Garg, Prashant; Fetzer, Thiemo

Working Paper

Causal Claims in Economics

I4R Discussion Paper Series, No. 183

Provided in Cooperation with:
The Institute for Replication (I4R)

Suggested Citation: Garg, Prashant; Fetzer, Thiemo (2024) : Causal Claims in Economics, I4R Discussion Paper Series, No. 183, Institute for Replication (I4R), s.l.

This Version is available at:
https://hdl.handle.net/10419/306280

INSTITUTE for REPLICATION

No. 183
I4R DISCUSSION PAPER SERIES

Causal Claims in Economics

Prashant Garg

Thiemo Fetzer

November 2024

INSTITUTE for REPLICATION

I4R DISCUSSION PAPER SERIES

I4R DP No. 183

Causal Claims in Economics

Prashant Garg[1], Thiemo Fetzer[1, 2, 3, 4, 5, 6]

[1]Imperial College London/Great Britain
[2]University of Warwick, Coventry/Great Britain
[3]University of Bonn/Germany
[4]Centre for Economic Policy Research (CEPR), London/Great Britian
[5]National Institute of Economic and Social Research (NIESR), London/Great Britain
[6]ECONtribute, Bonn/Germany

NOVEMBER 2024

E-Mail: joerg.peters@rwi-essen.de
RWI – Leibniz Institute for Economic Research
Hohenzollernstraße 1-3
45128 Essen/Germany
www.i4replication.org

ISSN: 2752-1931

Causal Claims in Economics*

Prashant Garg Thiemo Fetzer †

November 5, 2024

Click here for the latest version

Abstract

We analyze over 44,000 economics working papers from 1980–2023 using a custom language model to construct knowledge graphs mapping economic concepts and their relationships, distinguishing between general claims and those supported by causal inference methods. The share of causal claims within papers rose from about 4% in 1990 to 28% in 2020, reflecting the "credibility revolution." Our findings reveal a trade-off between factors enhancing publication in top journals and those driving citation impact. While employing causal inference methods, introducing novel causal relationships, and engaging with less central, specialized concepts increase the likelihood of publication in top 5 journals, these features do not necessarily lead to higher citation counts. Instead, papers focusing on central concepts tend to receive more citations once published. However, papers with intricate, interconnected causal narratives—measured by the complexity and depth of causal channels—are more likely to be both published in top journals and receive more citations. Finally, we observe a decline in reporting null results and increased use of private data, which may hinder transparency and replicability of economics research, highlighting the need for research practices that enhance both credibility and accessibility.

Keywords: KNOWLEDGE GRAPH, CREDIBILITY REVOLUTION, CAUSAL INFERENCE, NARRATIVE COMPLEXITY, NULL RESULTS, PRIVATE DATA, LARGE LANGUAGE MODELS
JEL Classification: A10, B41, C18, C80, D83

*We thank Samuel Kortum, Peter J. Lambert, Andrew Oswald, Carol Propper and Tommaso Valletti for helpful comments. We thank Lakshya Kavya for excellent research assistance. The authors declare no competing interests.

†Garg is based at Imperial College London. Email: prashant.garg@imperial.ac.uk. Fetzer is based at University of Warwick & Bonn and affiliated with CEPR, CAGE, NIESR, ECONtribute, Grantham Institute. Fetzer acknowledges funding by the Leverhulme Prize in Economics, a European Research Council Starting Grant (ERC, MEGEO, 101042703), and Deutsche Forschungsgemeinschaft (DFG, EXC 2126/1 – 390838866). This paper is part of the https://www.causal.claims/ project. Email: team@causal.claims

1 Introduction

Economic research has undergone a significant transformation over the past few decades, marked by an increased emphasis on establishing causal relationships through empirical methods. This "credibility revolution" has propelled the discipline toward more rigorous identification strategies, aiming to provide robust evidence for policy-making and theoretical advancement (Angrist & Pischke 2010). Pioneering work by researchers such as Orley Ashenfelter, Joshua Angrist, David Card, Guido Imbens and Alan Krueger introduced methodologies that enhanced causal identification, including natural experiments, regression discontinuity designs (RDDs), and instrumental variables (IVs) (Angrist & Imbens 1994, Imbens & Rubin 2015). These approaches address endogeneity concerns and provide more credible estimates of causal effects.

Leading journals now prioritize studies employing these methods over traditional correlational approaches (Card & DellaVigna 2013, Hamermesh 2013). Hamermesh (2013), for instance, documents a decline in purely theoretical articles and an increase in empirical studies utilizing self-generated data and experimental methods in top economics journals.[1] The credibility revolution has raised the bar for empirical work, emphasizing transparent reporting, careful consideration of identification assumptions, and rigorous sensitivity analyses (Angrist & Pischke 2008, Imbens & Rubin 2015). However, this focus on specific methodologies has sparked debates about the direction and priorities of economic research.[2] Studies like Currie et al. (2020) have also shown a significant rise in empirical methods across economics, supported by advancements in data and technology. Our study further contributes by breaking down these trends into specific causal inference methods and examining their usage across subfields.

Despite extensive discussions on methodological advancements, there is a lack of comprehensive analysis that quantifies how the structure and complexity of economic research have

[1] Hamermesh (2013) analyzes full-length articles published in the *AER*, *JPE*, and *QJE* from 1963 to 2011, highlighting shifts in authorship patterns, age structures, and methodological approaches.

[2] Some scholars argue that the intense focus on methodological innovation may overshadow genuine novelty in research questions. There's a perception that we might be answering the same questions repeatedly, merely applying different methods, which could suggest diminishing returns on new insights. This raises concerns that genuine novelty is becoming scarce and that "framing" and active "salesmanship" are increasingly important for dissemination. A global survey by Andre & Falk (2021) reveals that many economists believe research should become more policy-relevant, multidisciplinary, and disruptive, pursuing more diverse topics to address pressing societal issues.

evolved over time, particularly regarding the use of causal claims and narrative complexity across subfields. Our study addresses this gap by analyzing over 44,000 NBER and CEPR working papers using a custom large language model to extract structured information on the knowledge graphs of papers, including the methods used to evidence each claim—causal or otherwise—and the data employed in the analyses.

We introduce a novel approach by constructing a *knowledge graph* for each paper in our dataset. In these graphs, nodes represent economic concepts classified using JEL codes, and edges represent relationships from a *source* node to a *sink* node. This means that if a paper discusses how one economic concept relates to another, we capture this as a directional link between those concepts. Whether or not a claim is considered causal depends on the method used to substantiate it. Specifically, we identify an edge as a *causal edge* if the claim is evidenced using causal inference methods such as Difference-in-Differences (DiD), Instrumental Variables (IV), Randomized Controlled Trials (RCTs), Regression Discontinuity Designs (RDDs), Event Studies, or Synthetic Control

This graphical representation allows us to quantitatively assess the complexity and structure of narratives in economic research over time. We develop several key measures derived from these knowledge graphs, capturing different dimensions of narrative complexity, originality, and engagement with central or peripheral concepts in the field.

First, the number of edges represents the total number of relationships (edges) discussed in a paper's knowledge graph, reflecting the breadth of the narrative. We compute this measure for both the full knowledge graph and the causal subgraph. Second, the number of unique paths indicates the number of distinct pathways from source nodes to sink nodes, showing the variety of channels through which relationships occur; a higher number suggests more interconnected narratives with multiple mechanisms at play. Third, the longest path length measures the length of the longest chain of connected concepts, representing the depth of reasoning or the extent of argumentation in a paper. Fourth, the source-sink ratio captures the balance between the number of source nodes (originating concepts) and sink nodes (receiving concepts), indicating whether a paper focuses more on exploring various causes leading to a few effects or vice

versa. Fifth, the proportion of novel edges is the share of relationships in a paper that are not previously documented in the literature, capturing the originality of the research. Finally, the average eigenvector centrality measures how central the concepts in a paper are within the overall network of economic knowledge; concepts with higher centrality are more influential or connected within the field.

By examining both the full knowledge graphs (*All*) and the causal subgraphs (*Causal*), we can differentiate between general narrative complexity and the complexity specific to causal claims.

Our analysis reveals several notable patterns. First, the use of causal inference methods has significantly increased over time, with the average proportion of causal claims rising from about 4% in 1990 to nearly 28% in 2020, reflecting the impact of the credibility revolution in economics.

Second, we find a trade-off between factors that enhance publication in top journals and those that drive citation impact. Specifically, employing causal inference methods, introducing novel causal relationships, and engaging with less central, specialized concepts increase the likelihood of publication in top 5 journals. However, these features do not necessarily lead to higher citation counts once published. Instead, papers focusing on central, widely recognized concepts tend to receive more citations, highlighting a divergence between publication success and broader academic influence.

Third, narrative complexity—measured by the number of unique paths and the longest path length in the causal subgraph—is positively associated with both publication in top journals and increased citation counts, especially in top 5 and top 6–20 journals. This suggests that depth and complexity in causal narratives are valued both for publication success and academic impact.

These findings highlight the trade-off between methodological rigor, narrative structure, topic centrality, and the dissemination and recognition of research within economics. They suggest that while top journals favor innovative and methodologically rigorous research, broader academic impact may depend more on engaging with central, widely recognized topics. This divergence raises important questions about the direction and priorities of economic research.

Critics argue that the emphasis on specific empirical methods and complex narratives may

lead to overconfidence in results and potential overinterpretation, especially if the underlying assumptions are not fully met (Deaton 2010, Keane 2010). Additionally, the focus on identification sometimes comes at the expense of economic theory, resulting in studies that establish causal effects without adequately explaining the underlying mechanisms (Sims 2010, Heckman 2001). As Keane (2010) notes, the detachment from theoretical frameworks can limit the explanatory power of empirical findings. This underscores that while methodological rigor is essential for credible causal inference, it should not preclude the consideration of valuable evidence from diverse sources. Misapplication or overinterpretation of methods can lead to questionable conclusions. For instance, the use of instrumental variables relies on strong assumptions that the instrument affects the outcome only through the endogenous explanatory variable and is uncorrelated with the error term (Angrist & Imbens 1994). Violations of these assumptions, such as weak instruments or invalid exclusion restrictions, can produce biased estimates (Rosenbaum & Rosenbaum 2002). For example, Mellon (2021) highlights the challenges of using weather variables as instruments, identifying numerous potential exclusion-restriction violations.

RCTs have gained prominence as a gold standard for causal inference due to their internal validity. However, scholars like Deaton & Cartwright (2018) and Cartwright (2007) argue that RCTs may suffer from limited external validity and may not capture complex economic phenomena, and can further be subject to inducing demand effects (de Quidt et al. 2018). Generalizing findings from specific experimental settings without considering contextual differences can lead to misleading conclusions (Ravallion 2009).

