

The resulting field-level estimates reveal striking heterogeneity in how the social value of

science evolves over time. Among computer science subfields, machine learning exhibits near-exponential growth in social 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. In mathematics, mathematical optimization and statistics display steady upward trajectories, reflecting their growing importance as inputs to machine learning and engineering, while mathematical economics peaked in the mid-20th century and has since declined. Medical fields show relatively stable social values, and most physics subfields—particle physics, atomic physics, nuclear physics, and astronomy—trend downward over recent decades, consistent with their displacement in centrality rankings by newer fields in computer science and biology.

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 social 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 global innovation network spanning real economy sectors and scientific fields using text data, 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 extends the landscape by incorporating detailed scientific fields and tracing both domestic and international knowledge connections between science and technology in 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 rest of the paper is organized 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 social 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.[1] OpenAlex aggregates bibliographic records from major publishers, indexing services, and institutional repositories into a unified, deduplicated 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, producing a comprehensive and deduplicated database 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

[1] We access OpenAlex via https://openalex.org/ at November 2025.

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 linking each article to earlier literature. We use these citation relationships 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 classification system includes 284 distinct fields grouped under 19 root-level knowledge domains. We focus on the 209 STEM-related fields spanning natural sciences, engineering, mathematics, and computer science. Each article may be tagged with multiple fields; in such cases, we assign equal fractional weights to each assigned category, ensuring that multi-field publications contribute proportionally to each relevant domain.

We determine a publication's geographic origin from the institutional affiliations of its authors. When multiple authors are affiliated with institutions in different countries, we assign equal fractional weights to each country. Geographic assignment is essential for tracking scientific activity at the country-field-year level and for linking scientific output to economic outcomes in our cross-country analysis. To measure scientific output and quality, we compute both raw publication counts and citation-weighted measures at the country-field-year level.

Given the diversity of sources OpenAlex aggregates, we restrict our analysis to publications classified as research articles and books, excluding datasets and other non-research document types.[2] OpenAlex indexes publications in a wide range of languages, including English, German, French, and Spanish, among others. The language of each publication is determined algorithmically from its metadata. For our analysis, we include all languages covered by OpenAlex. Historically, English-language publications have accounted for ap-

[2]Specifically, we include items categorized as: 'article', 'preprint', 'report', 'libguides', 'paratext', 'peer-review', 'standard', 'review', 'letter', 'editorial', 'report-component', 'book-chapter', 'book', 'book-section', and 'dissertation'. We exclude works labeled as 'other', 'dataset', 'reference-entry', and 'grant'.

proximately 50% of records since the 1850s, with the share rising to more than 70% since the 1950s. Today, the vast majority of new scientific publications are in English, and metadata coverage is particularly strong for English-language material. Our dataset nonetheless remains inclusive of non-English publications, ensuring a broad and globally representative picture of scientific output over time.

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).[3] These open-source databases collectively form the underlying 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 information on more than 81 million patents from over 40 major patent authorities covering leading and developing countries from 1782–2025. 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 full-text information. 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 98% of patents containing a non-missing title. Abstract coverage is more limited, available for approximately 72% of patents overall; it begins in 1916 and increases 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 given their comprehensive availability. To validate the robustness of this approach, we compare an innovation network constructed using both titles and abstracts against one built

[3]We access DOCDB and INPADOC via https://publication-bdds.apps.epo.org/raw-data/products at November 2025.

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. For patents filed in languages other than English, DOCDB typically provides English translations of titles and abstracts, particularly for patents filed through international procedures such as the Patent Cooperation Treaty (PCT). For early patents lacking English translations, we supplement with translations from Google Patents.

Working with international patent data over a long historical period raises three challenges. The first is variation in technology classification systems across countries and time. Different patent offices have adopted different schemes, including the International Patent Classification (IPC), the Cooperative Patent Classification (CPC), and the U.S. Patent Classification (USPC). We use 3-digit IPC codes as our primary technology classification, which yields 123 distinct technological categories that can be further aggregated into eight 2-digit IPC groups.[4] Although the IPC system was formally established in 1971, it is now used by more than 100 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.

The second challenge is multi-filing: to protect intellectual property, inventors commonly file the same underlying invention with multiple patent authorities in different countries, forming a "patent family." To avoid double-counting, we retain only the first patent filed in each family and attribute all citations to that representative patent. We identify patent families using the INPADOC patent family information from DOCDB.

The third challenge concerns geographic attribution: filing a patent in a given country

[4]There are 131 technological categories at the 3-digit IPC level. We retain 123 after dropping 8 labeled as "other" within each 2-digit IPC group. These dropped categories account for less than 0.0005% of patents and have extensive missing periods.

does not necessarily indicate that the underlying inventive activity occurred there. We assign each patent to a geographic unit according to the country of residence of its inventor(s). If inventor information is unavailable, we use the country of the assignee(s) or, as a last resort, the country of the patent office. When a patent has inventors from multiple countries, we assign equal weights to each.

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[5], 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-level R&D expenditures to the country-sector-year level using Worldscope, a global firm-level database covering more than 0.11 million 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 (2021), 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

[5]We access the PatCit data at https://zenodo.org/record/4391095.

firms, we 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 bridge based on the joint appearance of Digital Object Identifiers (DOIs) across the two databases: for academic articles with both OpenAlex and Dimensions metadata, we match their respective field assignments and develop a crosswalk that maps 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.

2.4. Data on Production

We supplement the data above with production data from the World Input-Output Database (WIOD, Timmer et al., 2015). WIOD provides prices and quantities of inputs, outputs, and trade flows for 43 economies over 2000–2014, together representing more than 85% of world GDP. We obtain gross value-added and output for each country-sector-year. The sectoral classification in WIOD follows two-digit International Standard Industrial Classification (ISIC) Revision 4, which we project into IPC technology classes using the same SIC-to-IPC mapping described above for the R&D data. The production data serve two purposes in our analysis: they provide the consumption expenditure shares β_i used to parameterize the model's welfare weights across sectors, and they supply the sectoral value-added needed to calibrate the model's 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 intuition 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.

3.1. Modeling Innovation Network Using Textual Information

Table 1. Summary of Notations

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.

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, and this knowledge space has a fixed vocabulary of W distinct terms. Each research field i at time t, which comprises all documents newly created in that year (i.e., new academic papers, patents, etc.), 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 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 in terms of knowledge flow) 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. Denote the cosine similarity between the frequency vector of field i at time t and that of field j at time $t-1$:

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

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 similarities 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 solely from raw lagged similarity, one would erroneously 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.

It is worth discussing why this approach is helpful. 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 precisely 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

With the VAR structure in hand, 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. TFIDF rescales each component of the raw term-frequency vector by the inverse of the number of documents containing the corresponding word, thereby downweighting words that appear frequently across many documents while amplifying terms that are distinctive to a particular field. After computing TFIDF, we further rescale each document 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, the VAR system in (2) reduces to the following regression for each pair of fields (i, j):

$$v_{iwt} = \sum_j \omega_{ijt}^{\tau} v_{jw,t-\tau} + \gamma_{it} + \gamma_{iw} + \varepsilon_{iwt}, \qquad \tau = 1, \ldots, 10. \tag{3}$$

The dependent variable is the TFIDF frequency of word w in field i's word-frequency vector in year t. The explanatory variables are the lagged frequencies of word w in all other fields j (including field i itself) at lag τ. For a given lag τ, we estimate this equation for each field i at year t using data from the five-year window ending in year t, yielding $5 \cdot W$ observations—one for each word in each of the five years—with N right-hand-side variables and coefficients per regression.

We introduce field-by-word fixed effects γ_{iw} to absorb time-invariant vocabulary differences across fields. We also include field-by-time fixed effects γ_{it} to serve as a normalization of the TFIDF vector for each field in a given year. Because knowledge spillovers occur over time and the exact diffusion lag is a priori unclear, we adopt an averaging approach that

remains agnostic about the precise timing. Specifically, we estimate equation (3) separately for $\tau = 1, \ldots, 10$ and then average the estimated coefficients across lags to obtain

$$\omega_{ijt} = \frac{1}{10} \sum_{\tau=1}^{10} \omega_{ijt}^{\tau}.$$

This approach is equivalent to first averaging the lagged predictors $\bar{v}_{jwt} = \frac{1}{10} \sum_{\tau=1}^{10} v_{jw,t-\tau}$ and then running a single regression, which captures the cumulative effect of knowledge diffusion over the preceding ten years while remaining agnostic about the precise timing of individual spillovers.

Because each equation (3) involves a large number (N) of estimated coefficients for each field i, we impose a shrinkage constraint $||\boldsymbol{\omega}_{it}|| \leq \lambda$, where the tuning parameter λ is selected by cross-validation. Since the knowledge diffusion matrix $\boldsymbol{\Omega}_t$ is assumed to be non-negative with rows summing to one, we winsorize negative entries to zero and normalize so that $\mathbf{1}'\boldsymbol{\omega}_{it} = 1$.

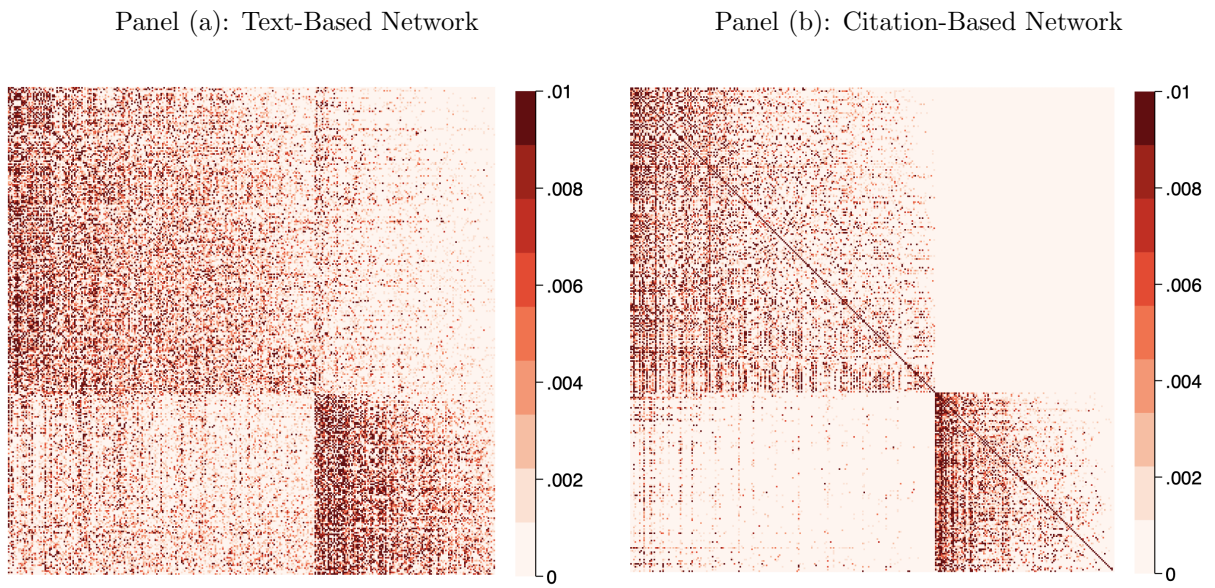
3.3. Network Description and Validation

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 heavily relies 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 remain 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 $\boldsymbol{\Omega}_{2010}$ 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. Text-Based and Citation-Based Innovation Network: Heatmap



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.

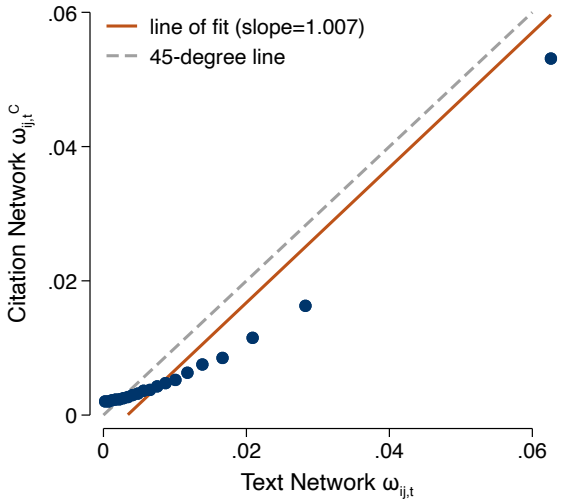
Figure 1, Panel (b) visualizes the citation-based network as a heatmap, combining patent-to-patent citations, paper-to-paper citations, and the patent-to-paper citations described in the previous section. Both networks exhibit similar within-domain block structure; however, our text-based network has 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 considerably sparse, and the patent-to-science quadrant is effectively empty—in sharp contrast to the rich cross-domain structure visible in the text-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.

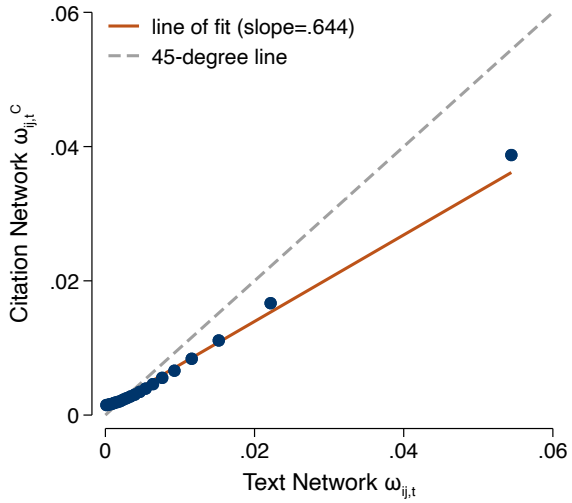
Validation: Alignment Between Text-Based and Citation-Based Networks We perform a series of tests to validate our measure's ability to capture knowledge flows. Figure 2 compares our text-based network with the citation-based network for within-patent fields (Panel a) and within-science fields (Panel 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.

Figure 2. Alignments Between the Text-Based and Citation-Based Network

Panel (a): Technology Class



Panel (b): Scientific Field



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 technological classes and Panel (b) considers the network within scientific fields.

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. The correlation also increases over time: the Pearson correlation within science rises from 0.19 in 1940 to 0.50 in 2020, with a similar pattern for within-patent correlations. This upward 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 Internet Appendix Section A.1.1.

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 innovation 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. These results confirm that our method successfully recovers directional knowledge flows and reconstructs the underlying network structure from text data.

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 our network captures meaningful knowledge

flows by examining whether terminology associated with major technological innovations spreads to fields identified as downstream recipients by our network.

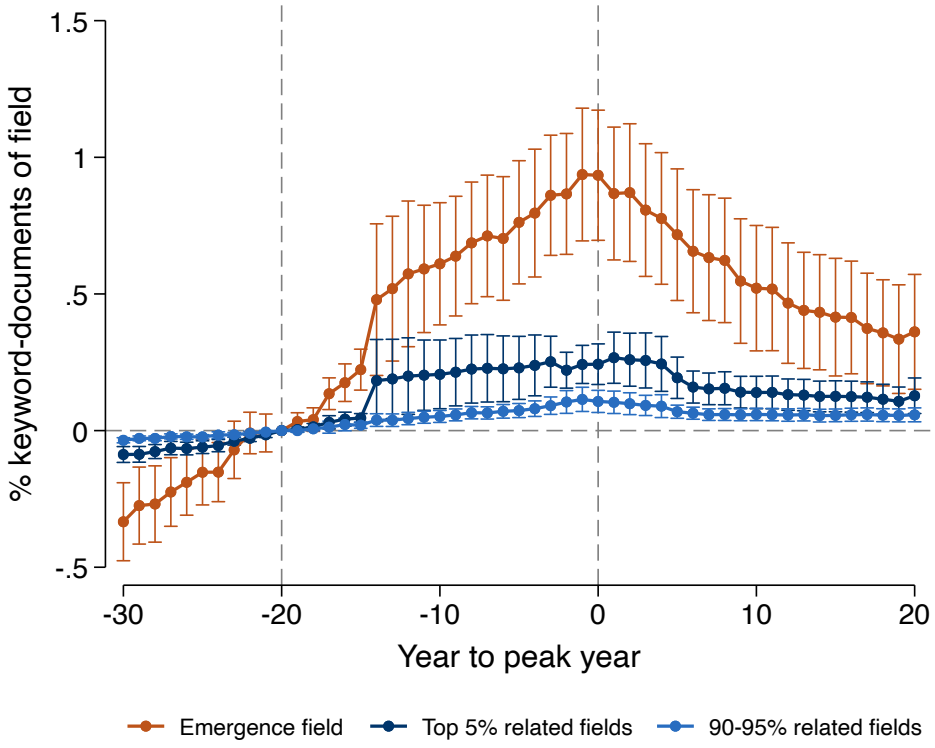
We implement this analysis in three steps. First, we identify major technological innovation episodes by applying a Markov regime-switching model to the Katz centrality time series for each field, detecting periods of rapid growth in a field's importance within the innovation network. This procedure identifies 79 distinct innovation episodes, 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 grey 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 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 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 that our text-based approach captures meaningful knowledge flows. Appendix Figure A.4 provides additional examples illustrating this diffusion process for specific innovation episodes.

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 distinctive 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.

4. Knowledge Evolution and Diffusion: Stylized Facts

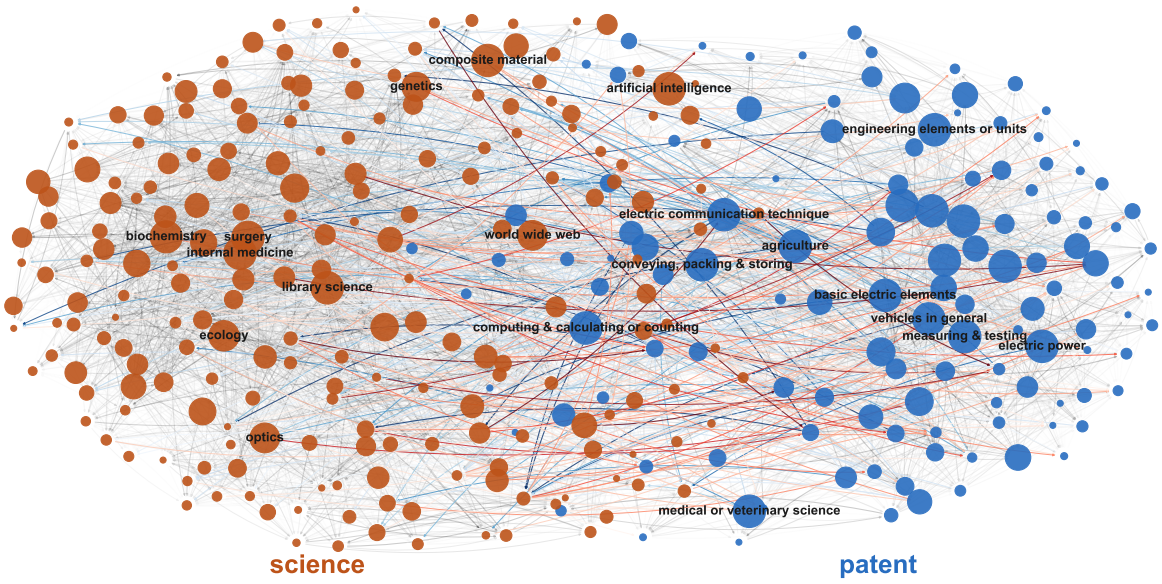
In this section, we document three stylized facts drawn from the innovation network that jointly motivate the key mechanisms in our theoretical framework and quantitative analysis: the structural dominance of science over technology in knowledge centrality; the heterogeneous and cyclical dynamics of 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: grey 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}.

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: grey 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}.

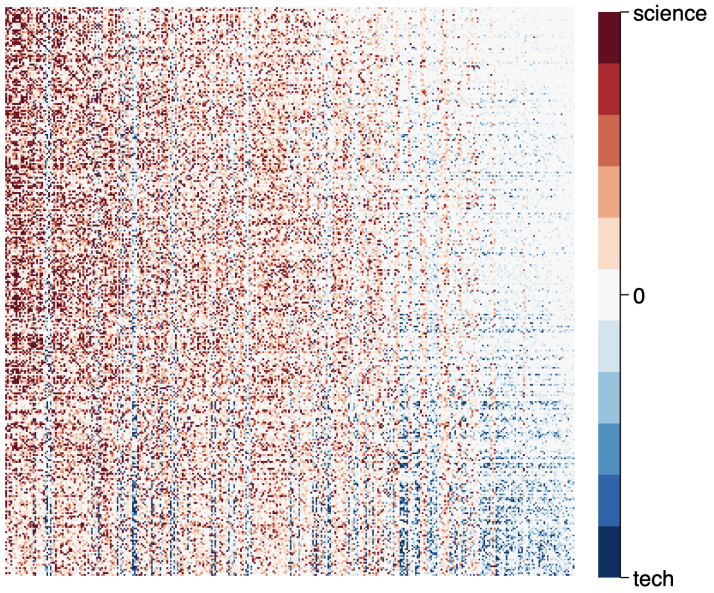
Cross-domain flows exhibit strong asymmetry. There are substantially more science-to-patent flows (red arrows) than patent-to-science flows (blue arrows), indicating that knowledge diffusion runs predominantly from science into technology rather than the reverse. Scientific fields such as the World Wide 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 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. This reflects the scale-free structure of knowledge spillovers. 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 2010. A field's Katz centrality aggregates all the direct and 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: nanotechnology, data science, chemical engineering, cell biology, and internet privacy hold the top five positions, each with Katz centrality above 0.011. The highest-ranked technology class—computing & calculating or counting—enters only at overall rank 17 with a Katz centrality of 0.0076, well below the top ten science fields. Other leading technology classes (basic electric elements, electric communication technique) similarly cluster below rank 20. This reflects the structural fact that science fields are more upstream: they influence both other science fields and downstream technology sectors, while technology sectors primarily influence other technology sectors. Appendix Table A.3 shows that this dominance of science in overall centrality rankings is not confined to 2010 but is a persistent feature of the post-WWII innovation network; Ap-

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 top-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.

pendix Figure A.3 traces the Katz centrality trajectories of the top fields in each domain over the full sample.

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 survival patterns reveal a modest level of instability, with distinct dynamics between science and technology. Patent fields experience relatively large short-run movements in their rankings, reflecting rapid and short-lived changes in centrality. These fluctuations are likely

Table 2. Top 10 Science and Technology Fields by Katz Centrality in 2010

Scientific Fields			Technological Classes		
Rank	Katz Centrality	Field Name	Rank	Katz Centrality	Field Name
1	0.0221	nanotechnology	17	0.0076	computing & calculating or counting
2	0.0179	data science	23	0.0071	basic electric elements
3	0.0129	chemical engineering	24	0.0069	electric communication technique
4	0.0124	cell biology	25	0.0068	biochemistry, beer, microbiology & enzymology
5	0.0124	internet privacy	47	0.0049	petroleum, gas or coke industries
6	0.0114	statistics	52	0.0048	medical or veterinary science
7	0.0112	engineering ethics	61	0.0044	information storage
8	0.0098	database	67	0.0041	lighting
9	0.0097	library science	69	0.0041	organic chemistry
10	0.0092	molecular biology	71	0.0041	measuring & testing

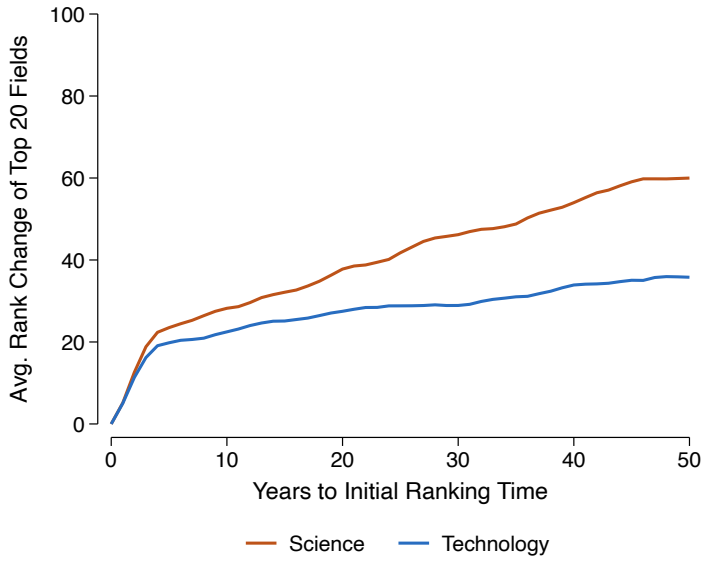
Notes. This table lists the top 10 scientific fields and the top 10 technology classes by Katz centrality in 2010, each selected from within their own domain. The Rank column reflects each field's position in the *combined* ordering across all 332 fields (209 science fields and 123 technology classes). Science fields occupy the top 16 positions in the combined ranking; the highest-ranked technology class enters at overall rank 17. Katz centrality is computed from the text-based innovation network $\mathbf{\Omega}_{2010}$ as described in Section 3; higher values indicate greater upstream influence in the knowledge diffusion process.

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 higher 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 over time, 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.

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

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.

model in state-space form, following Durbin and Koopman (2012).[6] 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 (grey) and their deterministic transition paths (blue, excluding the noise innovations) for each field.

[6]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:

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

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 60 years to exclude implausibly short- or long-run oscillations.

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 long-duration and smoother cycles with an average estimated periodicity of approximately 41 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 is consistent with historical episodes such as the post-war rise of nuclear physics, the genomics revolution, and the ongoing surge of machine learning.

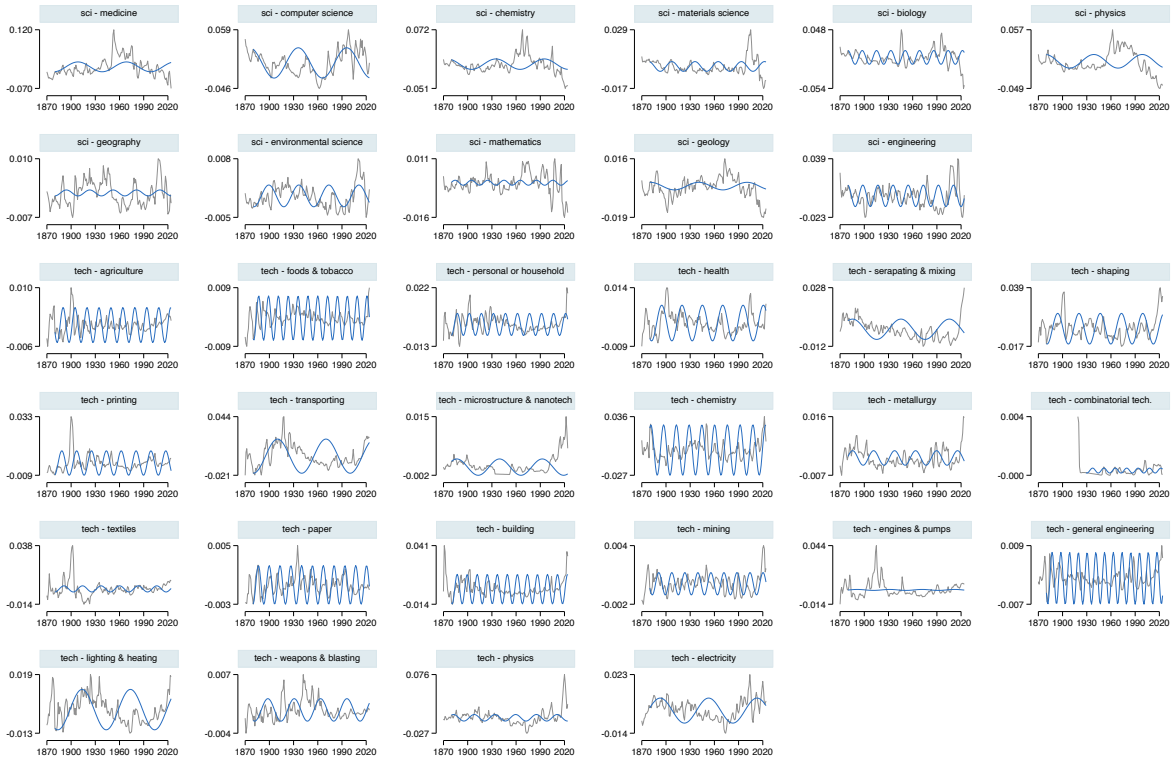
Technology fields, shown in the bottom panel, follow shorter-term and more volatile cycles with an average estimated period of approximately 29 years—roughly 12 years shorter than the science cycles. This faster rhythm likely reflects the tighter feedback loops between technological innovation and market 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.

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

We examine how breakthrough innovations affect a field's position within the innovation network. Our measure of field centrality is Katz centrality, which captures both direct and indirect knowledge flows. A field has high Katz centrality if it influences many other fields both directly and indirectly through chains of intermediate fields, meaning it occupies an upstream position in the knowledge diffusion process. This measure summarizes a field's overall reach and importance in shaping the evolution of knowledge across the entire innovation network—crucially including spillovers that are several steps removed from the source.