Moreover, the complexity of research outputs has increased as papers have become longer and include more coauthors (Card & DellaVigna 2013). This expansion reflects the need to address methodological rigor and to include detailed explanations of causal mechanisms, robustness checks, and theoretical integration. However, this rise in narrative complexity may also indicate that the presentation and promotion of research findings are becoming more important factors in dissemination.[3] Increased complexity can make it challenging for readers

[3]The emphasis on elaborate narratives and framing may encourage researchers to "oversell" their findings, potentially at the expense of clarity or transparency. This dynamic may lead to research being presented in a way that separates evidence from the story, which can be particularly problematic when findings are translated into media narratives.

and reviewers to critically assess the validity of the claims (Ioannidis 2005, Gelman & Loken 2014), and may contribute to overemphasizing research findings. The "garden of forking paths" metaphor illustrates how analytical flexibility can lead to false-positive findings even without intentional misconduct (Gelman & Loken 2014).[4] The American Statistical Association has highlighted the misinterpretation and misuse of p-values, advocating for a more nuanced understanding of statistical significance (Wasserstein & Lazar 2016). Simonsohn et al. (2014) propose the p-curve method as a tool to detect and correct for publication bias using only significant results, highlighting the pervasiveness of selective reporting.

Issues like p-hacking, publication bias, and the underreporting of null results further contribute to concerns about empirical research. Researchers may inadvertently engage in p-hacking by exploring various model specifications and reporting only those that yield significant results (Simmons et al. 2011). Publication bias, often referred to as the "file drawer problem," results from the tendency of journals to favor significant and novel results over null or replication studies (Rosenthal 1979, Sterling 1959). This creates a market where null results are undervalued, leading to a skewed literature that overrepresents positive findings and may affect the overall credibility of research. Brodeur et al. (2016) document how this bias leads to an overrepresentation of significant results in economics journals. Similarly, Chopra et al. (2024) find evidence for a substantial perceived penalty against null results, with researchers believing that studies reporting null findings have a lower chance of being published and are perceived as lower quality. Andrews & Kasy (2019) propose methods for identifying and correcting for publication bias, emphasizing the importance of accounting for selective publication in empirical research. Moreover, Frankel & Kasy (2022) discuss optimal publication rules given the scarcity of journal space, suggesting that journals should consider publishing findings that significantly shift prior beliefs, including precise null results.

Our analysis shows a decrease in the reporting of null results over time, from approximately 15% in 1980 to around 8.6% in 2023. This decline may reflect increased perceived professional

[4]The "garden of forking paths" refers to the many choices a researcher can make in data analysis, which can lead to a multitude of possible results. Without pre-specification, this can increase the likelihood of finding a significant result by chance.

norms to produce significant findings, possibly due to the publication process favoring positive results. The underreporting of null results can contribute to a skewed literature and may affect the overall credibility of research (Rosenthal 1979, Sterling 1959). Academic networks influence the dissemination and acceptance of research findings. Jackson (2010) and Newman (2003) discuss how social and citation networks shape academic discourse. The Matthew Effect, where established scholars receive disproportionate recognition, can amplify certain findings through increased citations and visibility (Merton 1968). Access to "clubs" or networks may thus be an important factor driving publication success, potentially overshadowing actual research quality. The emphasis on publishing in top journals further reinforces this effect, as noted by Heckman & Moktan (2020), who argue that reliance on top journals as a screening mechanism may not reliably identify the most creative or impactful research.[5]

Data availability is crucial for replication and verification of research findings. We observe a significant rise in the use of private or proprietary data sources, with the proportion of papers using private company data increasing from approximately 4% in 1980 to around 8.6% in 2023. In recent literature (post-2000), we find that fields with the highest use of private data are Behavioral Economics (14.0%), Finance (13.7%), and Industrial Organization (13.6%). While such data enhances analysis, it raises concerns about data accessibility, replicability, and transparency (Andreoli-Versbach & Mueller-Langer 2014). Private data companies could be strategic about whom and what data to provide access to the research community, potentially indirectly inducing it to shape narratives (Delbono et al. 2024).[6] Barrios et al. (2024) find that the use of private data significantly reduces trust in economics research among both economists and the general public. Their study shows that papers with conflicts of interest, such as reliance on proprietary data, are perceived as less credible, undermining the perceived value of the research. Best practices on replication are often still only an aspiration rather than a reality, partly due to the use of proprietary data. Data privacy regulations, such as the GDPR, further complicate

[5]The competitive nature of academic and the importance of high stakes publications for career progression may encourage behaviors that may be akin to overselling and encourage aggressive "salesmanship" that may come at the expense of substance and rigor.

[6]Further concerns arise if data companies extend research access to a narrow group of high profile academics as as part of their corporate social responsibility strategy.

data accessibility by imposing restrictions on the use and sharing of personal data (Fetzer 2022). This tension highlights the need for policies that balance privacy concerns with the benefits of data accessibility for scientific advancement.[7] In response to these challenges, Miguel (2021) documents the adoption of open science practices in economics, such as pre-registration and data sharing, noting a rapid transition toward increased transparency.

The relationship between theory and empirics is a central concern in economics. Critics argue that the focus on empirical identification has led to a neglect of theoretical development (Keane 2010, Heckman 2001). Without a solid theoretical foundation, empirical findings may lack coherence. Sims (2010) emphasizes that economics is not purely an experimental science and that theoretical models are essential for interpreting empirical results. Furthermore, Andre & Falk (2021) highlight that economists see value in research that is multidisciplinary and addresses diverse topics, suggesting a need to balance empirical rigor with theoretical and interdisciplinary approaches.

Emerging methodologies, such as machine learning and Bayesian inference, offer new tools for causal analysis. Machine learning techniques can handle high-dimensional data and uncover complex patterns but pose risks of overfitting and require careful interpretation (Athey & Imbens 2019, Chernozhukov et al. 2018). Bayesian methods provide a framework for incorporating prior information and uncertainty into causal inference (Rubin 1984). However, rapid technical progress can lead to a lack of quality training even among established researchers, making it challenging to keep pace with best practices.

Ethical considerations in empirical research extend beyond methodological rigor to include transparency about limitations, uncertainties, and the broader context of findings (Resnik 1998). Misleading claims can distort policy-making, erode public trust, and lead to poor allocation of resources. Ensuring integrity in research is a collective responsibility involving researchers, journals, institutions, and funding bodies. The media also plays a critical role in disseminating research findings, yet it can inadvertently perpetuate flawed research. Alabrese (2022) demonstrate that even retracted studies can continue to misinform the public if they receive significant media attention prior to their retraction, highlighting the shared responsibility of researchers

[7] For a discussion on navigating data privacy in research, see World Bank (2020).

and journalists in maintaining accuracy in scientific communication.

Moreover, the ways in which academics engage with the public—particularly through social media—can shape the perceived credibility of their work. Garg & Fetzer (2024) document systematic patterns in how academics discuss politically salient topics such as climate change, culture, and economics on platforms like Twitter. These expressions often diverge from public opinion in both focus and tone, influencing public perceptions of academia in ways that may not represent the broader academic consensus. Since only a subset of academics are active on social media, this may skew public understanding of academic priorities. Furthermore, Alabrese et al. (2024) find that academics who express strong political views online are often seen as less credible by the public, raising concerns about how personal political expression may affect trust in scientific research. These insights highlight the importance of responsible communication in preserving the integrity of both academic discourse and public trust in science.

These concerns gain even more urgency in light of recent global initiatives emphasizing the importance of evidence-based policy-making. International leaders and organizations have expressed alarm over the slow progress in achieving Sustainable Development Goals (SDGs), attributing part of the challenge to a lack of robust evidence to inform policy decisions. Despite substantial investments in public services, a "hidden" repository of underutilized studies exists that could inform better policy choices (Nature Editorial 2024). Recognizing these challenges, research councils and governments are investing in innovative solutions to enhance the accessibility and synthesis of existing research. For instance, in September 2024, the UK Economic and Social Research Council (ESRC) announced a significant investment in artificial intelligence to facilitate evidence synthesis for public policy, aiming to build a global infrastructure that provides useful evidence for policymakers.[8] Moreover, organizations like the Behavioural Insights Team have proposed blueprints for better international collaboration on evidence, emphasizing the importance of evidence synthesis and accessibility (Behavioural Insights Team 2024).[9]

[8]See the broader funding call at https://www.ukri.org/opportunity/transforming-global-evidence-ai-driven-evidence-synthesis-for-policymaking/

[9]See Behavioural Insights Team (2024) for detailed recommendations on international collaboration to improve evidence synthesis. Available at https://www.bi.team/publications/

Understanding these trends requires examining how economic research evolves over time. Angrist et al. (2017) analyze a large dataset of economics journal articles from 1980 to 2015, documenting shifts in research fields and styles. They find that the growth in empirical work reflects substantial shifts within fields rather than across them, with more empirical papers appearing in influential journals and receiving more citations. This evolution underscores the importance of methodological advancements and their impact on the discipline's focus, but also raises questions about whether new ideas are becoming harder to come by (Bloom et al. 2020, Park et al. 2023).

Our study contributes to this discussion by providing empirical evidence on the evolution of empirical methods and their differential adoption across subfields. We observe that methods such as DiD, IV, RCTs, and RDDs have seen substantial growth, reflecting the discipline's shift towards more rigorous identification strategies. Fields such as Urban, Health, Development, and Behavioral exhibit the most significant increases in the use of causal inference methods. Conversely, fields like Macroeconomics show more modest increases. This variation underscores how research questions, data availability, and methodological traditions influence the adoption of empirical methods across different areas of economics.

By constructing and analyzing the knowledge graphs of economic research, we offer a novel perspective on how the complexity and structure of narratives have changed over time and how they influence the dissemination and recognition of research findings. Our findings suggest that certain structural features and methodological choices in research are associated with successful publication outcomes, highlighting the evolving landscape of economic research in the era of the credibility revolution.

The remainder of this paper is organized as follows: Section 2 details the data and information retrieval methods used to extract and analyze the knowledge graph of papers. Section 3 introduces the knowledge graph of economics, discussing the measures of narrative complexity, the evolution of empirical methods, and their adoption across fields. Section 4 examines how the structure of a paper's knowledge graph relates to publication and citation outcomes. Section 5 explores challenges in replication and data accessibility, including the reporting of null results

international-collaboration-evidence/.

and the use of private data. Finally, Section 6 concludes with implications for research practices and the communication of economic knowledge.

2 Data and Methods

In this section, we present the data sources, extraction processes, and methods employed to examine causal claims within the economics literature. Further methodological details and technical specifications are provided in Appendix A.[10]

2.1 Working Paper Corpus

Our analysis is based on a comprehensive corpus of working papers from two primary sources: the National Bureau of Economic Research (NBER) and the Centre for Economic Policy Research (CEPR). The NBER dataset comprises 28,186 working papers, while the CEPR dataset includes 16,666 papers, resulting in a total sample of 44,852 papers. These papers span several decades and encompass various subfields of economics, providing a broad view of the research landscape.

To refine the sample and focus on relevant content, we applied specific filtering criteria. We included only papers containing more than 1,000 characters to exclude incomplete documents. Additionally, we limited the analysis to the first 30 pages of each paper, ensuring that we captured the sections most likely to contain causal claims, such as introductions, literature reviews, and empirical analyses.