Figure 8 presents local projections 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 documents in the same field (over the next five years) to its similarity to past documents in the same field (over the

Figure 7. Long-Run Cyclicality of Network Centrality

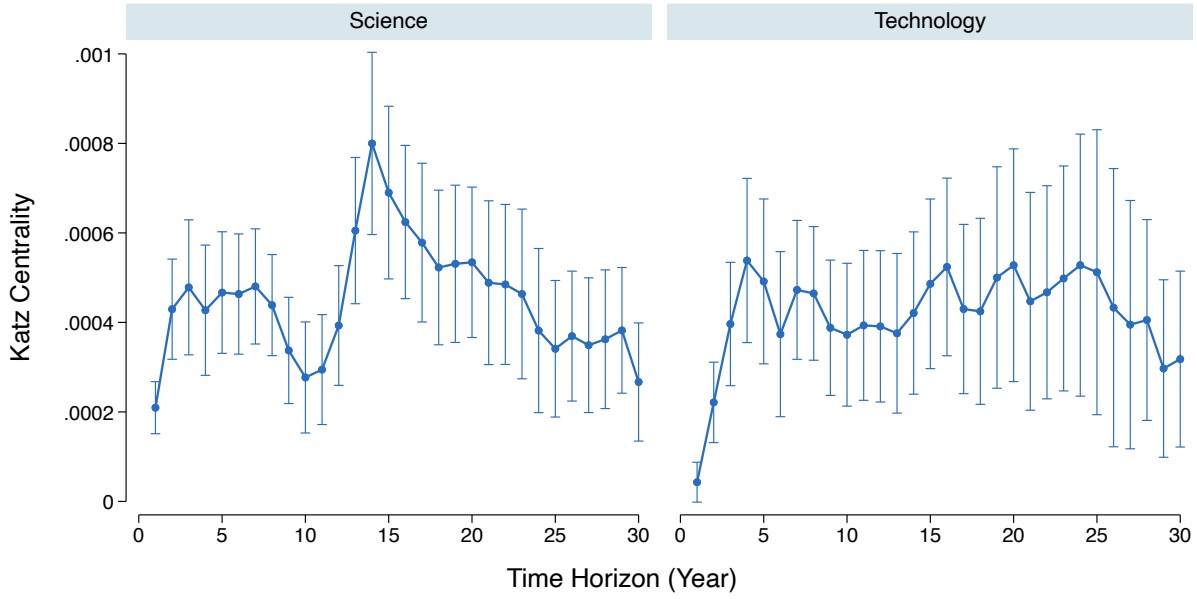


Notes. This figure shows the estimated cyclical components of Katz centrality from an unobserved-components time-series model (see footnote in Section 4), estimated separately for aggregate science fields (top panel) and aggregate technology classes (bottom panel). For each field, the grey lines show 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 60 years. Appendix Figure A.5 reports the R^2 and estimated periods of the models for each field.

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 removing cohort-year fixed effects to control for secular language shifts over time, we classify documents in the top 1% of the quality 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

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 cosine 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; documents in the top 1% of this distribution within each field-year are classified as breakthroughs. For each field-year, we count breakthrough innovations accumulated over the preceding ten years and compute its log. 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.

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 paradigm-shifting 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$_{i,t+5}$				Katz Centrality$_{i,t+20}$			
ln(Breakthrough, 99%)$_{i,t}$	0.00064***	0.00026***			0.00052***	0.00024***		
	(0.000)	(0.000)			(0.000)	(0.000)		
ln(Breakthrough, 90%)$_{i,t}$			0.00074***	0.00042***			0.00061***	0.00041***
			(0.000)	(0.000)			(0.000)	(0.000)
ln(Knowledge Stock)$_{i,t}$		0.00079***		0.00060***		0.00062***		0.00039***
		(0.000)		(0.000)		(0.000)		(0.000)
Observations	49,800	49,800	49,800	49,800	44,820	44,820	44,820	44,820
R-squared	0.32643	0.36923	0.35924	0.37873	0.29187	0.31216	0.31404	0.31990
No. of Fields	332	332	332	332	332	332	332	332
Field FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes. This table reports OLS estimates of the effect of breakthrough innovations on future Katz centrality at horizons of 5 years (Columns 1–4) and 20 years (Columns 5–8). The dependent variable is field i's Katz centrality in year $t + 5$ (Columns 1–4) or $t + 20$ (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 99th-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 top 1% (or top 10%) of the distribution of future-to-past textual similarity ratios within each field-year cohort, after removing cohort-year fixed effects to account for secular language evolution. 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 5 years (Columns 1–4) and 20 years (Columns 5–8). Across all specifications, the coefficients on the log count of breakthrough innovations—whether measured at the 99th or 90th percentile of the importance distribution—are positive, statistically significant at the 1% level, 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 table shows that both channels independently predict centrality, but that breakthroughs provide additional predictive power not captured by scale alone.

The estimated coefficients are also broadly stable across forecast horizons. The coefficient on log(Breakthrough, 99%) declines only modestly from 0.00064 at the 5-year horizon (Col-

umn 1) to 0.00052 at the 20-year horizon (Column 5), indicating that the centrality gains from breakthrough innovations are not purely transient—they persist and remain economically meaningful decades 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 nonparametric. Our main results characterize the social value of R&D in each sector as a function of this network.

5.1. Setup

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

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

The consumption good at each time t is a Cobb-Douglas aggregator over sectoral goods $\{y_{it}\}_{i=1}^K$. 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):

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

The collection of sectoral knowledge stocks $\{q_{it}\}_{i=1}^K$ 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 generate a flow of new innovation n_{it}:

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

$\mathcal{O}_i(\cdot)$ 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 helps innovation everywhere. 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 reflecting 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 at 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:

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

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 $K \times K$ matrix*

$\boldsymbol{\Omega}(t)$ *with entries*

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

The entry $\Omega_{ij}(t)$ measures the elasticity of sector i's knowledge aggregation 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: $\sum_{j=1}^{K} \Omega_{ij}(t) = 1$ for all i and t. The innovation network is allowed to vary over time, reflecting the evolving structure of knowledge flows across sectors.

A key object of this study, the matrix $\boldsymbol{\Omega}(t)$ represents a weighted directed graph in which economic sectors are the nodes. The entry $\omega_{ij}(t)$ captures the elasticity to which sector i's innovation benefits from sector j's existing knowledge stock. We refer to sector j as upstream to sector i and, conversely, i as downstream to 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 construction is not limited by any specific sector definition; innovation networks can be constructed across industrial sectors, technology classes, and scientific fields.

Interpreting Science. Among the K 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 increase from a marginal increase of R&D toward 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. In our empirical analysis later, we close the model by specifying a resource constraint and calibrate the absolute level to match existing estimates in the empirical literature.

To formalize, fix a time t and consider a given R&D allocation $\{x_{is}\}_{s\geq t}$. Hold the production allocation $\{\ell_{is}\}$ fixed. We inject an additional h_i units of R&D into sector i during a short time window $[t, t+\epsilon]$:

$$x_{is}^{\epsilon,\boldsymbol{h}} \equiv x_{is} + h_i \cdot \mathbf{1}_{t\leq s\leq t+\epsilon}.$$

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

$$\xi_{it} \equiv \lim_{\epsilon\to 0} \frac{1}{\epsilon} \frac{\partial V_t^{\epsilon,\boldsymbol{h}}}{\partial h_i}\bigg|_{\boldsymbol{h}=\mathbf{0}}. \tag{10}$$

Since the production allocation $\{\ell_{is}\}$ is held fixed, writing welfare (5) under the perturbed path gives

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

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

$$\xi_{it} = \lim_{\epsilon\to 0} \frac{1}{\epsilon} \int_t^\infty e^{-\rho(s-t)}\, \psi \sum_{j=1}^K \beta_j \frac{\partial \ln q_{js}^{\epsilon,\boldsymbol{h}}}{\partial h_i}\bigg|_{\boldsymbol{h}=\mathbf{0}} ds. \tag{11}$$

Equation (11) makes the mechanics transparent. Boosting R&D in sector i during $[t, t+\epsilon]$ 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 $\mathbf{\Phi}(t, s)$, which tracks how a perturbation to knowledge at time t ripples through the innovation network to affect knowledge at a later time $s \geq t$.

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*

$$\xi_{it} \propto_t \frac{\gamma_{it}}{x_{it}}, \tag{12}$$

where $\propto_t$ denotes proportionality across sectors i for each time t (the proportionality constant may vary with t), and $\boldsymbol{\gamma}_t = (\gamma_{1t}, \ldots, \gamma_{Kt})'$ is defined by

$$\boldsymbol{\gamma}_t' = \int_t^\infty \boldsymbol{\beta}' \, \mathbf{\Phi}(t, s) \, ds, \tag{13}$$

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

$$\frac{\partial \mathbf{\Phi}(t, s)}{\partial s} = -(\rho + \lambda) \left(\boldsymbol{I} - \frac{\mathbf{\Omega}(s)}{1 + \rho/\lambda} \right) \mathbf{\Phi}(t, s), \quad \mathbf{\Phi}(t, t) = \boldsymbol{I}. \tag{14}$$

Proof. See Appendix **??**. □

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.

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 $\mathbf{\Phi}(t, s)$ encodes this entire chain, discounted at the effective rate $\rho + \lambda$, with the innovation network $\mathbf{\Omega}(s)$ governing the spillover structure at each point in time. The integral in (13) is guaranteed to converge under our maintained assumptions: because $\mathbf{\Omega}(s)$ is row-stochastic for all s, the propagation matrix decays exponentially, $\|\mathbf{\Phi}(t, s)\|_\infty \leq e^{-\rho(s-t)}$, so the cumulative effect of knowledge spillovers remains bounded.[7] No additional assumptions—such as diagonalizability or commutativity of $\{\mathbf{\Omega}(s)\}$—are required.

An important feature of Proposition 1 is that the formula involves the observed innovation network $\mathbf{\Omega}(s)$ 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 $\mathbf{\Omega}(s)$ since they depend on knowledge stocks. However, this network response is itself proportional to the perturbation size h, so its welfare effect is second-order in h. The marginal value ξ_{it} involves only first-order derivatives and therefore 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 $\mathbf{\Omega}$ is time-invariant, the propagation matrix admits the closed form

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

This is precisely the framework developed in Liu and Ma (2025). The total social values become time-invariant and take the Leontief inverse form

$$\boldsymbol{\gamma}' \propto \boldsymbol{\beta}' \left(\boldsymbol{I} - \frac{\mathbf{\Omega}}{1 + \rho/\lambda}\right)^{-1}, \tag{16}$$

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

and the optimal allocation satisfies $\boldsymbol{x}^* \propto_t \boldsymbol{\gamma}$. The Leontief inverse decomposes the social value into how the consumer directly values each technology β_i and successively longer chains of knowledge spillovers, each discounted at rate ρ/λ per round of spillovers. Note that equation (16) is precisely the Katz centrality of the innovation network (with attenuation factor $1/(1+\rho/\lambda)$), weighted by the consumption vector $\boldsymbol{\beta}$. Intuitively, the social value of a sector under a static network is its Katz centrality: 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:

$$\boldsymbol{\gamma}_t' \propto_t \boldsymbol{\beta}' \left(\boldsymbol{I} + \frac{\boldsymbol{\Omega}_{t+1}}{1+\rho/\lambda} + \frac{\boldsymbol{\Omega}_{t+2}\,\boldsymbol{\Omega}_{t+1}}{(1+\rho/\lambda)^2} + \frac{\boldsymbol{\Omega}_{t+3}\,\boldsymbol{\Omega}_{t+2}\,\boldsymbol{\Omega}_{t+1}}{(1+\rho/\lambda)^3} + \cdots \right). \tag{17}$$

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 (17): the propagation matrix $\boldsymbol{\Phi}(t,s)$ in (13) plays the role of the discounted product $\boldsymbol{\Omega}_s \cdots \boldsymbol{\Omega}_{t+1}/(1+\rho/\lambda)^{s-t}$, 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; as we document there, the innovation network $\boldsymbol{\Omega}(t)$ 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*

$$x_{it} \propto_t \gamma_{it}. \tag{18}$$

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 **??** 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 $\boldsymbol{\Omega}(\cdot)$ 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 and 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. From Model to Data

Proposition 1 expresses the total social value of R&D in each sector as a time integral involving the propagation matrix $\mathbf{\Phi}(t, s)$, which solves a matrix differential equation driven by the time-varying innovation network $\mathbf{\Omega}(s)$. In general, the propagation matrix is difficult to compute. When the innovation network varies over time and its values at different dates fail to commute—i.e., $\mathbf{\Omega}(s)\mathbf{\Omega}(s') \neq \mathbf{\Omega}(s')\mathbf{\Omega}(s)$—the matrix exponential does not factor into a product of per-period terms. The solution must instead track the entire path of $\mathbf{\Omega}(s)$, resulting in an object analogous to the time-ordered exponential studied in quantum mechanics.[8] With $N = 332$ fields, directly computing or forecasting the $N \times N$ propagation matrix at each point in time would require predicting $N^2 \approx 110{,}000$ time series—an infeasible task.

We address this challenge by decomposing the innovation network using a joint diagonalization (JD) method. Let $\mathbf{\Omega}_t$ denote the $N \times N$ matrix of innovation linkages across N fields in year t. In our data, $N = 332$. JD seeks a shared K-dimensional basis, with $K \leq N$, such that each yearly network can be approximated by

$$\mathbf{\Omega}_t \approx \boldsymbol{U}\, \mathbf{\Sigma}(t)\, \boldsymbol{V}, \tag{19}$$

where $\boldsymbol{U} \in \mathbb{R}^{N \times K}$ collects K time-invariant basis vectors capturing common knowledge structure, $\mathbf{\Sigma}(t) = \text{diag}(\sigma_1(t), \ldots, \sigma_K(t)) \in \mathbb{R}^{K \times K}$ is a diagonal matrix of time-varying importance weights, and $\boldsymbol{V} = (\boldsymbol{U}^\top \boldsymbol{U})^{-1} \boldsymbol{U}^\top \in \mathbb{R}^{K \times N}$ is the Moore-Penrose pseudo-inverse of $\boldsymbol{U}$, satisfying $\boldsymbol{V}\boldsymbol{U} = \boldsymbol{I}_K$. The kth column of $\boldsymbol{U}$, denoted $\boldsymbol{u}_k$, 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. 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 the approximated networks commute at all dates:

$$\mathbf{\Omega}_s\, \mathbf{\Omega}_{s'} = \boldsymbol{U}\, \mathbf{\Sigma}(s)\, \mathbf{\Sigma}(s')\, \boldsymbol{V} = \mathbf{\Omega}_{s'}\, \mathbf{\Omega}_s,$$

[8] In quantum mechanics, the analogous object is called the *Dyson series* or the *Magnus expansion*. Exact evaluation requires computing the full path-ordered product, which is computationally prohibitive for large systems.

because the diagonal matrices $\boldsymbol{\Sigma}(s)$ and $\boldsymbol{\Sigma}(s')$ commute. Under this commutativity, the propagation matrix admits a spectral decomposition. In continuous time:

$$\boldsymbol{\Phi}(t,s) = \sum_{k=1}^{K} \boldsymbol{u}_k \exp\left[-(\rho+\lambda)(s-t) + \lambda \int_t^s \sigma_k(u)\, du\right] \boldsymbol{v}_k', \tag{20}$$

where $\boldsymbol{u}_k$ and $\boldsymbol{v}_k'$ denote the kth column of $\boldsymbol{U}$ and kth row of $\boldsymbol{V}$, respectively. Intuitively, each knowledge topic k contributes independently to the propagation of knowledge shocks: a perturbation to the knowledge stock at time t propagates through topic k at a rate governed by the cumulative eigenvalue $\int_t^s \sigma_k(u)\, du$, discounted at rate $\rho+\lambda$. Knowledge spillovers operating through topics with persistently high eigenvalues contribute more to the social value of R&D; those operating through topics with low eigenvalues contribute less. Substituting into the formula for total social value (equation 13):

$$\boldsymbol{\gamma}_t' = \sum_{k=1}^{K} (\boldsymbol{\beta}'\boldsymbol{u}_k) \left(\int_t^{\infty} e^{-(\rho+\lambda)(s-t)+\lambda \int_t^s \sigma_k(u)\, du}\, ds \right) \boldsymbol{v}_k'. \tag{21}$$

This expression captures the three forces 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 scalar integral summarizes the cumulative spillover potential of topic k, aggregating how its eigenvalue evolves over the entire future path. The term $\boldsymbol{v}_k'$ maps the topic-level value back to individual fields, distributing the social value according to each field's contribution to the topic. The discrete-time analogue is

$$\boldsymbol{\gamma}_t' \propto_t \sum_{k=1}^{K} (\boldsymbol{\beta}'\boldsymbol{u}_k) \left[1 + \frac{\sigma_k(t+1)}{1+\rho/\lambda} + \frac{\sigma_k(t+1)\,\sigma_k(t+2)}{(1+\rho/\lambda)^2} + \cdots\right] \boldsymbol{v}_k', \tag{22}$$

which extends the Leontief inverse formula (16) to the time-varying network setting. The first term captures the direct impact of knowledge on consumption; the second captures one round of spillovers at $t+1$, discounted at rate ρ/λ; successive terms capture higher rounds at $t+2, t+3, \ldots$, each weighted by the product of eigenvalues along the path. This is the time-varying analogue of the power series expansion of the Leontief inverse, with the crucial

difference that the spillover intensity at each horizon is now governed by the eigenvalue σ_k at that specific date rather than a constant.

Equations (21)–(22) represent a substantial dimensionality reduction. Instead of forecasting all $N^2 \approx 110{,}000$ entries of the innovation network at each future date, we need only forecast K eigenvalue time series. In our preferred specification with $K = 150$ knowledge topics, this reduces the prediction problem by a factor of approximately 700. As we document in the next subsection, the top 100 eigenvalues capture the bulk of the predictable variation in the network, further concentrating the effective dimensionality.

6. Estimates of the Economic Value of Science

Armed with our theoretical framework and innovation network measure, we next move to the primary goal of this paper: estimating the economic value of science.

6.1. Joint Diagonalization and the Structure of Innovation Network

We now describe the calculation of the JD decomposition (19) and characterize the structure of the resulting knowledge topics. Beyond simplifying the computation of the social value formula, 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.

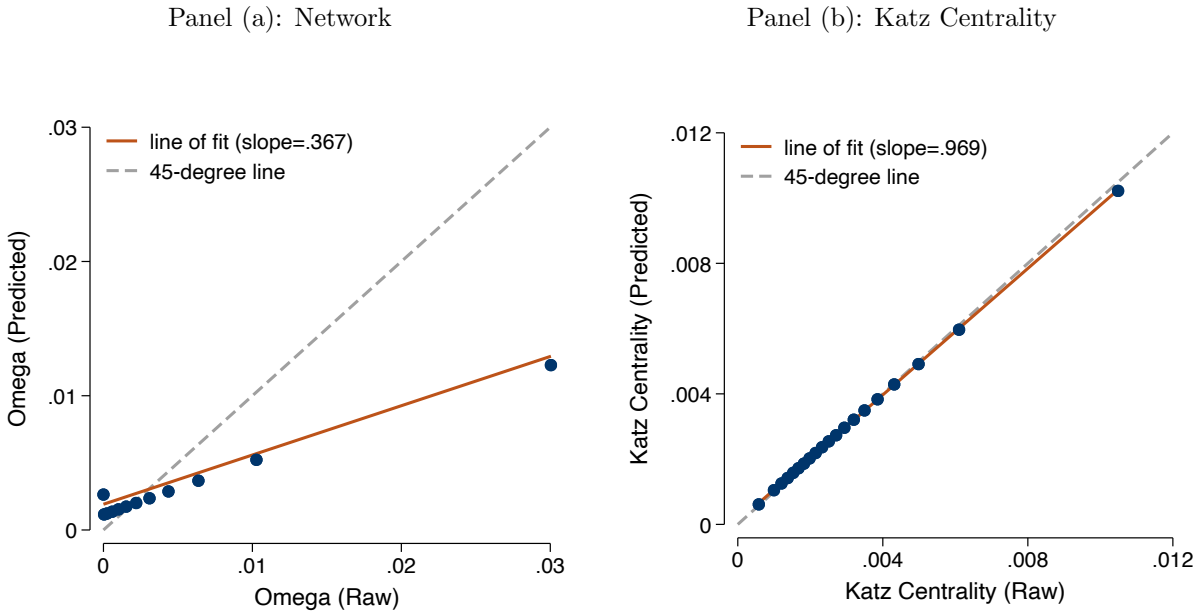
We estimate $\boldsymbol{U}$ and $\{\boldsymbol{\Sigma}(t)\}$ by solving a penalized reconstruction problem using stochastic gradient descent with Adam optimizer and backpropagation algorithm. Specifically, letting Katz_t denote the empirical Katz centrality vector and $\widehat{\text{Katz}}_t$ its model-implied counterpart, we choose $\boldsymbol{U}$ and $\{\boldsymbol{\Sigma}(t)\}$ to minimize

$$\min_{\boldsymbol{U},\{\boldsymbol{\Sigma}(t)\}} \; w_1 \left[\|\boldsymbol{\Omega}_t - \hat{\boldsymbol{\Omega}}_t\|_2^2 + \frac{1}{2} \|\min(\hat{\boldsymbol{\Omega}}_t, 0)\|_2^2 \right] + w_2 \left[\|\text{Katz}_t - \widehat{\text{Katz}}_t\|_2^2 + \frac{1}{2} \|\min(\widehat{\text{Katz}}_t, 0)\|_2^2 \right],$$

subject to the normalization and non-negativity constraint $0 \leq \phi_k(u) \leq 1$ for every field u and topic k. Intuitively, the first term ensures that the low-dimensional representation faithfully reconstructs the observed bilateral knowledge flows, while the second term preserves the global centrality structure of the network—the ranking of fields by their overall

upstream influence. The non-negativity constraint $\hat{\boldsymbol{\Omega}}_t \geq 0$ reflects the maintained assumption that knowledge spillovers are non-negative. We implement simple reparameterizations and rescaling to enforce the constraints and normalize the columns of $\boldsymbol{U}$.

Figure 9. The Empirical Performance of Joint Diagonalization



Notes. This figure shows the empirical performance of joint diagonalization. Panel (a) shows the network, and panel (b) shows the Katz centrality.

Figure 9 reports the empirical performance of the JD approximation, and Appendix Figure A.6 summarizes the associated model-selection procedure across different choices of K and the loss-weight parameter w_1. Our preferred specification sets $K = 150$ and $w_1 = 0.8$, which delivers a good trade-off between parsimony and fit.

Performance of the Joint Diagonalization Now we assess the ability of the joint diagonalization's performance at recovering the original network matrices. In Appendix Figure A.6, we vary w_1 from 0.6 to 1 in increments of 0.1 and, for each w_1, vary the number of factors K from 10 to 332 in increments of 10, reporting the correlation between the observed and reconstructed network $\boldsymbol{\Omega}_t$ and Katz centrality Katz_t for each combination. The network correlation rises sharply as K increases from 10 to 150, then plateaus at approximately 0.46, indicating that 150 factors are sufficient to capture most of the network matrices' systematic

variation.

The correlation between observed and reconstructed Katz centrality follows a similar pattern and becomes very high once w_1 deviates from 1, reaching about 0.98. In our preferred specification, the JD approximation accounts for approximately 98% of the variation in Katz centrality and about 46% of the variation in the network. Figure 9 illustrates the alignment between the JD approximation and the observed network and Katz centrality in this specification: the points lie close to a straight line, with slopes of 0.37 for the network and 0.97 for Katz centrality.

Interpreting $\boldsymbol{U}$ The resulting knowledge topics have a clear economic interpretation. Each column of $\boldsymbol{U}$ defines a "knowledge topic", characterized by its loadings on different fields. For instance, one factor loads heavily on machine learning, control engineering, and artificial intelligence while loading negatively on generic communication and computing categories; we label this factor *AI & Control Systems*. Another factor combines machine learning, computer security, and embedded systems and we interpret it as *Secure Embedded Systems*. A third factor loads strongly on world-wide-web and internet-privacy related technologies, which we describe as *Web Privacy & Computing*. Additional factors pick out clusters such as *Nuclear & Atomic Physics* and *Combustion Engine Technologies*.

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

6.2. Predictability of Joint-Diagonalized Innovation Network

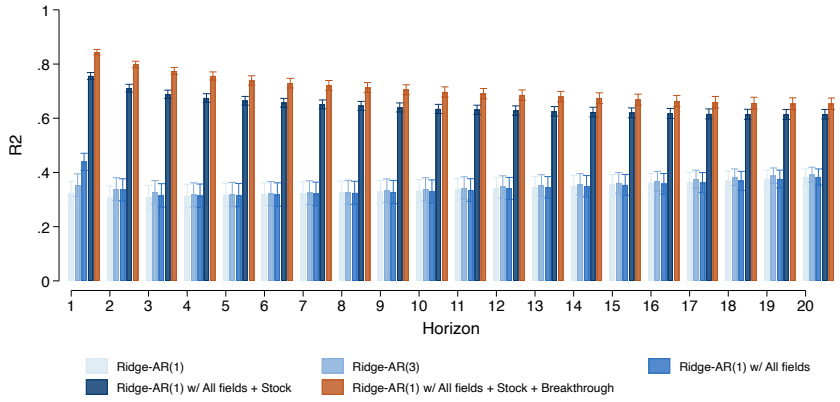
We next discuss the predictability of the innovation network. The JD operation reduced the prediction problem from predicting every component in the innovation network to predicting JD eigenvalues. The dimension is significantly reduced, from $N \times N = 110,224$ to $K = 150$. Conceptually, after JD, over-time variation in the eigenvalues $\sigma_k(t)$ translates into variation in the spillover matrix; if we can forecast $\sigma_k(t+u)$, we can recover the future evolution of the spillover structure $\boldsymbol{\Omega}_t$.

We assess how well future eigenvalues can be predicted from their own history and from

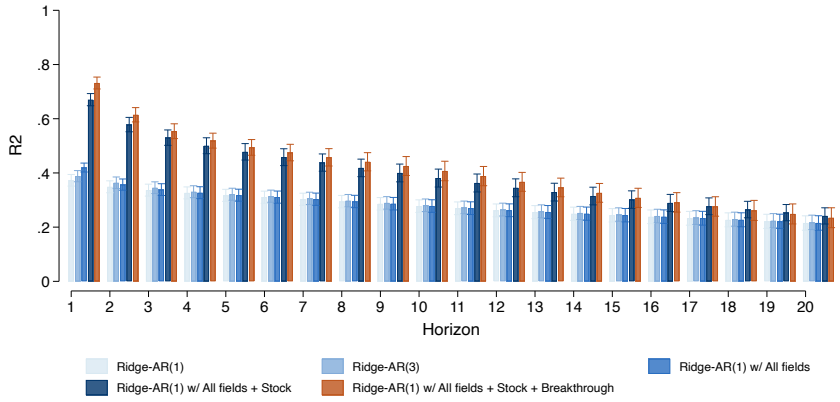
additional economic predictors. We implement two complementary approaches, focusing on both in-sample and out-of-sample performance. For the in-sample exercise, we estimate a time-series model for each eigenvalue factor k using the full sample data from 1870 to 2024, and then use the estimated model to forecast the path of eigenvalues for the next 50 years based on information available up to year t, computing the R^2 for each forecast horizon. For the out-of-sample exercise, we re-estimate the time-series model recursively using data only up to year t, generate forecasts for $t+u$, and again compute the corresponding R^2. We burn-in the first 30 years of the sample to ensure the models are well-estimated.

Figure 10. Predictability of Eigenvalues

Panel (a): In-Sample R2



Panel (b): Out-of-Sample R2



Notes. This figure shows the in-sample and out-sample R2 of the predictability of eigenvalues.

Figure 10 reports the in-sample and out-of-sample R^2 for specifications that differ in the

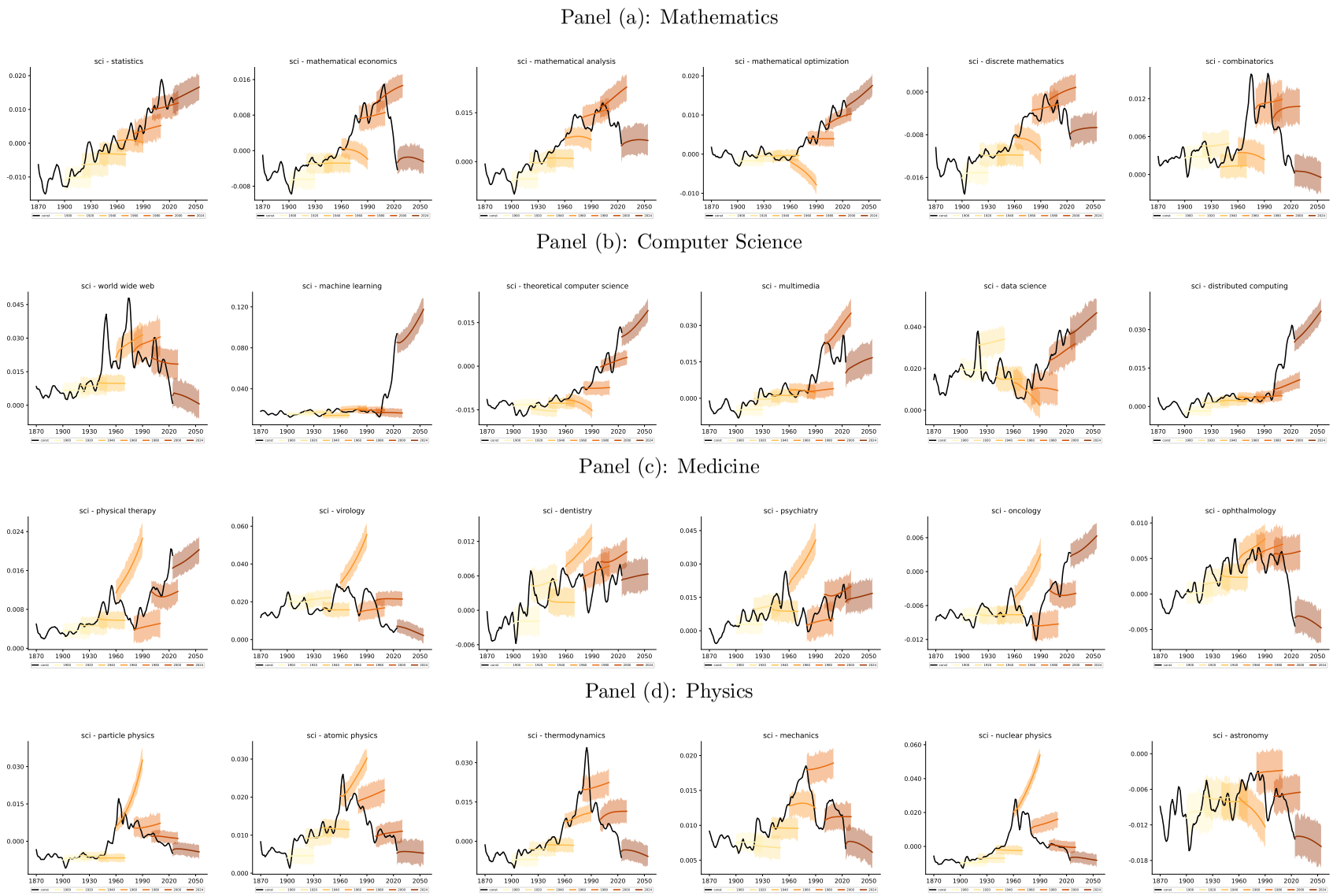
set of predictors, using our preferred Ridge model with penalty parameter $\lambda = 100$. We find that the eigenvalues are substantially predictable, and that economic fundamentals—knowledge stocks and breakthrough innovations—are the primary source of this predictability. A simple autoregressive specification captures about one-third of the in-sample variation ($R^2 \approx 0.33$), indicating modest but nontrivial persistence. Adding the knowledge stocks of all fields as exogenous predictors roughly doubles the explanatory power, raising the in-sample R^2 to approximately 0.75 and the out-of-sample R^2 to about 0.67 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. Including the count of breakthrough patents as additional predictors further improves performance: the in-sample R^2 reaches approximately 0.84 and the out-of-sample R^2 about 0.73 at short horizons, with meaningful gains at long horizons as well (in-sample $R^2 \approx 0.65$, out-of-sample $R^2 \approx 0.23$). This is consistent with the finding in Section 4 that breakthrough innovations reshape the network's centrality structure over extended horizons.