The corpus covers a wide range of economics subfields, including Labour Economics, Public Economics, Macroeconomics, Development Economics, and Finance. The papers employ diverse empirical strategies, such as Randomized Controlled Trials (RCTs), Instrumental Variables (IV), Difference-in-Differences (DiD), and Regression Discontinuity Designs (RDD), allowing us to examine methodological trends across the discipline.

[10] Our paper is part of a larger project on causal claims. Please visit https://www.causal.claims/ to (i) access our curated dataset, (ii) interactively search and view causal graph of papers and authors, (iii) search for causal claims across economics literature using our *Causal Claims Research Assistant (CClaRA)* which is an AI chatbot fine-tuned on our underlying knowledge graph data.

Pre-processing The preprocessing of the text data followed a structured pipeline aimed at cleaning and normalizing the text for analysis. The preprocessing steps included removing excessive whitespace, converting all characters to lowercase, and filtering out non-alphanumeric characters, keeping only spaces for readability. Additionally, we stripped leading and trailing whitespace to ensure uniformity in the text. These steps were helpful for the large language model to efficiently and accurately process the text, especially considering the large volume of data involved.

2.2 LLM based retrieval

We employed a multi-stage process using a large language model (LLM) to extract and analyze information from the working papers in our corpus.[11] We interacted with the LLM using carefully designed prompts that guided the model to extract the required information while adhering to a predefined JSON schema. The overall process is visually summarized in Figure 1, which illustrates the flow from input text to structured data extraction and subsequent analysis. This approach allowed us to efficiently process the text and extract detailed structured data necessary for our analysis while minimizing computational and human resources.

(Figure 1)

Our LLM-based retrieval process consists of the following stages:

Stage 1: Qualitative Summary Extraction In the first stage, we prompted the LLM to analyze the first 30 pages of each paper and extract a curated summary of key elements. This included the research questions as presented in the abstract, introduction, and full text; information on causal identification strategies used in the paper; details on data usage, accessibility, and acknowledgements; and metadata such as authors' names, institutional affiliations, fields of

[11]We used GPT-4o-mini, developed by OpenAI, a state-of-the-art LLM renowned for its capacity to perform complex language understanding and generation tasks with high accuracy. The model is pre-trained on a diverse corpus that includes books, academic papers, websites, and other text sources, encompassing a wide range of topics and detailed descriptions of economic concepts. This extensive pre-training enables the model to have a nuanced understanding and human-level judgment in various contexts. See Fetzer et al. (2024) for an innovative application of LLMs to build the vertical input-output linkages between products.

study, and methods used. This initial extraction provided a structured overview of each paper, which was used in subsequent stages to extract more detailed information. To ensure the reliability of our extraction process, we validate our information retrieval methods in Appendix C, demonstrating high accuracy and F1 scores across key empirical methods and fields of study.

Stage 2: Extraction of Causal Claims Using the curated summaries from Stage 1, particularly the sections on causal identification and causal claims, we prompted the LLM to extract detailed all knowledge links presented in each paper between two knowledge entities. The LLM identified source and sink variables as described by the authors, determined the types of relationships (e.g., direct causal effect, indirect causal effect, mediation, confounding, theoretised relationship, correlation), and recorded the empirical methods used to establish each link (e.g., RCT, IV, DiD, OLS, simulations). The result was an edge list per paper, where each row represents an edge with a source node (e.g., a cause) and a sink node (e.g., an effect). We also included relevant edge attributes, such as the method used to evidence that edge. While we collected additional attributes like the direction of effect, magnitude, and statistical significance, these features were experimental and are not used in the main analysis due to variation in reporting standards.[12]

Stage 3: Data Usage and Accessibility Extraction From the data-related summaries in Stage 1, we prompted the LLM to extract structured information regarding data sources and accessibility. Key elements included the ownership of the data (e.g., private company, public sector entity, researchers), data accessibility (e.g., freely accessible, restricted), and details on data granularity, units of analysis, temporal and geographical context. This information is crucial for assessing trends in data usage and the implications for transparency and replicability in economic research.

Matching Variables to Standardized Economic Concepts To facilitate systematic analysis and aggregation, we standardized the free-text descriptions of the source and sink variables by mapping them to official Journal of Economic Literature (JEL) codes. We created semantic embeddings for each JEL code's overall description, which concatenates the JEL description,

[12]These additional attributes were collected for exploratory purposes but were not included in the primary analysis.

guidelines, and keywords.[13] By generating vector embeddings of the variable descriptions and comparing them to the embeddings of JEL code descriptions, we identified the most relevant codes for each variable.[14] This process situates each causal claim within the broader context of economic research and allows us to construct a knowledge graph of economic research, mapping and documenting the frontier in causal evidence over time. This process is visually summarized in Figure 2, which illustrates our AI-driven approach to analyzing and mapping causal linkages between JEL codes.

(Figure 2)

Validation To validate our information retrieval methods, we conducted two exercises (details in Appendix C). First, we matched 307 papers with the annotated dataset from Brodeur et al. (2024), which classifies empirical methods and fields for 1,106 economics papers. Our retrieval achieved high accuracy and F1 scores, especially for RDD methods and fields like Macroeconomics and Urban Economics, demonstrating reliability across key dimensions.

In a second exercise, we compared our causal claims data to the Plausibly Exogenous Galore dataset,[15] a source documenting primary causal variables and exogenous variation for 1,435 papers. We matched 485 papers and aggregated our data at the paper level to align with their structure. This comparison yielded moderate similarity for causes and effects, likely due to the Plausibly Exogenous Galore dataset's focus on the most important causal link, whereas our dataset captures the full knowledge graph of each paper, including all causal links. This broader scope introduces variability in matching specific causes and effects. However, the source of exogenous variation showed high similarity, supporting the consistency of our approach in capturing essential causal elements across datasets.

[13] The JEL guidelines (available at https://www.aeaweb.org/jel/guide/jel.php) provide detailed descriptions of each code and are typically a paragraph long. Including keywords enhances the semantic specificity of the JEL codes.

[14] Fetzer et al. (2024) employs a similar methodology in comparing descriptions of product descriptions with Harmonized System (HS) codes to allow for a structured network.

[15] The Plausibly Exogenous Galore dataset is a curated list of plausibly exogenous variations in empirical economics research, maintained by Sangmin S. Oh. Available at https://www.notion.so/1a897b8106ca44eeaf31dcd5ae5a61b1?v=ff7dc75862c6427eb4243e91836e077e.

2.3 Citations and Publication Data

Matching Publication Outcomes to Working Papers To analyze the publication trajectories of the working papers, we matched each paper to its eventual publication outcome using multiple data sources. Our primary goal was to determine whether a working paper was published in a peer-reviewed journal and, if so, identify the journal and publication date. This information is essential for understanding the dissemination and impact of research within the economics discipline.

We utilized four data sources to obtain publication information. First, we used official metadata from the NBER, which provides publication data collected via author submissions and automated scraping from RePEc.[16] While comprehensive, the dataset includes duplicates and primarily covers NBER papers. The second source was a large language model (LLM) prompted to retrieve publication outcomes based on its knowledge, yielding results for a small subset of NBER and CEPR papers. Third, we used the OpenAlex repository, matching titles of working papers and prioritizing published versions when multiple matches existed. Finally, we incorporated data from Baumann & Wohlrabe (2020), which provides manually verified publication outcomes for NBER and CEPR papers between 2000 and 2012, matched up to mid-2019.

To ensure consistency, we standardized journal names across these sources using the SCImago Journal Rank (SJR) lists for the fields of "Economics, Econometrics and Finance" and "Business, Management and Accounting." After removing generic journal names, this list included 2,367 unique journals.

Our matching process followed a hierarchical approach, prioritizing verified data. We first checked for a match in the dataset from Baumann & Wohlrabe (2020), followed by a search in OpenAlex for exact title matches. If no match was found, we consulted the NBER metadata, and finally used the LLM retrieval method for remaining papers. This approach ensured a comprehensive and reliable matching of publication outcomes. In total, the dataset from Baumann & Wohlrabe (2020) provided publication outcomes for 9,139 papers, OpenAlex

[16] This is available at https://www2.nber.org/wp_metadata/.

matched 10,840 papers, NBER metadata contributed 15,872 matches, and the LLM retrieval identified outcomes for 1,707 papers. By consolidating these sources, we obtained a detailed picture of the publication trajectories of a substantial number of working papers.

Despite this extensive coverage, certain limitations remain. The NBER metadata may be incomplete due to reliance on author submissions, and the Baumann & Wohlrabe (2020) dataset only covers papers up to 2012, matched to 2019. Additionally, errors may arise due to title similarities or data entry issues. However, by leveraging multiple sources, we minimized these limitations, resulting in a robust dataset for further analysis.

Citations Data To extend our analysis, we collected citations data for the working papers using three primary sources: RePEc's CiteEc service (https://citec.repec.org/), Baumann & Wohlrabe (2020), and the OpenAlex repository. We prioritized the citations data from RePEc's CiteEc service, which provides up-to-date (as of November 2024) citation counts for a large number of economics papers. For papers not included in CiteEc, we used the manually verified citations from Baumann & Wohlrabe (2020), which provides citation counts for NBER and CEPR papers published between 2000 and mid-2019. For any remaining papers, we obtained citation counts from OpenAlex, matched by exact paper titles. By merging these sources and prioritizing in this order, we assembled citations data for approximately 94.6% of our total sample, and 97.7% of the pre-2020 sample used in our analysis. This extensive coverage enables us to incorporate citations as a measure of research impact.

3 The Knowledge Graph of Economics

Over the past four decades, economic research has undergone a profound transformation, characterized by an increasing emphasis on establishing credible causal relationships using empirical methods—a shift often referred to as the "credibility revolution." Scholars such as Angrist & Pischke (2008) have highlighted the importance of rigorous econometric techniques designed to enhance causal inference, describing them as *Mostly Harmless Econometrics*. To systematically capture and analyze this evolution, we introduce a graphical approach by constructing a *knowl-*

edge graph for each paper in our dataset. This method allows us to quantitatively assess the complexity and structure of claims in economic research over time and observe changes in the adoption of causal inference methods.

3.1 Graphical Framework

For each paper p, we define a directed graph $G_p = (V_p, E_p)$, where V_p is the set of nodes representing economic concepts, classified using JEL codes, and E_p is the set of directed edges representing claims from a *source node* to a *sink node*. We use the terms "source" and "sink" to denote the direction of the claim within the paper, without presupposing causality. Whether a claim is interpreted as causal depends on the attributes of the edge connecting the nodes.

An important attribute of each edge $e \in E_p$ is whether the claim was evidenced using a *causal inference method*. We classify an edge as a *causal edge* if the associated claim in the paper was evidenced using one of the following methods: Difference-in-Differences (DiD), Instrumental Variables (IV/2SLS), Randomized Controlled Trials (RCTs/Experiments), Regression Discontinuity Design (RDD), Event Study, or Synthetic Control. This classification allows us to distinguish between causal and non-causal claims within the network. With this definition, approximately 19% of all claims in our dataset are classified as causal edges.