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 a 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 10^4 in factors of 10. The choice $\lambda = 100$ provides a favorable balance between fit and regularization.

6.3. The Economic Value of Science

We now present our estimates of the economic value of science across individual scientific fields. The estimation proceeds in three steps. First, using the JD decomposition (19), we forecast the eigenvalues $\{\sigma_k(t+1), \sigma_k(t+2), \ldots\}$ over the next 50 years using the Ridge prediction model from the previous subsection. Second, we substitute the forecasted eigenvalues into the discrete-time social value formula (22) to compute $\boldsymbol{\gamma}_t$ for each field. Third, we construct confidence intervals using a bootstrap procedure that resamples from the prediction residuals.

Figure 11. The Economic Value of Science



Notes. This figure shows the estimated economic value of science for each scientific field. Each panel includes six selected scientific fields from the same major field, including mathematics, computer science, medicine, and physics.

Figure 11 displays the estimated social value for 24 selected scientific fields, organized into four panels by major discipline: mathematics, computer science, medicine, and physics. In each subplot, the black line plots the field's Katz centrality over time. As we established in Section 5.2, Katz centrality corresponds to the social value that would obtain if the innovation network at a given date t were held constant forever—it captures the "static" component of the social value, abstracting from future changes in the network structure. The colored lines originating at different dates show our out-of-sample forecasts of the social value: each curve, starting from year t, uses data available up to t to predict the eigenvalue paths and compute the forward-looking social value. At its starting year t, the gap between a forecast curve and the Katz centrality 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 social value evolves as the predicted network changes unfold.

Several patterns emerge from the figure. Within mathematics, mathematical economics exhibits a pronounced rise through the mid-20th century—consistent with the golden era of general equilibrium theory and game theory—followed by a sustained decline as these foundational frameworks matured and their marginal contribution to downstream fields diminished. Mathematical optimization, by contrast, shows a steady rise, reflecting its growing importance as a building block for machine learning, operations research, and engineering applications. Statistics displays a similar upward trajectory, consistent with the expanding role of statistical methods across the sciences and technology.

The computer science panel reveals the most dramatic dynamics. Machine learning exhibits near-exponential growth in social value, with the forecast curves projecting continued rapid increases—reflecting its role as a general-purpose scientific tool with broad downstream applications. The World Wide Web shows the opposite pattern: a sharp rise in the 1990s followed by a sustained decline, suggesting that its foundational contribution to knowledge diffusion has been largely absorbed into downstream fields. Data science and distributed computing show more moderate but sustained growth.

In medicine, the selected fields display relatively stable social values, with virology and oncology showing modest upward trends. The physics panel reveals a general pattern of

declining social value across most subfields—particle physics, atomic physics, nuclear physics, and astronomy all trend downward over recent decades. This is consistent with the centrality dynamics documented in Section 4: as newer fields in computer science and biology rise in the centrality rankings, established physics subfields are gradually displaced. Mechanics and thermodynamics show more stable trajectories, likely reflecting their continued relevance to engineering applications.

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, with the forecasts diverging 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 lose precision over time. The fact 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.

7. Conclusion

We develop a framework for measuring the economic value of scientific research by combining a text-based innovation network with an endogenous growth model. Using 463 million scientific articles and 81 million patents, we construct an annual network of knowledge flows among 209 scientific fields and 123 technology classes from 1870 to 2024. We show that the social value of R&D in any sector—including scientific fields that do not directly produce consumption goods—is fully characterized by its discounted influence on downstream sectors through this network. In the static case, social value reduces to Katz centrality; the time-varying formulation extends this to a forward-looking measure reflecting the evolving structure of knowledge flows.

The findings connect to a long-standing question in R&D policy: how should governments divide research funding between basic science and applied technology? The United States alone directs over $180 billion annually to R&D, yet the allocation across fields has been guided largely by precedent and advocacy rather than systematic welfare analysis. A core difficulty is that basic science does not produce marketable goods, so its contribution cannot be priced by standard methods. Our framework addresses this gap by tracing how scientific

knowledge propagates through the innovation network into downstream technologies and consumer welfare, providing the first field-level, welfare-grounded accounting of what basic science contributes to economic growth. The finding that science's welfare share substantially exceeds its R&D share gives quantitative content to the case for basic research funding articulated since Bush (1945) and formalized in the spillover literature (Griliches, 1958; Jaffe, 1986; Jones and Williams, 1998)

Our estimates also reveal that the relative value of scientific fields is not static. Machine learning and mathematical optimization have risen sharply; some classical physics subfields have declined as their foundational contributions were absorbed downstream. These dynamics suggest that fixed science funding allocations may themselves be inefficient, and that periodic reassessment informed by the evolving structure of knowledge flows could improve returns to public research investment.

Our analysis is subject to important caveats. The welfare formula is derived for marginal reallocations; applying it to large counterfactuals extends a local approximation globally and is best interpreted as an upper bound. The text-based network measures terminology diffusion, which need not perfectly coincide with genuine knowledge adoption; TFIDF further misses synonymous concepts and has uneven historical coverage. Welfare calculations are sensitive to the discount rate ρ, and this sensitivity is asymmetric—science's share responds more to ρ than technology's because its value accrues through long indirect chains. The qualitative finding that science is underweighted relative to its social value is robust across specifications; the precise magnitude is not. Finally, the model omits creative destruction and directed technical change—structural boundaries that limit its ability to assess whether and how markets would self-correct.

Two extensions seem essential. First, endogenizing the innovation network by modeling how funding decisions reshape knowledge flows—required to evaluate large policy changes rather than marginal ones. Second, connecting field-level estimates to micro-level mechanisms of knowledge production: scientists, teams, institutions, and funding channels. Our framework provides a welfare-grounded accounting of science's indirect contribution to growth—not a substitute for expert judgment in science policy, but 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.

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, 2021. and Nan Li,"Growth Through Intersectoral Knowledge ....

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.

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."

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.

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

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 Endogeneous 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.* Vol. 27863, National Bureau of Economic Research.

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*, Forthcoming.

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. 2021. "Technological Obsolescence." *Working Paper*.

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.

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

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.

Timmer, Marcel P, Erik Dietzenbacher, Bart Los, Robert Stehrer, and Gaaitzen J De Vries. 2015. "An illustrated user guide to the world input–output database: the case of global automotive production." *Review of International Economics*, 23(3): 575–605.

Wang, Dashun, and Albert-László 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. 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 $\mathbf{\Omega} \in \mathbb{R}^{N \times N}$ 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 $\boldsymbol{V}_t \equiv [\boldsymbol{v}_{1t}, \cdots, \boldsymbol{v}_{Nt}]'$ 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 $\mathbf{\Omega}_{i,\cdot}\boldsymbol{V}_{t-1} + u'_{it}$. Here, $\mathbf{\Omega}_{i,\cdot}$ is the i-th row of true innovation network $\mathbf{\Omega}$, 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 words distributions across all fields in the previous period, with weights determined by the true innovation network Ω. We then obtain the term-frequency matrix 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 simluated 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 compuated as $\hat{\boldsymbol{\Omega}}_t = \boldsymbol{P}_{t,t-1} \boldsymbol{P}_{t-1}^{-1}$.

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 (a) of the right panel, following a structure of lower triangular matrix.

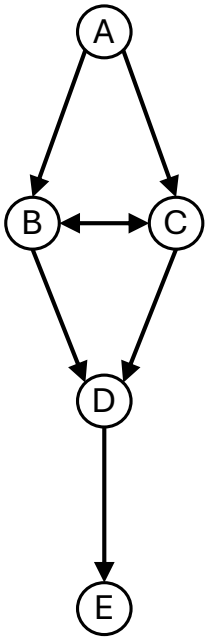
Panel (b) shows the estimated innovation network recovered using our methodology. The estimated matrix accurately approximates the structure of the true network (i.e., upstream vs downstream fields), with a correlation coefficient of 0.93. For example, field A maintains a high self-weight and significant 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 derived from the estimated network also closely track those from the true network, preserving the rank order and resulting a correlation coefficient of 0.99.

Panel (c) compares 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 is mixed of different information, including the persistent similarities between fields instead of 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 (a): 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): 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 (c): 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, estimated network using our methodology, and the cosine similarity matrix.

A.1.2. 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 Here is a note on how to link 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 assume that each document $d = 1, \ldots, D_{i,t}$ follows a multinomial count distribution with parameter $v_{i,t}$:

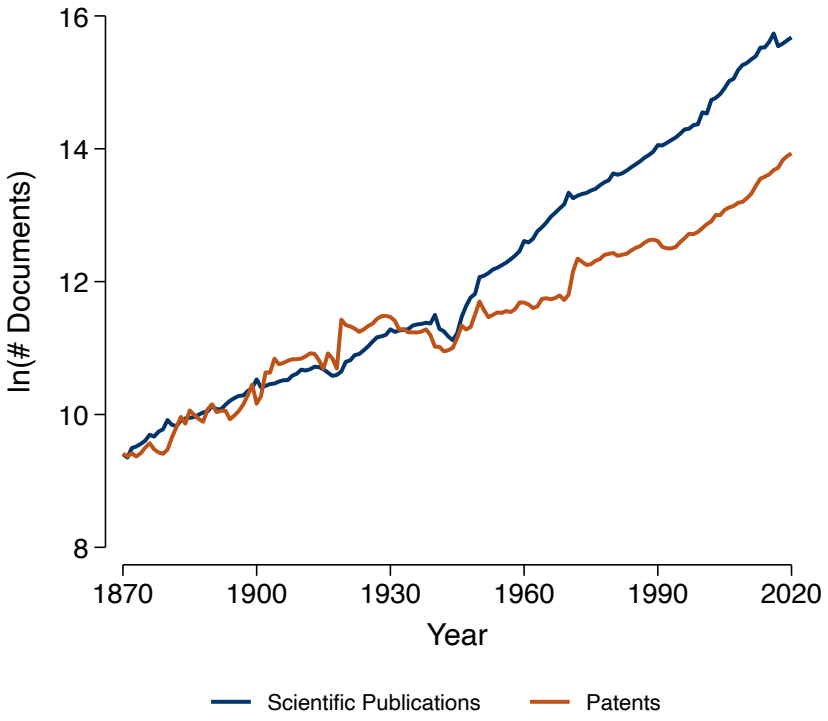
$$x_d \sim \text{Multinomial}(v_{i,t}, n_d), \tag{A1}$$

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 like 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. Basically, it models the data generating process for text with a set of probabilistic rules.

A.2. Additional Results

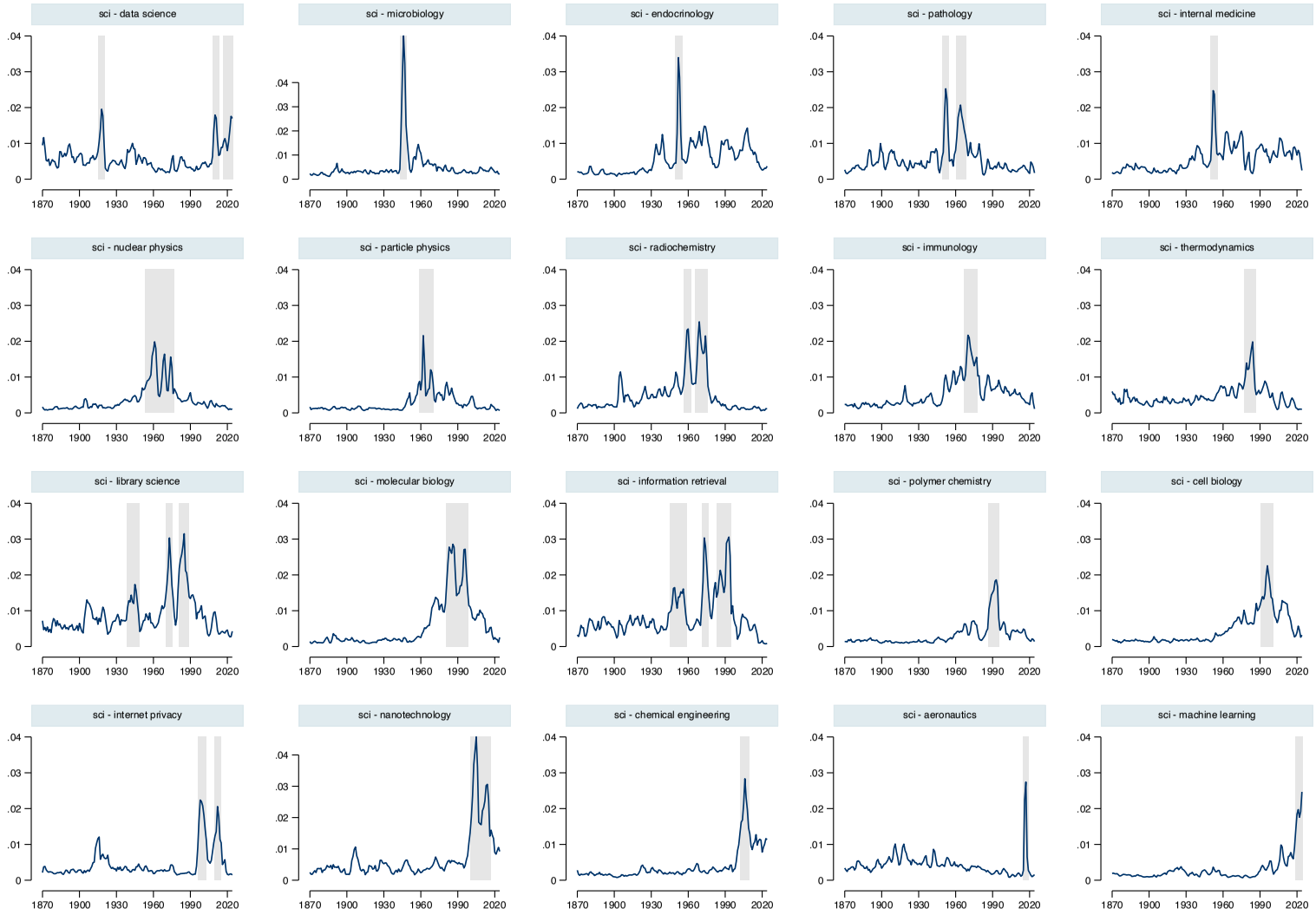
Figure A.2. Number of Patents and Publications Over Time