The network G_p thus includes all claims, encompassing both causal and non-causal edges. For instance, theoretical relationships between two concepts are represented as edges but are not considered causal unless they are supported by the specified empirical methods. This comprehensive approach enables us to analyze the overall structure of a paper's argumentation and the role of causal inference methods within it.

We consider several key measures derived from these claim networks. The *number of edges*, denoted as $|E_p|$, represents the total number of claims made in a paper, reflecting the breadth of the narrative. An increase in $|E_p|$ suggests a more complex narrative with multiple interrelated claims. Similarly, the *number of causal edges*, denoted as $|E_p^{\text{causal}}|$, represents the total number of claims evidenced using causal inference methods, indicating the depth of causal analysis within the paper.

Other measures include the *number of unique paths* in G_p, denoted as P_p, which is the total number of distinct directed paths between all pairs of nodes, excluding self-loops. This captures the interconnectedness of the narrative within the paper; a higher P_p indicates a more intertwined argument structure. The *longest path length* in G_p, denoted as L_p, represents the length of the longest directed path, indicating the depth of reasoning in the paper. We compute both P_p and L_p for the overall network and for the subnetwork consisting only of causal edges, denoted as P_p^{causal} and L_p^{causal}, respectively.

Illustrative examples To concretely illustrate our graphical framework and the measures derived from it, we examine four landmark economic papers: Chetty et al. (2014), Banerjee et al. (2015), Gabaix (2011), and Goldberg et al. (2010). These papers cover a diverse range of topics and methodologies, showcasing the versatility of our approach. The visual representations of these knowledge graphs are provided in Figure 3 and 4, which highlights the varying structures and complexities of the narratives in these influential papers.

In Chetty et al. (2014), the authors investigate the geography of intergenerational mobility in the United States using administrative records. The paper presents a comprehensive analysis of how various factors correlate with upward mobility. The knowledge graph for this paper (Figure 3a) includes seven edges, all of which are non-causal. The relationships mapped include how *parent income* (JEL code **D31**) influences *child income rank* (**J13**), and how factors such as *lower residential segregation* (**R23**), *less income inequality* (**D31**), *better primary schools* (**I21**), *greater social capital* (**Z13**), and *more stable family structures* (**J12**) are associated with *higher upward mobility* (**J62**). The knowledge graph measures for this paper are as follows: the *number of edges* $|E_p| = 7$ (all non-causal), the *number of unique paths* $P_p = 6$, and the *longest path length* $L_p = 1$. The relatively high number of edges indicates a broad exploration of factors affecting upward mobility, while the longest path length of 1 reflects that the relationships are primarily direct associations rather than extended causal chains.

Banerjee et al. (2015) report on a randomized evaluation of the impact of introducing microfinance in a new market. Utilizing Randomized Controlled Trials (RCTs), the authors assess how access to microfinance affects various economic outcomes for households in Hyderabad, India.

The knowledge graph for this paper (Figure 3b) includes eight edges, all of which are causal, evidenced through RCTs. The causal relationships include how *the introduction of microfinance* (**G21**) leads to *households having a microcredit loan* (**D14**), which in turn influences *new business creation* (**L26**) and *investment in existing businesses* (**G31**). These investments impact *average monthly per capita expenditure* (**D12**), while microcredit affects *expenditure on durable goods* (**E21**) and *expenditure on temptation goods* (**D12**). The authors also examine the effect of microcredit on *development outcomes* such as health, education, and women's empowerment (**I15**). The knowledge graph measures are: the *number of edges* $|E_p| = 8$ (all causal), the *number of unique paths* $P_p = 12$, and the *longest path length* $L_p = 3$. The high number of causal edges and unique paths indicates a complex causal narrative with multiple interconnected outcomes, while the longest path length reflects deeper causal chains.

(Figure 3)

In Gabaix (2011), the author proposes that idiosyncratic firm-level fluctuations can explain a significant part of aggregate economic shocks, introducing the "granular" hypothesis. The paper develops a theoretical framework and provides empirical evidence supporting the idea that shocks to large firms contribute to aggregate volatility. The knowledge graph (Figure 4a) includes six edges, all of which are non-causal, representing theoretical relationships such as how *idiosyncratic shocks to large firms* (**D21**) contribute to *aggregate volatility* (**E32**) and *GDP fluctuations* (**F44**), and how the *fat-tailed distribution of firm sizes* (**L11**) implies that *idiosyncratic shocks do not average out* (**E32**). The knowledge graph measures are: the *number of edges* $|E_p| = 6$ (all non-causal), the *number of unique paths* $P_p = 11$, and the *longest path length* $L_p = 3$. The relatively high number of unique paths and the longest path length indicate a complex theoretical narrative with multiple interconnected concepts and deeper reasoning chains.

Finally, Goldberg et al. (2010) examine the impact of imported intermediate inputs on domestic product growth in India. The authors use empirical methods, including Difference-in-Differences (DiD), to establish causal relationships between declines in input tariffs and firm performance. The knowledge graph (Figure 4b) for this paper includes five edges, three of which are causal. The causal relationships include how *declines in input tariffs* (**F14**) lead

to *increased firm product scope* (**L25**) and *improved firm performance* (**L25**), and how *increased availability of new imported inputs* (**O39**) causes *relaxed technological constraints for domestic firms* (**O33**). The knowledge graph measures are: the *number of edges* $|E_p| = 5$, the *number of causal edges* $|E_p^{\text{causal}}| = 3$, the *number of unique paths* $P_p = 5$, and the *longest path length* $L_p = 2$. These measures reflect a focused exploration of specific causal relationships, with a moderate level of narrative complexity.

(Figure 4)

These examples demonstrate the diversity in the structure and complexity of knowledge graphs across different types of economic research. They illustrate how our measures capture key aspects of the narratives, such as the breadth of topics covered, the depth of causal analysis, and the interconnectedness of concepts.

3.2 Observing the Credibility Revolution in Economic Research

To analyze the evolution of the use of causal inference methods in economic research, we focus on the *proportion of causal edges* in papers over time. This measure reflects the extent to which economists have increasingly adopted rigorous causal inference methods in their work, indicative of the credibility revolution.

Figure 5**(a)** displays the average proportion of causal edges per paper from 1980 to 2023. The data show a significant increase over time. In 1990, the average proportion of causal edges was approximately 4.2%. By 2000, it had risen modestly to around 8.4%. However, the increase became more pronounced in the subsequent decades: by 2010, the average proportion reached approximately 17.1%, and by 2020, it had climbed to around 27.8%. This upward trend indicates that economic papers have increasingly incorporated causal inference methods to substantiate their claims over the past three decades.

This substantial increase suggests a growing emphasis on establishing credible causal relationships in economic research. The proliferation of empirical methods and a heightened focus on causal identification strategies have contributed to this trend, reflecting the impact of the credibility revolution in economics.

We also examine how this proportion varies across different fields within economics. Figure 5**(b)** presents the average proportion of causal edges by field, comparing the pre-2000 and post-2000 periods. The data reveal that most fields have experienced substantial increases in the average proportion of causal edges in the post-2000 period.

Fields such as *Urban*, *Health*, *Development*, and *Behavioral* exhibit the highest increases. Urban increased from approximately 4.7% pre-2000 to 33.2% post-2000, marking one of the largest gains. Health saw an increase from 10.1% to 37.6%, achieving the highest post-2000 level among all fields. Development rose from 4.4% to 31.0%, and Behavioral increased from 3.6% to 29.6%. These fields, which often address policy-relevant questions and benefit from natural experiments or data conducive to causal analysis, have embraced causal inference methods more extensively.

Conversely, some fields experienced smaller increases or maintained lower levels. Macroeconomics increased modestly from 3.3% to 8.4%, reflecting a more cautious adoption of causal inference methods, possibly due to challenges in experimental design and identification strategies in macroeconomic contexts. Econometrics saw an increase from 6.3% to 11.0%, and Finance rose from 2.9% to 17.0%. These patterns suggest that the adoption of causal inference methods has varied across fields, influenced by the nature of the research questions, data availability, and methodological traditions within each field.

(Figure 5)

3.3 Evolution of Empirical Methods in Economic Research

To explore the increasing focus towards causal inference, we show time trends across methods and fields. Figure 6 illustrates the adoption of prominent empirical methods in NBER and CEPR working papers from 1980 to 2023. Methods such as DiD, IV, RCTs, and RDDs have seen substantial growth, reflecting the discipline's shift towards more rigorous identification strategies. [17]

DiD has become increasingly prevalent, rising from approximately 4% of papers in 1980 to

[17] A paper might use multiple methods and might be part of multiple fields, it's not mutually exclsuive etc– write it in the right way

over 15% in recent years. This growth underscores DiD's utility in exploiting policy changes and natural experiments to identify causal effects. IV methods have also seen a steady increase, from around 2% of papers in 1980 to over 6% by 2023, highlighting their role in addressing endogeneity through exogenous instruments (Angrist & Imbens 1994). The adoption of RCTs has accelerated since the early 2000s, increasing from less than 1% of papers in 2000 to over 7% in 2023, signaling the increasing feasibility and acceptance of experimental designs in economics. Similarly, RDD usage has grown from near zero in the 1980s to over 2% in recent years.

Conversely, the proportion of theoretical and non-empirical work has declined significantly, from approximately 20% of papers in 1980 to under 10% in 2023, indicating a broader emphasis on empirical analysis. The use of simulations has also decreased, from over 6% in 1980 to around 2–4% in recent years, possibly due to the availability of richer datasets and more sophisticated empirical methods.

(Figure 6)

These trends are not uniform across subfields. Figure 7 presents the distribution of empirical methods by field. Fields such as *Labour*, *Public*, and *Urban* heavily utilize DiD, with over 12% of papers employing this method in Labour and Public, and over 16% in Urban. RCTs are particularly prominent in *Behavioral* and *Development*, where they are used in over 20% and 11% of papers respectively, reflecting the feasibility and policy relevance of experimental interventions in these areas. This trend aligns with the findings of Currie et al. (2020), who documented a broad increase in empirical approaches due to the availability of big data and advanced computing.

In contrast, fields like *Macroeconomics*, *IO*, and *Finance* rely more on theoretical models and simulations to infer causal relationships from observational data. Theoretical approaches account for around 18% of papers in Macroeconomics and 12% in Finance. Structural estimation and simulations remain important in Macroeconomics and Industrial Organization, where complex theoretical models may be essential for understanding aggregate phenomena and market dynamics.

(Figure 7)

3.4 Additional Measures of Knowledge Graph Structure

Beyond the primary measures of claim network complexity, we explore other aspects of the knowledge graphs that capture different dimensions of research contributions. Three such measures are the *source-sink ratio*, the *proportion of novel edges*, and the *average eigenvector centrality*.

The *source-sink ratio*, denoted as R_p, quantifies the balance between the number of unique source nodes and unique sink nodes in a paper's knowledge graph G_p. It is defined as:

$$R_p = \frac{|\{v \in V_p \,|\, \text{out-degree}(v) > 0\}|}{|\{v \in V_p \,|\, \text{in-degree}(v) > 0\}| + \varepsilon},$$

where ε is a small constant to avoid division by zero. A value of $R_p = 1$ indicates an equal number of unique source and sink nodes, reflecting a balanced exploration of relationships in the paper. A ratio greater than one suggests a focus on multiple sources leading to fewer sinks, while a ratio less than one indicates that a few sources lead to multiple sinks.