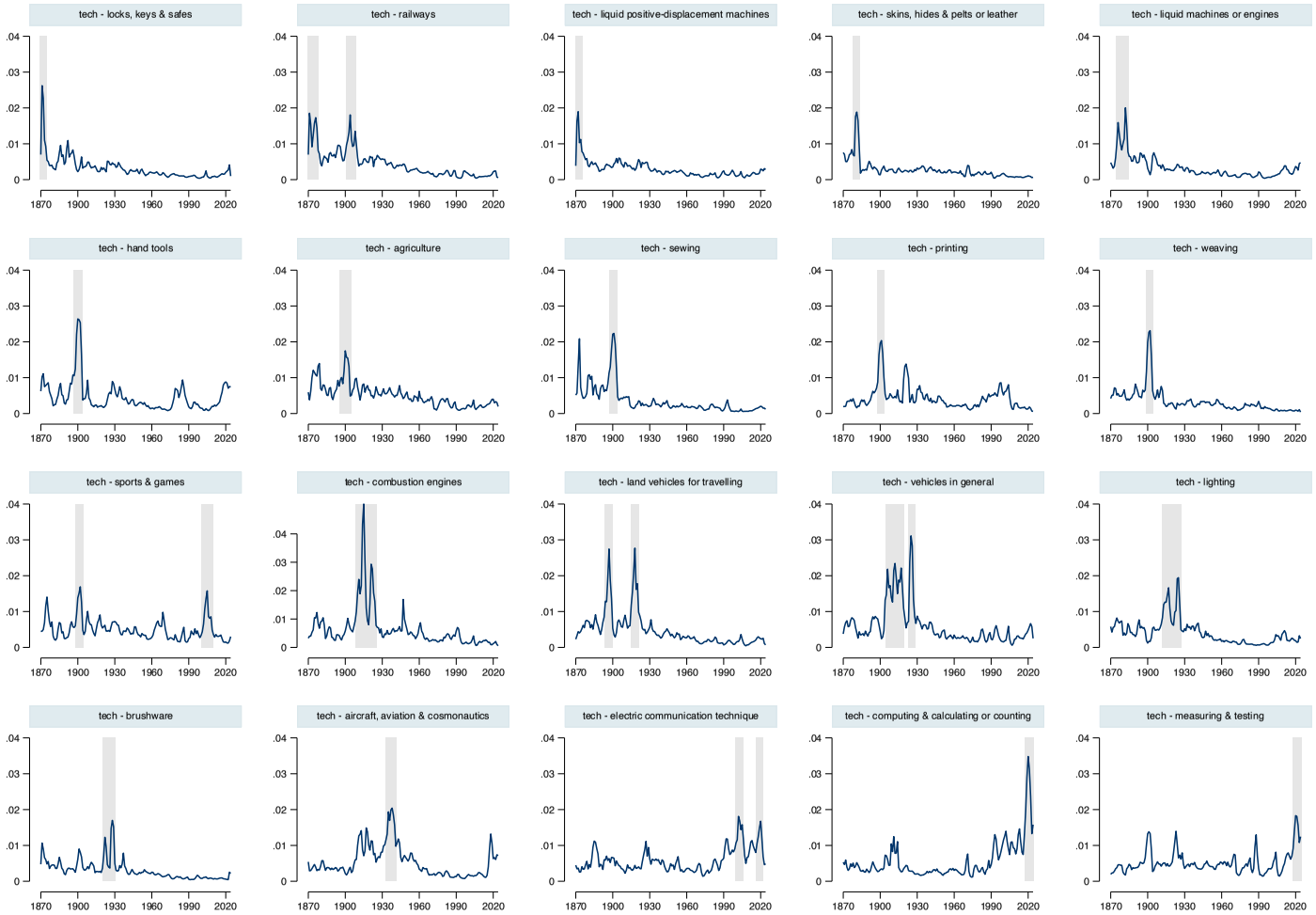
Notes. This figure shows the number of patents (solid line) and articles (dashed line) over time.

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

Panel (a): Science Fields



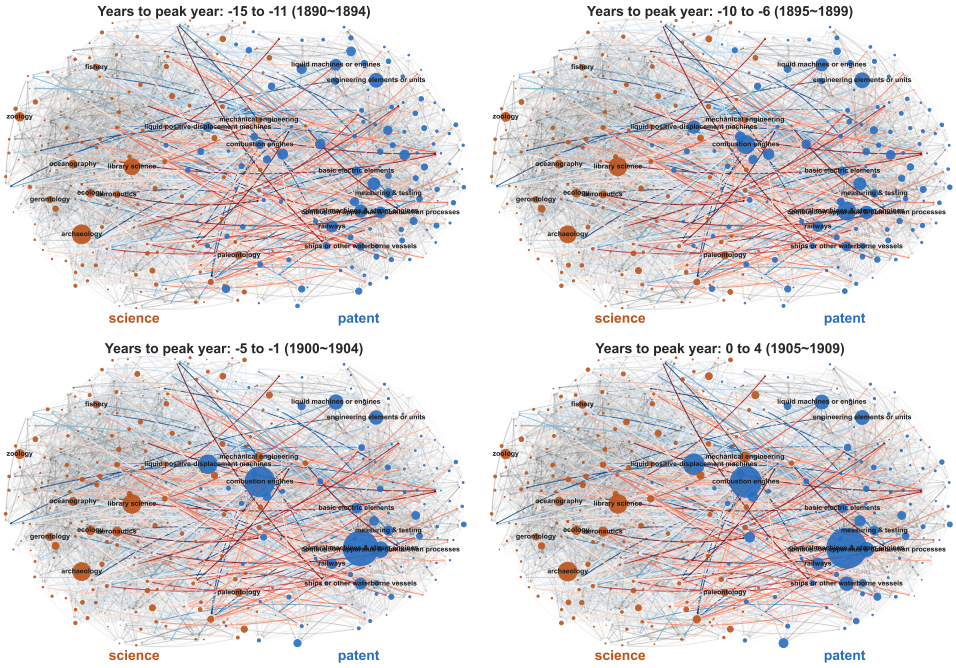
Panel (b): Technology Fields



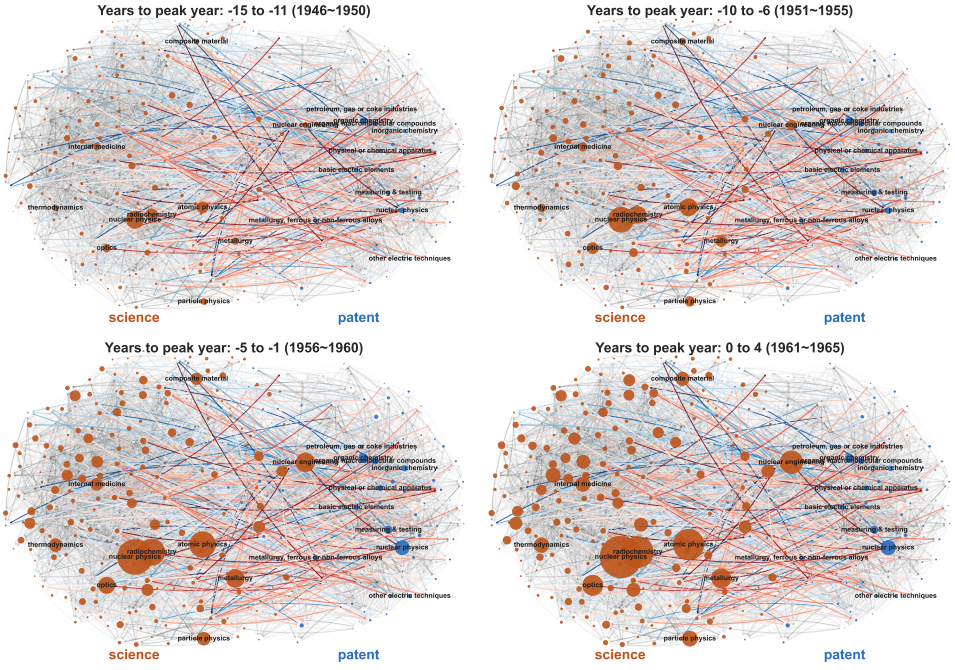
Notes. This figure shows the Katz centrality of science and technology fields over time. We show the top 20 fields sorted by their maximum Katz centrality across all years. The grey area represents the major technological innovation waves.

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

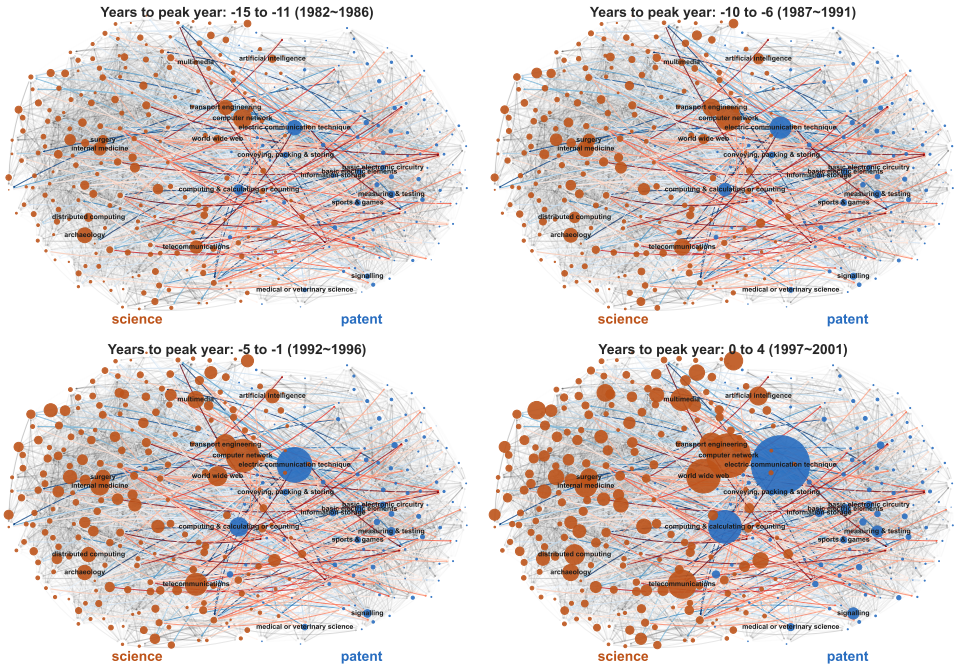
Panel (a): Tech - General Machines & Steam Engines (1905)



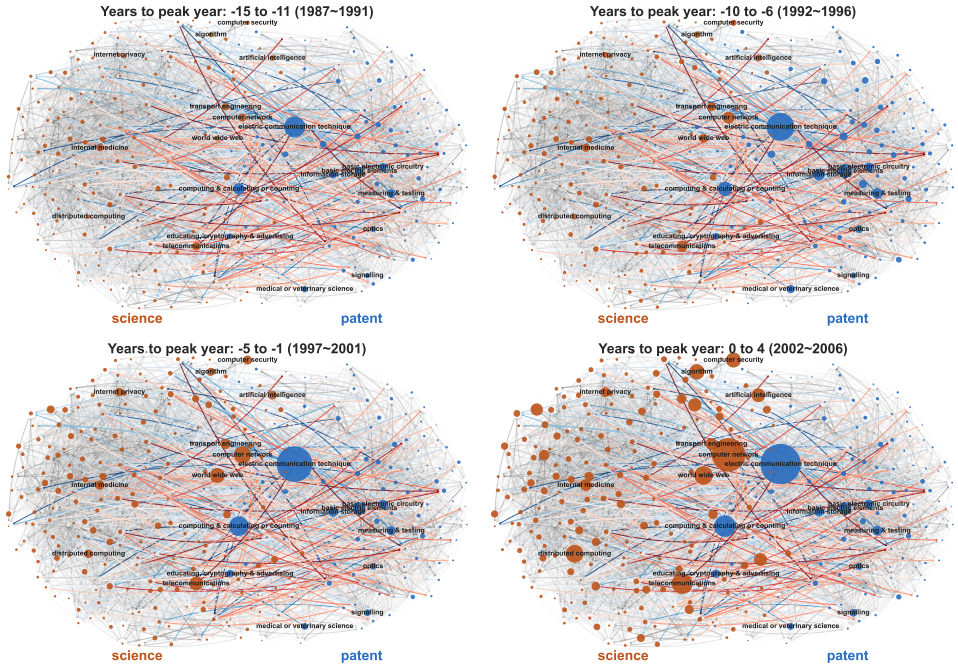
Panel (b): Sci - Nuclear Physics (1961)



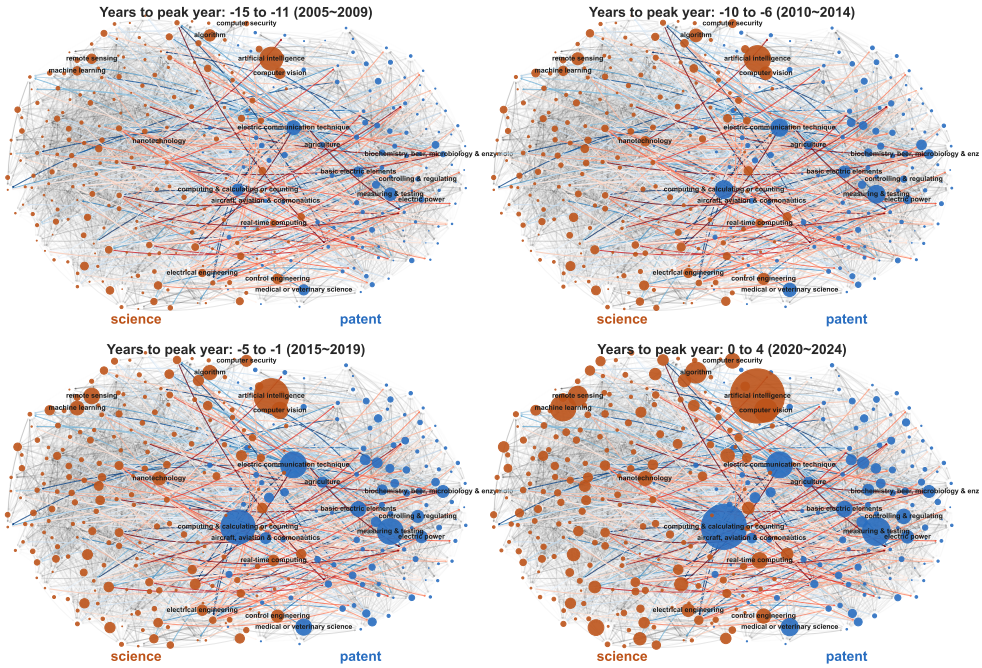
Panel (c): Sci - Computer Network (1997)



Panel (d): Tech - Electric Communication Technique (2002)



Panel (e): Tech - Measuring & Testing (2020)