The *proportion of novel edges*, denoted as U_p, measures the originality of a paper's claims by calculating the proportion of relationships that are novel compared to all prior research in our dataset. It is defined as:

$$U_p = \frac{|\{e \in E_p \,|\, e \notin \bigcup_{q<p} E_q\}|}{|E_p|},$$

where the papers are ordered chronologically, and $q < p$ represents all papers preceding paper p. A higher U_p indicates that a paper introduces more new relationships not previously documented, contributing to the novelty of the research.

The *average eigenvector centrality*, denoted as C_p, captures the influence or importance of the nodes within a paper's knowledge graph. Eigenvector centrality is a measure from network theory that assigns relative scores to all nodes in a network based on the principle that connections to high-scoring nodes contribute more to a node's score. For each node v in the knowledge graph G_p, the eigenvector centrality $c(v)$ is calculated, and the average eigenvector centrality for the paper is given by:

$$C_p = \frac{1}{|V_p|} \sum_{v \in V_p} c(v).$$

A higher average eigenvector centrality indicates that a paper's knowledge graph involves nodes that are more central or influential within the overall network of economic concepts, as defined by our dataset.

Illustrative examples To illustrate these measures, we revisit the four landmark papers, examining their differences in terms of source-sink ratio, proportion of novel edges, and average eigenvector centrality.

In Chetty et al. (2014), a source-sink ratio of $R_p = 2.5$ reflects a focus on multiple sources—residential segregation, income inequality, school quality, social capital, and family stability—affecting a single main outcome, upward mobility. With a proportion of novel edges $U_p = 0$, the relationships are already established in existing literature at the JEL code level. An average eigenvector centrality of $C_p = 0.14$ indicates engagement with relatively central economic concepts like income distribution and intergenerational mobility.

By contrast, Banerjee et al. (2015) exhibits a source-sink ratio of $R_p \approx 0.71$, signifying that a few sources, primarily the introduction of microfinance, lead to multiple effects including borrowing behavior, business creation, investment in existing businesses, expenditure patterns, and development outcomes. A high proportion of novel edges, $U_p = 0.75$, suggests substantial originality in the relationships examined. The average eigenvector centrality of $C_p = 0.187$, slightly higher than in Chetty et al., reflects engagement with central concepts, possibly due to the policy relevance of microfinance.

Similarly, Gabaix (2011) has a source-sink ratio of $R_p \approx 0.8$, indicating a balance between sources and sinks, with idiosyncratic shocks to large firms leading to aggregate outcomes like volatility and GDP fluctuations. The proportion of novel edges $U_p = 0.33$ shows some new contributions, while an average eigenvector centrality of $C_p = 0.09$ suggests engagement with less central, more specialized concepts, aligning with its innovative approach to macroeconomic fluctuations.

In Goldberg et al. (2010), the source-sink ratio is $R_p = 2$, with multiple sources—declines in input tariffs and increased availability of imported inputs—affecting firm outcomes like product scope and performance. A proportion of novel edges $U_p = 0.25$ reflects the introduction of some

new relationships, and an average eigenvector centrality of $C_p = 0.11$ indicates engagement with moderately central concepts such as trade liberalization and firm performance.

These examples demonstrate how variations in the structural features of knowledge graphs capture different dimensions of research contributions, including narrative focus, originality, and engagement with central or specialized topics. Understanding these measures provides deeper insights into how the complexity and structure of research narratives influence both publication success and academic impact.

Most central concepts in economics To gain deeper insights, we compute eigenvector centrality scores for both the overall knowledge graph and the subgraph consisting only of causal edges. This allows us to compare the centrality of economic concepts when considering all claims versus only those substantiated using causal inference methods.

Figure 8 displays the top 20 JEL codes ranked by their eigenvector centrality in both the overall and causal knowledge graphs. The comparison reveals interesting patterns in how central economic concepts differ based on the type of claims.

In the overall knowledge graph, the nodes with the highest eigenvector centrality scores include **G21** (Banks and Mortgages), **J31** (Wage Structure), and **I24** (Education and Inequality), reflecting their prominence in the economic literature. However, when focusing on the causal knowledge graph, the nodes with the highest centrality scores shift towards **I24** (Education and Inequality), **J13** (Fertility and Family), and **I21** (Analysis of Education).

This shift indicates that while certain economic concepts are central overall, the focus of causal inference methods tends to concentrate on specific areas, such as those related to education, family, and health economics. The disproportionate representation of these topics in the causal knowledge graph suggests that researchers employing causal inference methods may be addressing questions that are more amenable to experimental or quasi-experimental designs, such as those in education and health policy.

The distribution of eigenvector centralities is highly skewed in both graphs. In the overall knowledge graph, the mean eigenvector centrality is 0.0402, with a maximum of 0.898 for **G21**. In the causal knowledge graph, the centralities are more concentrated among certain nodes, with

I24 achieving the highest score normalized to 1.0. This normalization facilitates the comparison between the two graphs.

By incorporating average eigenvector centrality as a measure, we capture the extent to which a paper's claims involve central or peripheral concepts in economics. Papers engaging with highly central nodes in the causal knowledge graph may be contributing to well-established areas of research using rigorous methods, while those involving less central nodes might be exploring more novel or specialized topics.

These additional network measures provide deeper insights into how different research designs, topics, and narratives influence the structure of knowledge graphs. They illustrate the utility of our graphical approach in capturing and quantifying not only the complexity but also the balance, originality, and centrality of economic research.

(Figure 8)

By incorporating average eigenvector centrality as a measure, we capture the extent to which a paper's causal narrative involves central or peripheral concepts in economics. Papers engaging with highly central nodes may be building upon well-established ideas, while those involving less central nodes might be exploring more novel or specialized topics.

4 Knowledge Graph Predictors of Publication and Citations

In this section, we investigate how the structural properties of papers' knowledge graphs relate to their dissemination and impact within the academic community. Specifically, we analyze whether papers with certain narrative complexities and features are more likely to be published in higher-ranked journals and receive more citations. By examining both publication outcomes and citation counts, we aim to understand how the construction of research narratives influences both immediate recognition through publication in prestigious outlets and long-term influence measured by citations

4.1 Publication Outcomes

For this exercise, our sample includes papers for which we have publication data, focusing on three dependent variables: indicators for whether a paper is published in a top 5 economics journal, whether it is published in a top 6–20 journal, and whether it is published in a top 21–100 journal.[18] Our key independent variables are the knowledge graph measures defined earlier, calculated for both the full knowledge graphs (*All*) and the subgraphs consisting only of causal edges (*Causal*). These measures include the number of unique paths, the longest path length, the source-sink ratio, the proportion of novel edges, the average eigenvector centrality, the number of edges, and the proportion of edges that are causal. We include both the raw measures and their logarithmic transformations where appropriate. Year fixed effects are included in some specifications to control for temporal trends.

Figure 9 summarizes the regression results. Each panel corresponds to one of the publication outcome variables: publication in a top 5 journal, publication in a top 6–20 journal, and publication in a top 21–100 journal. Within each panel, we plot the estimated coefficients for the knowledge graph measures, with and without year fixed effects, along with 95% confidence intervals. The results for both the *All* and *Causal* versions of the measures are presented, allowing for comparison between the full knowledge graph and the causal subgraph.

Our analysis reveals several notable patterns:

First, papers with a higher proportion of causal edges (*Share Edges Causal*) are more likely to be published in top 5 journals. The coefficient on the proportion of causal edges is positive and statistically significant for top 5 publications, suggesting that the extent to which a paper employs causal inference methods is positively associated with placement in the most prestigious journals. This relationship holds both with and without year fixed effects.

Second, narrative complexity, as captured by the number of unique paths and the longest path length, is generally associated with higher publication outcomes. For the full knowledge graph measures (*All*), the log number of unique paths and the log longest path length are

[18] For context on types of papers published in top 5 journals, see Appendix Figure A1, which displays the proportion of working papers that were eventually published in the top 5, broken down by field and method. We find that certain fields and methods have higher publication rates in top journals, with field-method combinations such as Theoretical methods in Behavioural, Structural in IO or RCTs in Urban.

positively and significantly associated with publication in top 5 and top 6–20 journals. This suggests that papers with more complex narratives, involving multiple pathways and deeper chains of reasoning, are favored in higher-ranked journals. The positive association indicates that exploring various channels through which relationships occur enhances the likelihood of publication in top journals.

Interestingly, when focusing on the causal subgraph measures, the number of causal edges does not exhibit the same positive association with publication outcomes. The coefficients on the log number of edges for the causal subgraph are not significant predictors of top journal publications. This may reflect the challenge of rigorously evidencing a large number of causal relationships. Journals may value depth over breadth in causal claims, favoring papers that provide thorough analysis of fewer causal links.

Third, the type of complexity matters. While the overall number of relationships in a paper is positively correlated with publication in top journals, an increase in the source-sink ratio (i.e., the balance between source nodes and sink nodes) is associated differently across the *All* and *Causal* measures. For the full knowledge graph, a higher source-sink ratio is negatively associated with top 5 publication, suggesting that papers focusing on a few causes leading to multiple effects are favored. Conversely, for the causal subgraph, a higher source-sink ratio is positively associated with top 5 publication, indicating that papers exploring multiple causal factors leading to fewer outcomes are more likely to be published in top journals.

Fourth, the proportion of novel edges (*Share of New Edges*) shows a positive association with top 5 publication when considering the causal subgraph. Papers introducing new causal relationships that have not been previously documented are more likely to be published in top 5 journals. However, this pattern does not hold for publications in top 6–20 journals, where the coefficients are not statistically significant or even negative. This suggests that top journals may place a premium on novelty in causal claims, while other high-ranking journals may prefer research that builds upon established relationships.

Fifth, the average eigenvector centrality of nodes in a paper's knowledge graph is negatively associated with top 5 publication when considering both the full graph and the causal subgraph.

This implies that papers engaging with less central or more specialized concepts are more likely to be published in top 5 journals. In contrast, for publications in top 6–20 journals, the average eigenvector centrality is positively associated with publication outcomes, indicating a preference for papers focusing on more central concepts in the discipline. This difference suggests that top journals may favor innovative research exploring less examined areas, while other journals may prefer contributions that reinforce or expand upon central themes in economics.

Lastly, the number of edges in the full knowledge graph is a positive predictor of top 5 and top 6–20 publication outcomes. This further supports the notion that higher narrative complexity and breadth of relationships are valued in prestigious journals. However, the number of edges in the causal subgraph is not a significant predictor, reinforcing the idea that depth and rigorous evidence in causal claims are more critical than mere quantity.[19]

It is important to note that these results are correlational and should not be interpreted as causal effects of knowledge graph structures on publication outcomes. The observed associations may reflect editorial preferences, the nature of research favored by top journals, or other unobserved factors influencing both the structure of papers and their publication success. Our findings highlight patterns in the data but do not establish causality or imply that journals explicitly prefer certain types of papers based on these measures.

(Figure 9)

In summary, our findings suggest that certain structural features of papers, as captured by our knowledge graph measures, are associated with successful publication outcomes in economics journals. Higher narrative complexity, deeper causal chains, and the introduction of novel causal relationships are positively related to publication in top journals, particularly the top 5. Conversely, focusing on central concepts may be more advantageous for publication in top 6–20 journals. These patterns highlight how the structural composition of research narratives may influence dissemination and recognition in the field of economics.

[19]This may have implications for the topology of research findings that are published in leading journals, with a skew towards linear path length and depth rather than, e.g. high dimensional consistency across breadth of outcomes or measures.

4.2 Citation Counts

Building upon our previous analysis of publication outcomes, we now examine how knowledge graph structural properties relate to the citation impact of papers. Citations serve as a key indicator of a paper's influence and reception within the academic community. Understanding the factors that contribute to higher citation counts can provide insights into the dissemination and impact of research findings.

Our sample includes papers published up to the year 2020 to allow sufficient time for citation accumulation.[20] The distribution of citation counts in our sample is highly skewed, with a mean of 30.46 and a median of 10.5. Due to this skewness, we apply a logarithmic transformation to the citation counts using $\log(citations + 1)$ to normalize the distribution.[21]

To explore how citation distributions vary across journal categories, we create a kernel density plot of the citation percentiles for papers published in Top 5, Top 6–20, and Top 21–100 journals (see Appendix Figure A2). The plot reveals that, in general, papers published in higher-ranked journals tend to receive more citations. However, the most highly cited papers (above the 90th percentile) are slightly more evenly distributed across top 5 and top 6–20 journals. This suggests that while top journals generally publish papers that garner more citations, exceptionally influential papers can emerge from a wide range of field journals. This pattern may indicate that highly impactful research can "cut through" traditional signals of journal prestige, resonating with the academic community regardless of the publication venue.[22]

To investigate the relationship between knowledge graph measures and citation counts, we run regressions with the transformed citation counts as the dependent variable. We use the same set of independent variables as in Section 4.1, including both the *All* and *Causal* knowledge graph measures. We split the sample into four groups: all journals, Top 5 journals, Top 6–20

[20] Citations typically accrue over time; thus, including only papers published before 2020 helps ensure that citation counts are more reflective of the paper's impact.

[21] We also experimented with alternative transformations such as percentiles and deciles. The results are qualitatively similar, indicating robustness to different normalization methods.

[22] As summarized in Appendix Table A1, the mean citation percentile for Top 5 journals is 62.2, with a median of 68, while Top 6–20 journals have a mean of 57.1 and a median of 61. Top 21–100 journals show a mean of 52.1 and a median of 53. These statistics indicate that, on average, papers published in Top 5 journals have higher citation impact compared to those in lower-ranked journals. However, the overlap in distributions suggests that papers published in non-top 5 journals can achieve comparable citation impact.

journals, and Top 21–100 journals. This approach allows us to assess whether the predictors of citation impact differ by journal category.

Figure 10 presents the regression results. Each panel corresponds to one of the journal categories, and within each panel, we display the estimated coefficients for the knowledge graph measures, both with and without year fixed effects. The figure illustrates several key patterns.

First, for papers published in both Top 5 and Top 6–20 journals, higher values of certain knowledge graph measures are associated with increased citation counts. Specifically, the number of unique paths and the longest path length in the *Causal* subgraph show positive and significant coefficients. This suggests that within these journals, papers that explore complex causal narratives with multiple pathways and deeper causal chains tend to receive more citations. While the magnitude of the coefficients is slightly larger for Top 6–20 journals, the difference is not substantial. Such papers may resonate with the academic community due to their comprehensive exploration of causal mechanisms and rigorous identification strategies. This finding aligns with the broader trends of the credibility revolution in economics, where rigorous causal inference has become highly valued (Angrist & Pischke 2010, Imbens & Rubin 2015).

Second, the proportion of novel edges (*Share of New Edges*), particularly non-causal ones, shows a negative association with citation counts in most journals. This suggests that introducing new non-causal relationships may not enhance, and may even reduce, the citation impact of papers. It may be that building upon established relationships resonates more with the academic community, leading to higher citations. Although we find that the proportion of novel causal edges is positively associated with the likelihood of publication in top 5 journals (see Figure 9), this does not translate into higher citation counts once published. This indicates that while originality in causal claims may enhance publication success in top journals, it does not necessarily lead to greater academic influence as measured by citations.

Third, the average eigenvector centrality of nodes in the *full* knowledge graph shows a positive association with citation counts in most journals, including the Top 5, whereas the relationship is not statistically significant for Top 6–20 journals. This indicates that, once published, papers focusing on more central concepts tend to receive more citations in most journals, sug-

gesting that engaging with widely recognized topics enhances a paper's impact. Interestingly, this contrasts with our earlier finding that average eigenvector centrality is negatively associated with the likelihood of publication in Top 5 journals (Figure 9). This suggests that while top journals favor innovative research exploring less-examined areas, papers addressing more central concepts receive more citations once published. One possible interpretation is that top journals prioritize novelty and specialization at the publication stage, but within those journals, papers on central topics garner more attention and citations. In contrast, Top 6–20 journals, which are often top field journals, prefer contributions focusing on more central concepts within their specific areas, but engaging with central concepts does not necessarily enhance citation impact in these journals.

Moreover, across journal categories, we find that the proportion of causal edges does not have a significant positive association with citation counts. This suggests that while employing causal inference methods may enhance the likelihood of publication in top-tier journals (Figure 9), it does not necessarily translate into higher citation impact once published. This may reflect a trade-off between the methodological rigor required for precise identification and the broader relevance or generalizability of the findings. Papers focusing on narrowly defined, precisely identified causal relationships might be valued by journal editors and referees for their methodological contributions, but may not attract as many citations if they address less pressing or less generalizable issues. This result highlights the ongoing debate in the economics profession about the balance between precise identification and addressing big, policy-relevant questions.[23]

(Figure 10)

These findings highlight that predictors of citation impact vary across journal categories. In both Top 5 and Top 6–20 journals, embracing narrative complexity and developing complex causal narratives enhance a paper's citation impact. This may be due to the visibility and

[23]This speaks to discussions on the direction of the economics profession. For instance, see this tweet by Dani Rodrik: "From any rational, decision-theoretic standpoint, the Economics profession under-invests in imperfectly identified analyses of big/important/relevant questions relative to well-identified but comparatively uninteresting questions".

credibility of these journals, where innovative and comprehensive research is more likely to be recognized and cited. Researchers may place greater trust in papers published in higher-ranked journals because of rigorous peer-review processes and high methodological standards (Card & DellaVigna 2013, Hamermesh 2013). Moreover, the increased citation impact of such papers may reflect the academic community's appreciation for research that not only employs empirical methods but also delves deeply into causal mechanisms.

In sum, the relationship between knowledge graph measures and citation impact is nuanced, influenced by the interplay of methodological rigor, narrative complexity, and journal prestige. Our analysis reveals that the structural features of a paper's knowledge graph are linked to its citation impact, varying by journal category. In Top 5 and Top 6–20 journals, papers with greater narrative complexity, deeper causal chains, and complex causal narratives receive more citations, highlighting the importance of both research content and structure, alongside publication venue, in determining influence.

These results align with broader discussions on the evolution of economic research and drivers of academic impact. The credibility revolution's emphasis on rigorous causal inference appears reflected in citation patterns, especially in higher-ranked journals. Yet, concerns remain about overemphasizing methodology at the expense of theoretical innovation (Heckman 2001, Keane 2010); balancing methodological rigor with substantive contributions continues to be a critical challenge for the discipline.

5 Challenges in Replication and Data Accessibility

While the increased use of causal inference methods has enhanced the credibility of economic research, concerns about replicability and transparency persist. Specifically, the decline in reporting null results and the growing reliance on proprietary data may hinder the replication and verification of studies. In this section, we examine trends in the reporting of null results and the use of private sector data in economic research.

5.1 Reporting of Null Results

Reporting null results is vital for scientific transparency and reducing publication bias (Rosenthal 1979, Sterling 1959). However, null results are often underreported due to perceived lower publication likelihood (Brodeur et al. 2016, Chopra et al. 2024). Figure 11**(a)** shows that the average share of null result claims per paper decreased from approximately 15% in 1980 to about 8.6% in 2023. This decline suggests increased pressure to produce significant findings or possible publication bias favoring positive results.

In our knowledge graph framework, the share of null edges for each paper p is:

$$\text{Null Edge Share}_p = \frac{|\{e \in E_p \,|\, \text{edge } e \text{ represents a null result}\}|}{|E_p|}.$$

Variation Across Fields Figure 11**(b)** compares the average share of null results by field pre- and post-2000. Most fields show a decrease post-2000. Econometrics and Behavioral Economics report higher shares of null results, decreasing from 16.05% to 12.34% and from 11.71% to 11.02%, respectively. Fields like Finance and IO report lower shares, decreasing from 11.44% to 7.93% and from 10.05% to 7.51%. These differences may reflect varying research practices and publication norms across fields. Fields heavily relying on experimental or quasi-experimental methods, such as Behavioral and Econometrics, may be more likely to report null results due to their methodologies. Conversely, fields like Finance and IO may face stronger publication biases against null findings or focus on research questions more likely to yield significant results.

Variation Across Methods Figure 11**(c)** displays the average share of null results by empirical method pre- and post-2000. Methods like RCTs, DiD, and RDD have higher shares of null results. RCTs increased slightly from 13.15% to 13.57%, while DiD decreased from 12.80% to 9.94%. Structural Estimation and Theoretical work report lower shares, decreasing from 9.78% to 6.12% and from 10.85% to 8.95%, respectively. These patterns may reflect the differing nature of these methods. Experimental and quasi-experimental methods like RCTs and RDDs, designed for rigorous causal inference, may often result in null findings when interventions do not produce significant effects. Transparent reporting of such results is important to avoid publication bias.

Interaction Between Field and Method Figure 11**(d)** presents a heatmap of null result shares by field and method. High shares are observed in combinations like RCTs in Labour Economics (16.07%) and Development Economics (16.16%), and RDDs in Health Economics (18.27%). Lower shares are seen in Structural methods within Finance (5.46%) and IO (3.20%). These findings suggest that both field and method influence the reporting of null results, with experimental and quasi-experimental methods in certain fields more likely to report null findings. Emphasizing the publication of null results is essential to maintain scientific integrity and avoid publication bias.

(Figure 11)

5.2 Private Sector Data and Accessibility

Data availability is crucial for replication and verification of research findings. While open data is professed as an ideal, it is not widely practiced in economics (Andreoli-Versbach & Mueller-Langer 2014). The use of proprietary data exacerbates the problem, limiting other researchers' ability to replicate studies or test alternative hypotheses.[24] Data privacy regulations like the GDPR have introduced additional barriers to data sharing. This tension highlights the need for policies that balance privacy concerns with the benefits of data accessibility for scientific advancement (Fetzer 2022).[25] Moreover, Barrios et al. (2024) find that the use of private data significantly reduces trust in economics research among both economists and the general public. Their study shows that papers with conflicts of interest, such as reliance on proprietary data, are perceived as less credible, undermining the perceived value of the research. In response to these challenges, Miguel (2021) documents the adoption of open science practices in economics, such as pre-registration and data sharing, noting a rapid transition toward increased transparency.