Panel (f): Sci - Data Science (2023)

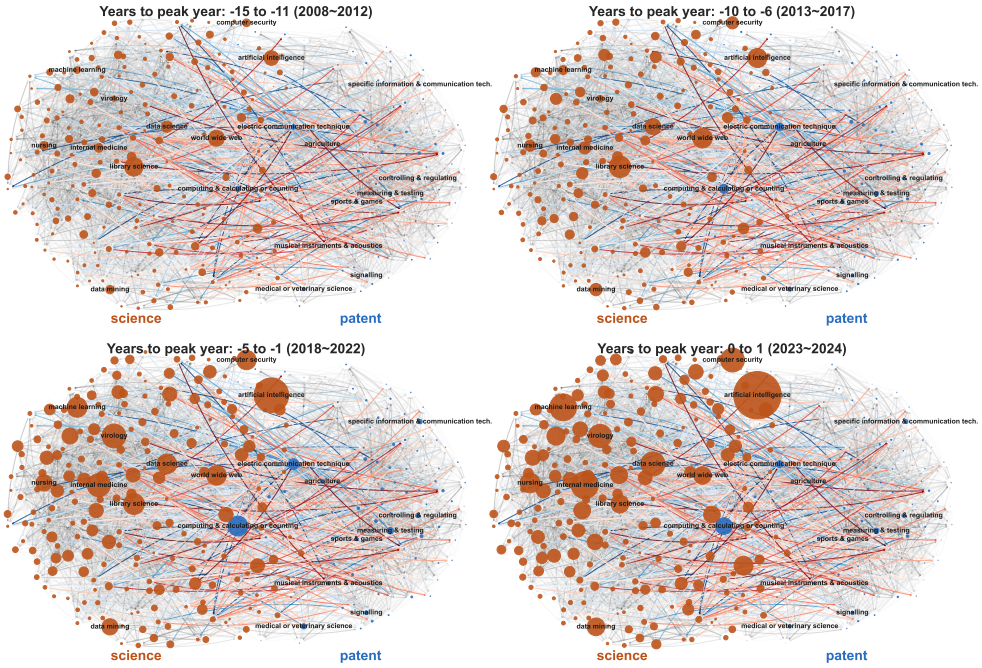
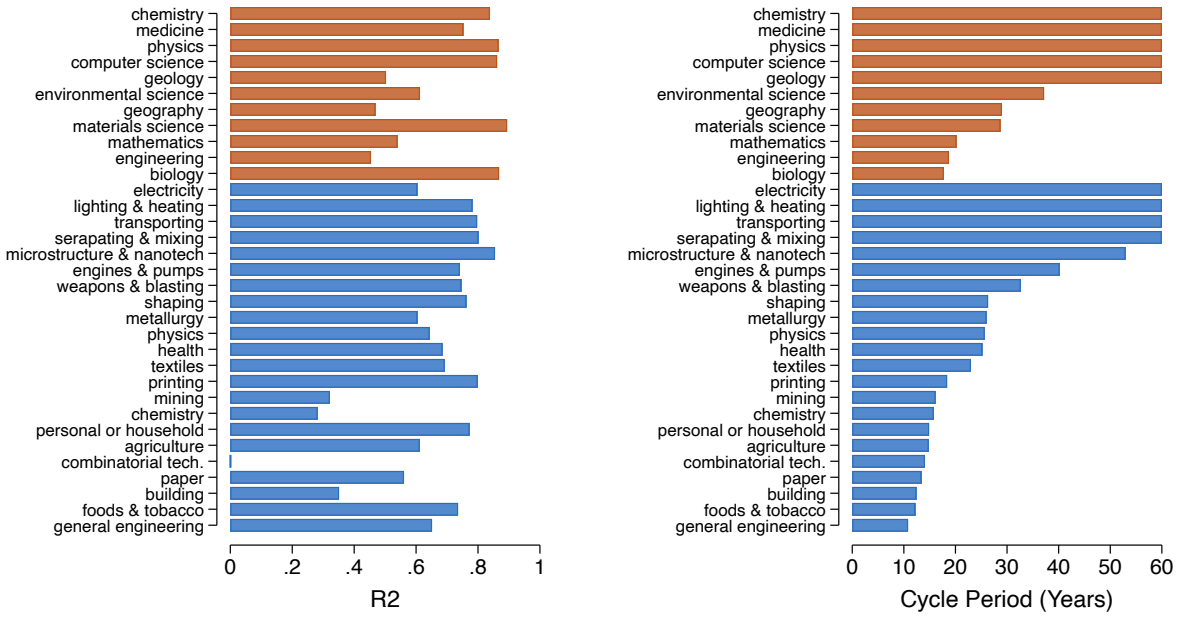


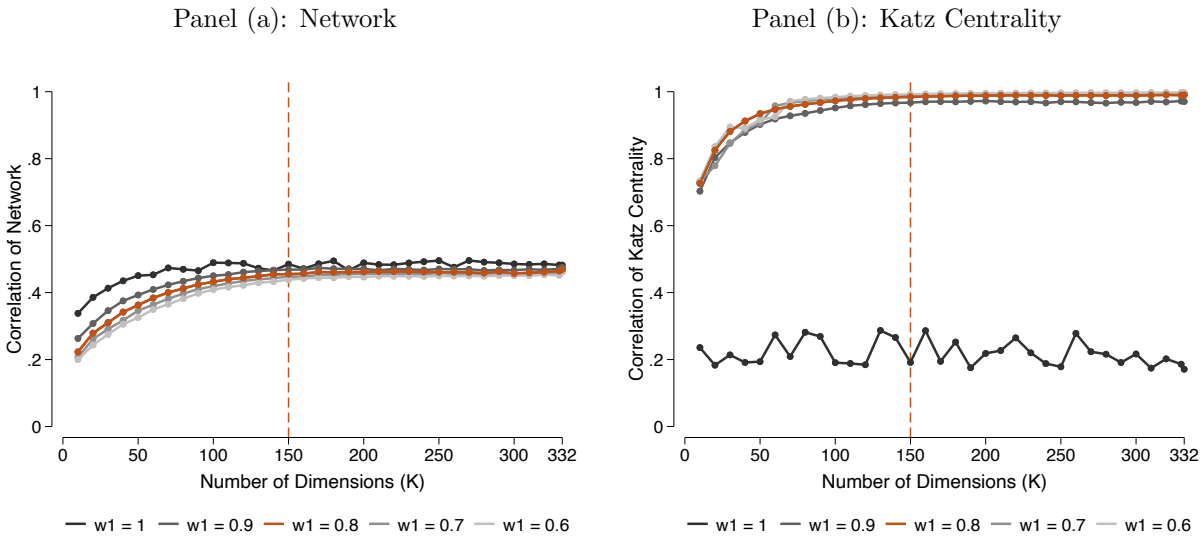
Figure A.5. Model Performance and Estimated Periods of Long-Run Dynamics

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



Notes. This figure shows the model performance and estimated periods of long-run dynamics in unobserved components time series model. Panel (a) shows the R^2 of the model, and panel (b) shows the estimated periods of the model.

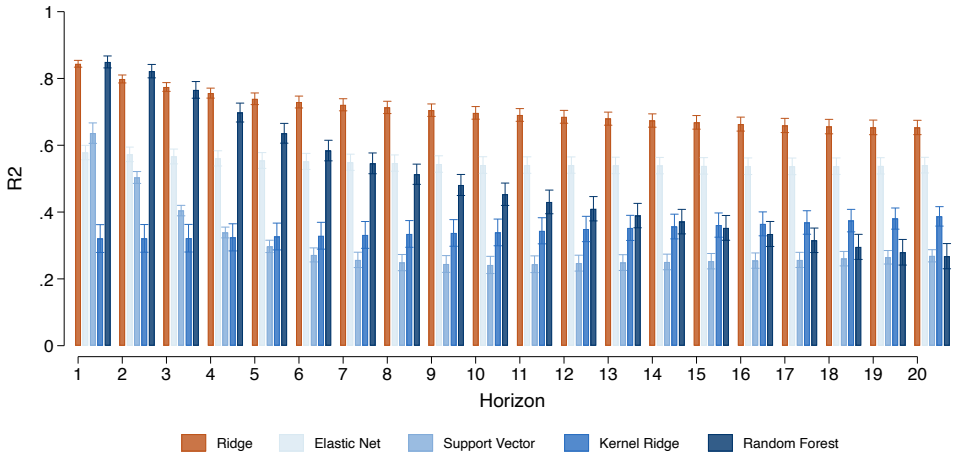
Figure A.6. Parameter Selection of Joint Diagonalization



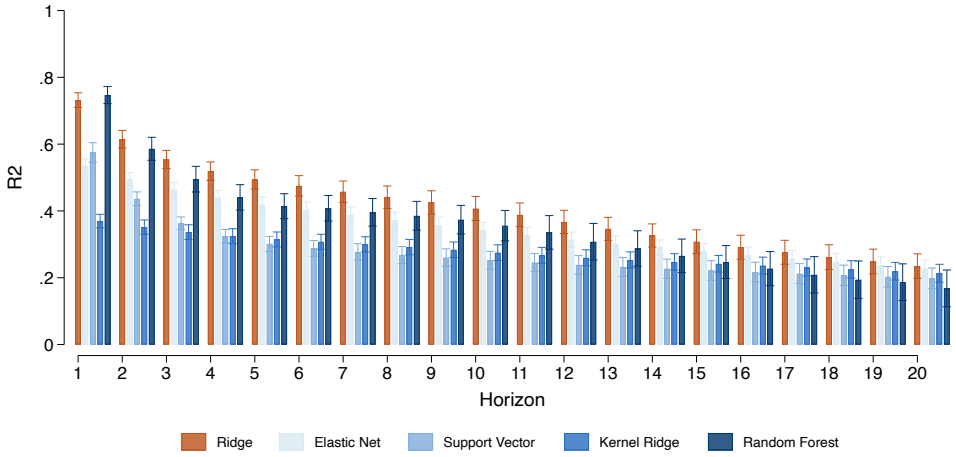
Notes. This figure shows the parameter selection of joint diagonalization. Panel (a) shows the network, and panel (b) shows the Katz centrality.

Figure A.7. Predictability of Eigenvalues: Model Selection

Panel (a): In-Sample R2



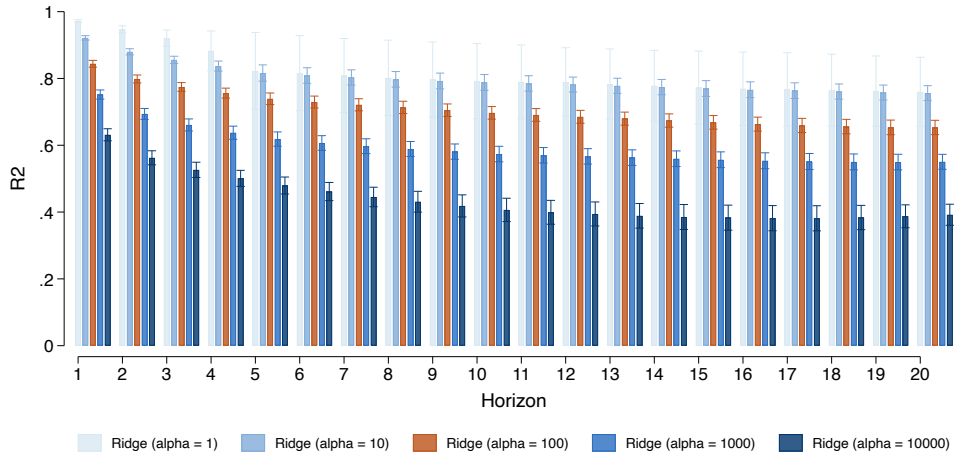
Panel (b): Out-of-Sample R2



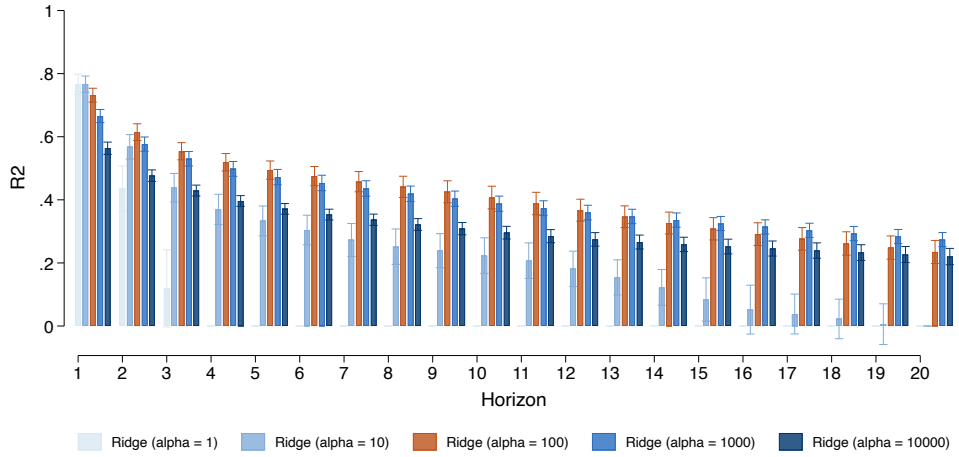
Notes. This figure shows the in-sample and out-sample R2 of the eigenvalue predictability using alternative models.

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

Panel (a): In-Sample R2



Panel (b): Out-of-Sample R2



Notes. This figure shows the in-sample and out-sample R2 of the eigenvalue predictability using alternative penalty in Ridge model.

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

Panel (a): Correlation of Innovation Network

	Scientific Field		Tecnnology Class	
Year	Pearson	Spearman	Pearson	Spearman
1940	0.19	0.10	0.08	0.07
1950	0.22	0.08	0.11	0.17
1960	0.31	0.19	0.09	0.15
1970	0.35	0.19	0.09	0.28
1980	0.35	0.21	0.21	0.14
1990	0.39	0.25	0.13	0.25
2000	0.35	0.24	0.22	0.17
2010	0.42	0.25	0.26	0.34
2020	0.50	0.25	0.46	0.49

Panel (b): Correlation of Innovation Centrality

	Scientific Field		Tecnnology Class	
Year	Pearson	Spearman	Pearson	Spearman
1940	0.64	0.64	0.02	0.06
1950	0.47	0.54	0.26	0.48
1960	0.52	0.74	0.51	0.57
1970	0.54	0.73	0.56	0.63
1980	0.57	0.59	0.36	0.53
1990	0.57	0.60	0.64	0.66
2000	0.31	0.47	0.74	0.58
2010	0.23	0.23	0.75	0.76
2020	0.23	0.29	0.80	0.78

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

Table A.2. List of Technological Innovation Episodes

Field Name	Emergence Year	Peak Year	Decline Year	Max Katz Centrality	Keywords
tech - locks, keys & safes	1870	1871	1874	0.0262	improved lock, padlock, sash, improved gate, improved hinge
tech - railways	1870	1871	1878	0.0185	improved car, improved car coupler, cab, railroad car, car coupler
tech - ships or other waterborne vessels	1870	1871	1874	0.0165	rudder, sail, improved anchor, steering apparatus, pontoon
tech - liquid positive-displacement machines	1870	1872	1875	0.0190	improved pump, blower, ejector, water
tech - skins, hides & pelts or leather	1878	1881	1883	0.0189	skiving, tanning, leather, hide, unhairing
tech - liquid machines or engines	1875	1882	1884	0.0201	windmill, horsepower, governor, wind engine, improved motor
tech - land vehicles for travelling	1894	1897	1899	0.0275	bicycle, velocipede, sulky, sleigh, thill
tech - hand tools	1897	1900	1903	0.0264	pipe wrench, wrench, vise, pliers, hammer
tech - agriculture	1896	1900	1904	0.0175	horseshoe, harrow, threshing, corn harvesting, mower
tech - writing or drawing implements	1898	1900	1903	0.0158	penholder, pencil, ink, plotter, eraser
tech - sewing	1898	1901	1903	0.0224	union, quilt, buttonhole, plait
tech - printing	1898	1901	1903	0.0204	typewriting machine, writing machine, typewriting, printing machine, typewriter
tech - weaving	1899	1902	1904	0.0231	mechanical loom, sett, Draper, temple, carpet
tech - sports & games	1899	1902	1904	0.0169	puzzle, skate, toy, game apparatus, game
tech - railways	1901	1904	1908	0.0181	truck, car truck, railway car, locomotive, railway signal
tech - general machines & steam engines	1903	1905	1907	0.0156	turbine, rotary engine, valve gear, water motor, muffler
tech - vehicles in general	1905	1912	1919	0.0234	trolley, tire, car fender, electric railway, caster
tech - combustion engines	1909	1915	1925	0.0505	carburetor, carburetor, explosive engine, explosive, gas engine
tech - land vehicles for travelling	1915	1918	1921	0.0277	tractor, sleigh, perambulator, bicycle, motorcycle
sci - data science	1916	1918	1920	0.0196	war, medico-legal, plague, tuberculosis, insurance
tech - vehicles in general	1923	1925	1928	0.0311	bumper, tractor, windshield, fob, tire
tech - lighting	1912	1925	1927	0.0195	lantern, headlight, incandescent, chandelier, candlestick
tech - brushware	1921	1928	1930	0.0170	toothbrush, broom, paintbrush, duster, cleaner
tech - aircraft, aviation & cosmonautics	1933	1938	1941	0.0204	aeroplane, airship, aircraft, parachute, propeller
sci - food science	1931	1938	1946	0.0152	milk, supplement, vitamin, sugar, pasteurization
tech - petroleum, gas or coke industries	1931	1942	1945	0.0153	gasification, hydrocarbon treatment, hydrocarbon oil, distillation, motor fuel
sci - world wide web	1944	1945	1950	0.0183	obituary, memorandum, correction, masthead, advertising index
sci - library science	1939	1945	1948	0.0173	bibliography, masthead, editorial, technology, appointment
sci - microbiology	1944	1946	1948	0.0593	penicillin, dysentery, bacteriology, bacilli, bacillus
sci - information retrieval	1946	1949	1958	0.0165	bibliography, patent, public received, appointment, item
sci - endocrinology	1950	1952	1955	0.0340	cortisone, ACTH, diabetes, vitamin, thiouracil
sci - pathology	1950	1952	1954	0.0253	epidemiology, tuberculosis, silicosis, pathology, osteopathy
sci - internal medicine	1950	1952	1955	0.0247	cortisone, ACTH, thiouracil, vitamin, penicillin
sci - intensive care medicine	1953	1955	1959	0.0151	penicillin, antibiotic, streptomycin, pneumonia, tuberculosis
sci - radiochemistry	1957	1960	1962	0.0234	radium, radioactive, isotope, plutonium, thorium
sci - nuclear physics	1954	1961	1976	0.0199	meson, fission, atom bomb, bomb, betatron
sci - particle physics	1959	1962	1970	0.0216	meson, parity, pion, decay, neutrino
sci - pathology	1961	1964	1968	0.0207	tuberculosis, epidemiology, pathology, pulmonary tuberculosis, cytology
sci - combinatorial chemistry	1965	1968	1969	0.0153	phenazine, pyridone, heterocycle, reagent, steroid
sci - radiochemistry	1966	1969	1975	0.0254	radioactive, fallout, radiochemistry, plutonium, determination

Field Name	Emergence Year	Peak Year	Decline Year	Max Katz Centrality	Keywords
sci - immunology	1967	1970	1977	0.0217	immunoglobulin, malaria, immunology, allergy, lymphocyte
tech - organic chemistry	1970	1972	1975	0.0169	manufacturing process, preparation process, steroid, phenyl, compound
sci - information retrieval	1972	1973	1976	0.0303	classification, miscellanea, synthesis, feedback, abstract
sci - library science	1971	1973	1975	0.0303	bibliography, subject index, select bibliography, classification, abstract
sci - world wide web	1965	1973	1974	0.0184	author index, subject index, literature, directory, bibliography
sci - stereochemistry	1964	1973	1976	0.0169	alkaloid, terpenoid, stereochemistry, conformation, absolute configuration
sci - family medicine	1973	1976	1977	0.0152	physician, editorial, pharmacy, book received, doctor
sci - chromatography	1978	1982	1985	0.0172	HPLC, chromatography, reversed-phase, mass spectrometry, erratum
sci - thermodynamics	1978	1984	1986	0.0198	thermodynamics, Bernoulli, boiling, heat transfer, equilibrium
sci - library science	1981	1985	1988	0.0315	bibliography, subject index, newsletter, conference, editorial board
sci - molecular biology	1981	1986	1998	0.0286	monoclonal, monoclonal antibody, antibody, clone, RNA
sci - information retrieval	1984	1993	1994	0.0306	Cheminform, announcements, notes, calendar, editorial
sci - polymer chemistry	1987	1993	1995	0.0186	polymerization, complex, copolymer, methacrylate, graft
sci - cell biology	1991	1996	2000	0.0226	embryo, kinase, transcription, cytokine, growth factor
sci - genetics	1988	1996	1999	0.0163	mutation, RFLP, plasmid, PCR, cDNA
sci - computer network	1994	1997	1999	0.0179	ATM, queue, ATM network, ISDN, multimedia
sci - multimedia	1994	1997	2001	0.0162	multimedia, hypermedia, session, interactive video, television
sci - internet privacy	1997	1998	2002	0.0224	internet, cyberspace, confidentiality, secret, email
sci - computational biology	1996	2000	2002	0.0170	genome, superfamily, biochemistry, clone, conference
sci - cancer research	1995	2000	2002	0.0162	melanoma, apoptosis, oncogene, mutation, leukemia
sci - world wide web	2000	2002	2003	0.0184	internet, website, intranet, wide web, world wide web
tech - electric communication technique	2000	2002	2005	0.0181	packet, ATM, internet, picture, video signal
sci - nanotechnology	2001	2005	2016	0.0457	nanotube, carbon nanotube, nanoparticle, sputtering
sci - computational chemistry	2001	2005	2006	0.0177	ab initio, density functional, ab initio study, sigmatropic rearrangement, fullerene
sci - organic chemistry	2002	2005	2006	0.0159	silyl, catalysis, alkene, aromatic, zeolite
tech - sports & games	2001	2005	2009	0.0158	game, skate, golf, game machine, club
sci - chemical engineering	2003	2006	2009	0.0283	nanotube, nanoparticle, coal, carbon nanotube, nano
sci - computer network	2004	2006	2008	0.0173	ATM, ATM network, ad hoc, multicast, QoS
sci - data science	2009	2010	2013	0.0179	data mining, proteomics, informatics, ontology, bioinformatics
sci - internet privacy	2010	2012	2014	0.0206	internet, privacy, website, email, HIPAA
sci - distributed computing	2011	2015	2024	0.0153	cloud, cloud computing, sensor network, peer, multicast
sci - aeronautics	2015	2017	2019	0.0274	NASA, aviation, aircraft, flight, crew
sci - embedded system	2016	2018	2021	0.0153	wearable, IoT, Arduino, Nexus, Android
tech - computing & calculating or counting	2018	2020	2024	0.0348	neural network, neural, machine learning, convolution, fingerprint
tech - measuring & testing	2018	2020	2024	0.0183	unmanned, radar, robot, detection method, fault
tech - electric communication technique	2017	2020	2021	0.0168	IoT, blockchain, cloud, mobile terminal, base station
tech - nanotechnology	2013	2022	2024	0.0157	graphene, nanotube, carbon nanotube, nano, nanoparticle
sci - data science	2017	2023	2024	0.0176	research guide, LibGuides, big data, COVID, data mining
sci - machine learning	2019	2024	2024	0.0247	machine learning, deep learning, convolution, neural, Markov

Notes. This table lists the technological innovation episode with its emergence year, peak year, decline year, maximum Katz centrality, and keywords.

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
10	0.0078	information retrieval	1	0.0176	skins, hides & pelts or leather
11	0.0074	environmental chemistry	2	0.0136	physical or chemical apparatus
12	0.0073	library science	3	0.0106	sewing
18	0.0066	thermodynamics	4	0.0103	horology
32	0.0052	aeronautics	5	0.0101	liquid machines or engines
34	0.0051	artificial intelligence	6	0.0095	cements
38	0.0050	parallel computing	7	0.0089	combustion engines
40	0.0049	data science	8	0.0087	steam generation
45	0.0045	optics	9	0.0081	dyes, paints, polishes, natural resins & adhesives
50	0.0043	meteorology	13	0.0073	land vehicles for travelling

Panel (b): Top 10 Fields in 1900

Scientific Fields			Technological Classes		
Rank	Katz Centrality	Field Name	Rank	Katz Centrality	Field Name
14	0.0093	family medicine	1	0.0264	hand tools
16	0.0091	virology	2	0.0222	sewing
21	0.0078	pathology	3	0.0199	weaving
26	0.0070	data science	4	0.0196	printing
31	0.0065	library science	5	0.0175	agriculture
46	0.0051	world wide web	6	0.0158	writing or drawing implements
56	0.0048	gynecology	7	0.0138	sports & games
57	0.0047	information retrieval	8	0.0133	measuring & testing
64	0.0045	veterinary medicine	9	0.0131	steam generation
68	0.0040	arithmetic	10	0.0118	horology

Panel (c): Top 10 Fields in 1920

Scientific Fields			Technological Classes		
Rank	Katz Centrality	Field Name	Rank	Katz Centrality	Field Name
5	0.0102	library science	1	0.0178	land vehicles for travelling
7	0.0099	data science	2	0.0154	combustion engines
10	0.0080	family medicine	3	0.0132	printing
12	0.0075	aeronautics	4	0.0105	heating & ranges & ventilating
14	0.0071	information retrieval	6	0.0099	other electric techniques
20	0.0064	astrophysics	8	0.0087	aircraft, aviation & cosmonautics
21	0.0063	internet privacy	9	0.0083	vehicles in general
24	0.0058	optometry	11	0.0076	combustion apparatus & combustion processes
35	0.0051	stereochemistry	13	0.0072	general machines & steam engines
37	0.0049	climatology	15	0.0068	optics

Panel (d): Top 10 Fields in 1940

Scientific Fields			Technological Classes		
Rank	Katz Centrality	Field Name	Rank	Katz Centrality	Field Name
2	0.0128	library science	1	0.0157	aircraft, aviation & cosmonautics
3	0.0103	archaeology	8	0.0079	combustion apparatus & combustion processes
4	0.0097	endocrinology	13	0.0072	dyes, paints, polishes, natural resins & adhesives
5	0.0091	internal medicine	14	0.0070	vehicles in general
6	0.0091	food science	16	0.0068	petroleum, gas or coke industries
7	0.0083	operating system	18	0.0065	lighting
9	0.0076	pathology	19	0.0063	conveying, packing & storing
10	0.0075	data science	22	0.0061	furniture
11	0.0075	physiology	23	0.0060	life-saving & fire-fighting
12	0.0073	biochemistry	24	0.0060	electric communication technique

Panel (e): Top 10 Fields in 1960

Scientific Fields			Technological Classes		
Rank	Katz Centrality	Field Name	Rank	Katz Centrality	Field Name
1	0.0234	radiochemistry	13	0.0078	organic chemistry
2	0.0175	nuclear physics	16	0.0072	combustion engines
3	0.0142	atomic physics	25	0.0063	biochemistry, beer, microbiology & enzymology
4	0.0121	virology	27	0.0062	agriculture
5	0.0108	microbiology	28	0.0061	organic macromolecular compounds
6	0.0100	pediatrics	34	0.0056	nuclear physics
7	0.0099	biochemistry	38	0.0052	measuring & testing
8	0.0090	nuclear engineering	40	0.0049	combustion apparatus & combustion processes
9	0.0086	metallurgy	45	0.0043	textiles
10	0.0081	pathology	46	0.0043	basic electric elements

Panel (f): Top 10 Fields in 1980

Scientific Fields			Technological Classes		
Rank	Katz Centrality	Field Name	Rank	Katz Centrality	Field Name
1	0.0160	chromatography	9	0.0080	conveying, packing & storing
2	0.0159	library science	17	0.0068	hand tools
3	0.0121	thermodynamics	22	0.0064	combustion apparatus & combustion processes
4	0.0120	molecular biology	34	0.0053	fertilisers
5	0.0109	information retrieval	42	0.0049	measuring & testing
6	0.0102	biochemistry	54	0.0043	petroleum, gas or coke industries
7	0.0091	radiology	69	0.0040	electric power
8	0.0083	waste management	83	0.0037	agriculture
10	0.0075	surgery	89	0.0036	combustion engines
11	0.0075	pathology	97	0.0034	machine tools

Panel (g): Top 10 Fields in 2000

Scientific Fields			Technological Classes		
Rank	Katz Centrality	Field Name	Rank	Katz Centrality	Field Name
1	0.0208	internet privacy	8	0.0115	computing & calculating or counting
2	0.0170	computational biology	10	0.0107	electric communication technique
3	0.0162	cancer research	12	0.0090	basic electric elements
4	0.0151	nanotechnology	15	0.0079	grinding & polishing
5	0.0133	cell biology	24	0.0074	photography
6	0.0133	optoelectronics	42	0.0052	printing
7	0.0118	multimedia	50	0.0046	biochemistry, beer, microbiology & enzymology
9	0.0111	molecular biology	59	0.0044	optics
11	0.0091	family medicine	62	0.0043	medical or veterinary science
13	0.0082	chemical engineering	74	0.0037	checking-devices

Panel (h): Top 10 Fields in 2020

Scientific Fields			Technological Classes		
Rank	Katz Centrality	Field Name	Rank	Katz Centrality	Field Name
2	0.0185	machine learning	1	0.0348	computing & calculating or counting
7	0.0107	mechanical engineering	3	0.0183	measuring & testing
9	0.0104	computer security	4	0.0168	electric communication technique
10	0.0096	embedded system	5	0.0144	musical instruments & acoustics
15	0.0090	genetics	6	0.0132	controlling & regulating
18	0.0089	nanotechnology	8	0.0105	physical or chemical apparatus
19	0.0089	nuclear chemistry	11	0.0095	signalling
21	0.0088	distributed computing	12	0.0091	nanotechnology
25	0.0082	theoretical computer science	13	0.0091	machine tools
26	0.0080	data science	14	0.0091	conveying, packing & storing

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