Figure 12**(a)** shows that the proportion of papers using private company data rose from approximately 3.97% in 1980 to around 8.61% in 2023, with a notable increase post-2000. This

[24]The rise of private data in research raises concerns about data ownership and property rights, potentially creating barriers to data access and replication. This trend reflects broader challenges related to data governance and the emergence of data markets, where data is treated as a valuable asset controlled by private entities.

[25]For a discussion on navigating data privacy in research, see World Bank (2020).

reflects greater availability of granular data from private companies and increased collaboration between researchers and private entities.

Variation Across Fields Figure 12**(b)** compares the use of private data by field pre- and post-2000. Fields like Finance, IO, and Behavioral Economics exhibit higher proportions post-2000. For instance, Finance increased from 6.33% to 13.66%, IO from 5.92% to 13.60%, and Behavioral from 2.90% to 13.97%. These increases may reflect the nature of research in these fields, which often relies on firm-level or experimental data that is not publicly available. The use of proprietary datasets allows researchers to conduct detailed analyses of financial markets, consumer behavior, and firm dynamics. In contrast, fields like Economic History and Econometrics have lower proportions of private data usage. For example, Economic History shows an increase from 0.56% pre-2000 to only 1.52% post-2000. This may be due to the reliance on historical data sources and publicly available datasets in these fields.

Variation Across Methods Figure 12**(c)** shows the average proportion of private data usage by method pre- and post-2010. Methods such as Event Studies, DiD, and IV are associated with higher private data usage. Event Studies increased from 11.65% to 15.25%, DiD from 7.83% to 12.08%, and IV from 3.98% to 9.75%. These methods often require detailed data on firm events, policy changes, or instrumental variables that may be proprietary or collected by private companies. The increasing reliance on these methods may contribute to the greater use of private data in economic research.

Interaction Between Field and Method Figure 12**(d)** presents a heatmap of private data usage by field and method, categorized into Low (below 4%), Medium (4% to 10%), and High (above 10%). High usage is observed in combinations like DiD in Behavioral Economics (28.85%) and Finance (20.06%), and Structural Estimation in IO (27.50%). Lower usage is seen in RCTs within Economic History and Econometrics.

The increasing reliance on proprietary data raises concerns about replicability and transparency. Balancing data accessibility with privacy and proprietary rights remains a critical

challenge. Policies promoting data sharing and transparency, while respecting privacy, are essential for advancing scientific knowledge.

(Figure 12)

6 Conclusion

This study analyzes over 44,000 NBER and CEPR working papers from 1980 to 2023 using a custom language model to construct knowledge graphs that map economic concepts (JEL codes) and their relationships, distinguishing between general claims and those substantiated with causal inference methods. We find a significant increase in the use of causal inference methods, with the average proportion of causal claims rising from about 4% in 1990 to nearly 28% in 2020, reflecting the growing influence of the credibility revolution in economics.

Our analysis reveals a divergence between factors influencing publication in top journals and those driving citation impact. While papers employing causal inference methods, introducing novel causal relationships, and engaging with less central, specialized concepts are more likely to be published in top 5 journals, these features do not necessarily translate into higher citation counts. Instead, papers focusing on central concepts tend to receive more citations once published. However, narrative complexity—measured by the number of unique paths and longest path length in the causal subgraph—is positively associated with both publication in top journals and increased citation counts, suggesting that depth and complexity in causal narratives are valued for both publication success and academic influence.

We also find that the balance between source and sink nodes affects outcomes differently depending on the graph considered. In the full knowledge graph, papers focusing on few sources leading to multiple sinks (few causes to many effects) are favored in top journals. Conversely, in the causal subgraph, papers exploring multiple causal factors leading to fewer outcomes are more likely to be published in top journals and receive higher citations.

Additionally, we observe a decline in the reporting of null results and an increased use of proprietary data, doubling from about 4% in 1980 to over 8% in 2023, raising concerns about

transparency and replicability.

Our findings highlight some dimensions of the trade-off between methodological rigor, narrative structure, topic centrality, and their differential effects on publication success and academic influence. They suggest a trade-off between pursuing innovative, specialized topics for publication in top journals and engaging with central, widely recognized concepts for broader impact. Encouraging transparency, fostering the reporting of null results, and balancing methodological rigor with broader relevance are essential for enhancing the credibility and impact of economic research.

References

Alabrese, E. (2022), 'Bad science: Retractions and media coverage'.

Alabrese, E., Capozza, F. & Garg, P. (2024), 'Politicized scientists: Credibility cost of political expression on twitter'.

Andre, P. & Falk, A. (2021), What's worth knowing in economics? a global survey among economists. Working Paper.

Andreoli-Versbach, P. & Mueller-Langer, F. (2014), 'Open access to data: An ideal professed but not practised', *Research Policy* **43**(9), 1621–1633.

Andrews, I. & Kasy, M. (2019), 'Identification of and correction for publication bias', *American Economic Review* **109**(8), 2766–2794.

Angrist, J., Azoulay, P., Ellison, G., Hill, R. & Lu, S. F. (2017), 'Economic research evolves: Fields and styles', *American Economic Review: Papers and Proceedings* **107**(5), 293–297.

Angrist, J. D. & Pischke, J.-S. (2008), *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press, Princeton, NJ.

Angrist, J. D. & Pischke, J.-S. (2010), 'The credibility revolution in empirical economics: How

better research design is taking the con out of econometrics', *Journal of Economic Perspectives* **24**(2), 3–30.

Angrist, J. & Imbens, G. (1994), 'Identification and estimation of local average treatment effects'.

Athey, S. & Imbens, G. W. (2019), 'Machine learning methods that economists should know about', *Annual Review of Economics* **11**(1), 685–725.

Banerjee, A., Duflo, E., Glennerster, R. & Kinnan, C. (2015), 'The miracle of microfinance? evidence from a randomized evaluation', *American economic journal: Applied economics* **7**(1), 22–53.

Barrios, J., Lancieri, F. M., Levy, J., Singh, S., Valletti, T. M. & Zingales, L. (2024), The conflict-of-interest discount in the marketplace of ideas, Technical report, New Working Paper Series.

Baumann, A. & Wohlrabe, K. (2020), 'Where have all the working papers gone? evidence from four major economics working paper series', *Scientometrics* **124**(3), 2433–2441.

Behavioural Insights Team (2024), 'A blueprint for better international collaboration on evidence'.
URL: *provide URL here*

Bloom, N., Jones, C. I., Van Reenen, J. & Webb, M. (2020), 'Are ideas getting harder to find?', *American Economic Review* **110**(4), 1104–1144.

Brodeur, A., Cook, N. & Neisser, C. (2024), 'P-hacking, data type and data-sharing policy', *The Economic Journal* **134**(659), 985–1018.

Brodeur, A., Lé, M., Sangnier, M. & Zylberberg, Y. (2016), 'Star wars: The empirics strike back', *American Economic Journal: Applied Economics* **8**(1), 1–32.

Card, D. & DellaVigna, S. (2013), 'Nine facts about top journals in economics', *Journal of Economic Literature* **51**(1), 144–161.

Cartwright, N. (2007), 'Are rcts the gold standard?', *BioSocieties* **2**(1), 11–20.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. & Robins, J. M. (2018), 'Double/debiased machine learning for treatment and structural parameters', *The Econometrics Journal* **21**(1), C1–C68.

Chetty, R., Hendren, N., Kline, P. & Saez, E. (2014), 'Where is the land of opportunity? the geography of intergenerational mobility in the united states', *The quarterly journal of economics* **129**(4), 1553–1623.

Chopra, F., Haaland, I., Roth, C. & Stegmann, A. (2024), 'The null result penalty', *The Economic Journal* **134**(657), 193–219.

Currie, J., Kleven, H. & Zwiers, E. (2020), Technology and big data are changing economics: Mining text to track methods, *in* 'AEA Papers and Proceedings', Vol. 110, American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, pp. 42–48.

de Quidt, J., Haushofer, J. & Roth, C. (2018), 'Measuring and bounding experimenter demand', *American Economic Review* **108**(11), 3266–3302.
URL: *https://www.aeaweb.org/articles?id=10.1257/aer.20171330*

Deaton, A. (2010), 'Instruments, randomization, and learning about development', *Journal of Economic Literature* **48**(2), 424–455.

Deaton, A. & Cartwright, N. (2018), 'Understanding and misunderstanding randomized controlled trials', *Social Science Medicine* **210**, 2–21.

Delbono, F., Reggiani, C. & Sandrini, L. (2024), 'Strategic data sales with partial segment profiling', *Information Economics and Policy* **68**, 101102.

Fetzer, T. (2022), 'Thiemo Fetzer Discussion of: COVID and Income Inequality', *Economic Policy* **37**(109), 201–203.
URL: *https://doi.org/10.1093/epolic/eiac016*

Fetzer, T., Lambert, J. P., Garg, P. & Feld, B. (2024), Ai-generated production networks: Measurement and applications to global trade, Technical report, Working Paper.

Frankel, A. & Kasy, M. (2022), 'Which findings should be published?', *American Economic Journal: Microeconomics* **14**(1), 1–38.

Gabaix, X. (2011), 'The granular origins of aggregate fluctuations', *Econometrica* **79**(3), 733–772.

Garg, P. & Fetzer, T. (2024), 'Political expression of academics on social media'.

Gelman, A. & Loken, E. (2014), 'The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time', *Department of Statistics, Columbia University* **348**, 1–17.

Goldberg, P. K., Khandelwal, A. K., Pavcnik, N. & Topalova, P. (2010), 'Imported intermediate inputs and domestic product growth: Evidence from india', *The Quarterly journal of economics* **125**(4), 1727–1767.

Hamermesh, D. S. (2013), 'Six decades of top economics publishing: Who and how?', *Journal of Economic Literature* **51**(1), 162–172.

Heckman, J. J. (2001), 'Micro data, heterogeneity, and the evaluation of public policy: Nobel lecture', *Journal of Political Economy* **109**(4), 673–748.

Heckman, J. J. & Moktan, S. (2020), 'Publishing and promotion in economics: The tyranny of the top five', *Journal of Economic Literature* **58**(2), 419–470.

Imbens, G. W. & Rubin, D. B. (2015), *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, Cambridge University Press.

Ioannidis, J. P. (2005), 'Why most published research findings are false', *PLOS Medicine* **2**(8), e124.

Jackson, M. O. (2010), *Social and Economic Networks*, Princeton University Press, Princeton, NJ.

Keane, M. P. (2010), 'A structural perspective on the experimentalist school', *Journal of Economic Perspectives* **24**(2), 47–58.

Mellon, J. (2021), 'Rain, rain, go away: 194 potential exclusion-restriction violations for studies using weather as an instrumental variable', *American Journal of Political Science* .

Merton, R. K. (1968), 'The matthew effect in science', *Science* **159**(3810), 56–63.

Miguel, E. (2021), 'Evidence on research transparency in economics', *Journal of Economic Perspectives* **35**(3), 193–214.

Nature Editorial (2024), 'Unearthing 'hidden' science would help to tackle the world's biggest problems', *Nature* **633**(7930), 493. Editorial.
URL: *https://doi.org/10.1038/d41586-024-02991-5*

Newman, M. E. (2003), 'The structure and function of complex networks', *SIAM Review* **45**(2), 167–256.

Park, M., Leahey, E. & Funk, R. J. (2023), 'Papers and patents are becoming less disruptive over time', *Nature* **613**(7942), 138–144.

Ravallion, M. (2009), 'Should the randomized controlled trial be the gold standard for development research?', *World Bank Research Observer* **24**(1), 30–53.

Resnik, D. B. (1998), *The Ethics of Science: An Introduction*, Routledge, London, UK.

Rosenbaum, P. R. & Rosenbaum, P. R. (2002), *Overt bias in observational studies*, Springer.

Rosenthal, R. (1979), 'The file drawer problem and tolerance for null results', *Psychological Bulletin* **86**(3), 638.

Rubin, D. B. (1984), 'Bayesianly justifiable and relevant frequency calculations for the applied statistician', *Annals of Statistics* **12**(4), 1151–1172.

Simmons, J. P., Nelson, L. D. & Simonsohn, U. (2011), 'False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant', *Psychological Science* **22**(11), 1359–1366.

Simonsohn, U., Nelson, L. D. & Simmons, J. P. (2014), 'p-curve and effect size: Correcting for publication bias using only significant results', *Perspectives on Psychological Science* **9**(6), 666–681.

Sims, C. A. (2010), 'But economics is not an experimental science', *Journal of Economic Perspectives* **24**(2), 59–68.

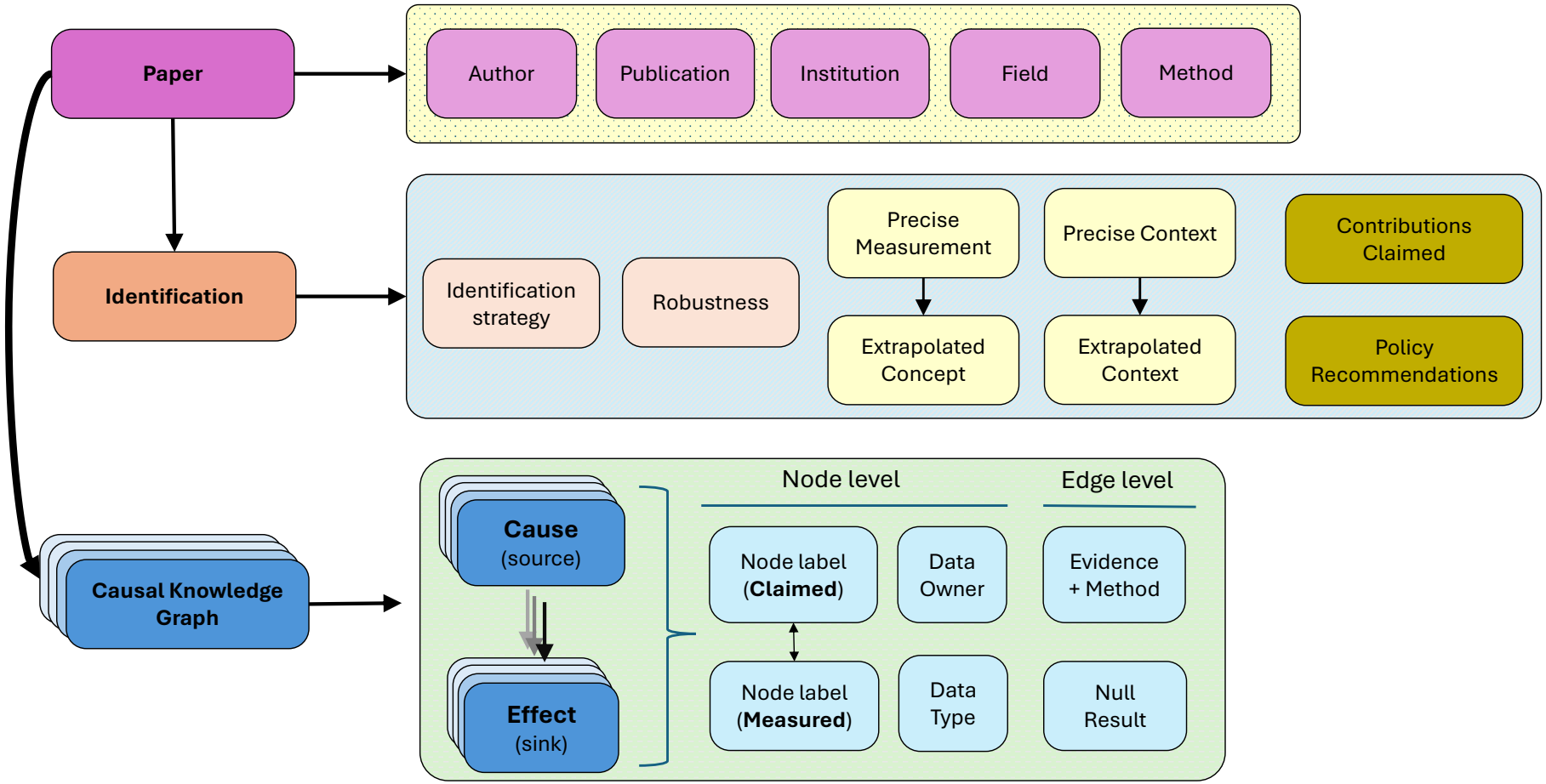
Sterling, T. D. (1959), 'Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa', *Journal of the American Statistical Association* **54**(285), 30–34.

Wasserstein, R. L. & Lazar, N. A. (2016), 'The asa's statement on p-values: Context, process, and purpose', *The American Statistician* **70**(2), 129–133.

World Bank (2020), 'Data privacy principles: Gdpr and beyond', *World Bank Publications* .
URL: *https://www.worldbank.org/en/data/dataprivacy*

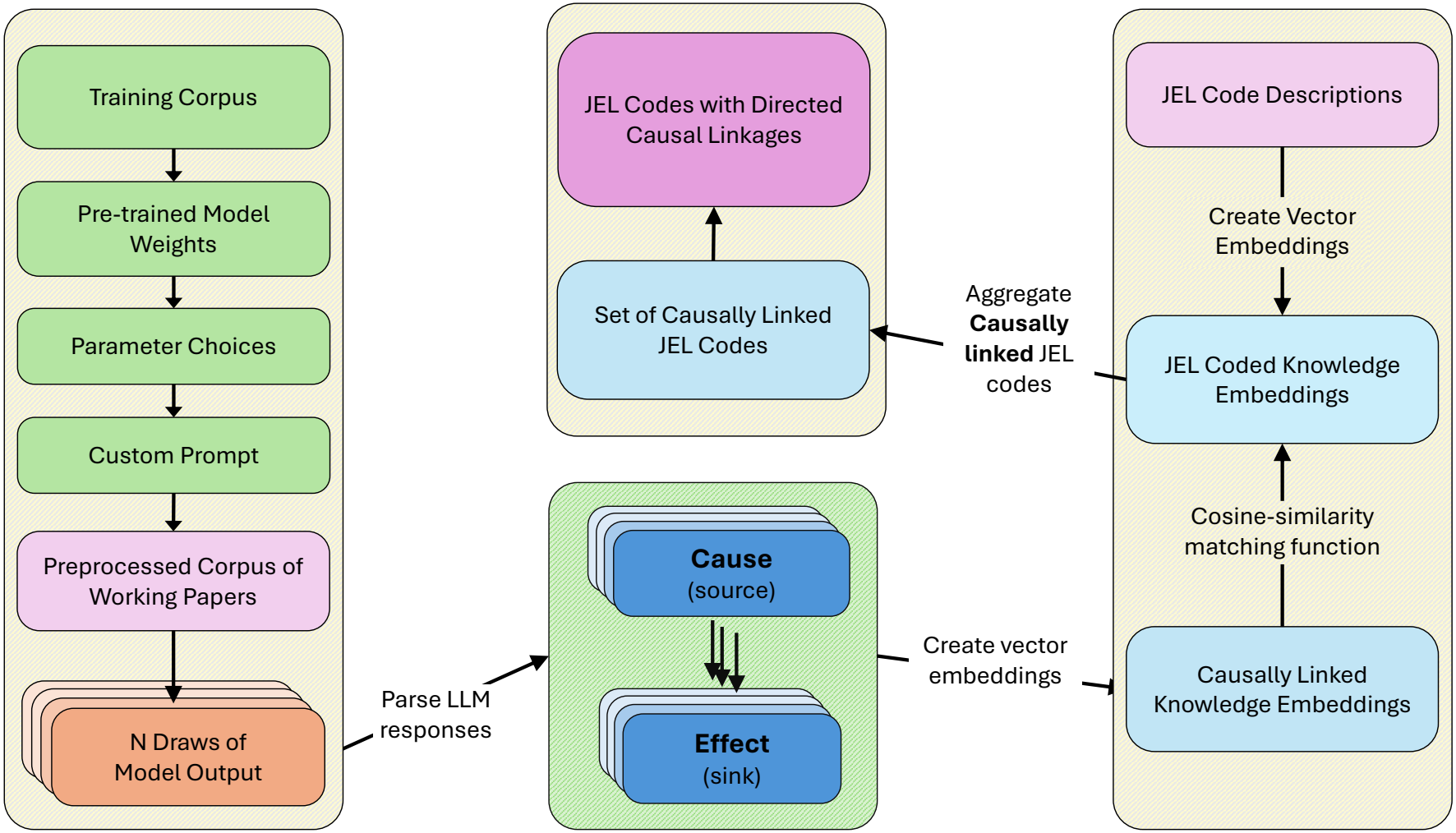
7 Figures and Tables

Figure 1: Retrieval of Concepts Using AI



Note: This flowchart illustrates our AI-powered approach to retrieving, assessing, and mapping causal claims and contributions from academic papers. The process begins with academic papers, from which the LLM extracts fields such as Author, Publication, Institution, Field, Method, and Data/Code Availability. These aspects feed into two main branches: **Identification** and **Causal Claims**. The **Identification** branch focuses on elements like Identification Strategy and Robustness Checks. The analysis extends to understanding precise measurements and contexts, as well as extrapolated concepts and contexts, leading to insights on contributions claimed and policy recommendations. The **Causal Claims** branch involves analyzing the causal and non-causal relationships identified in the papers, consisting of arrays of source (or cause) and sink (or effect) variables. The analysis operates across three levels. First, for each source or sink node, we consider the source of sink as claimed by the author and as measured in the paper, including the type the owner of the data used. Second, for each source-sink edge, we examine the method(s) used to evidence a claim, and whether null result was found. Third, at the graph level, we assess graphical measures like the number of steps taken from source to sink, the descriptions of these steps, and the overall complexity of the underlying narrative.

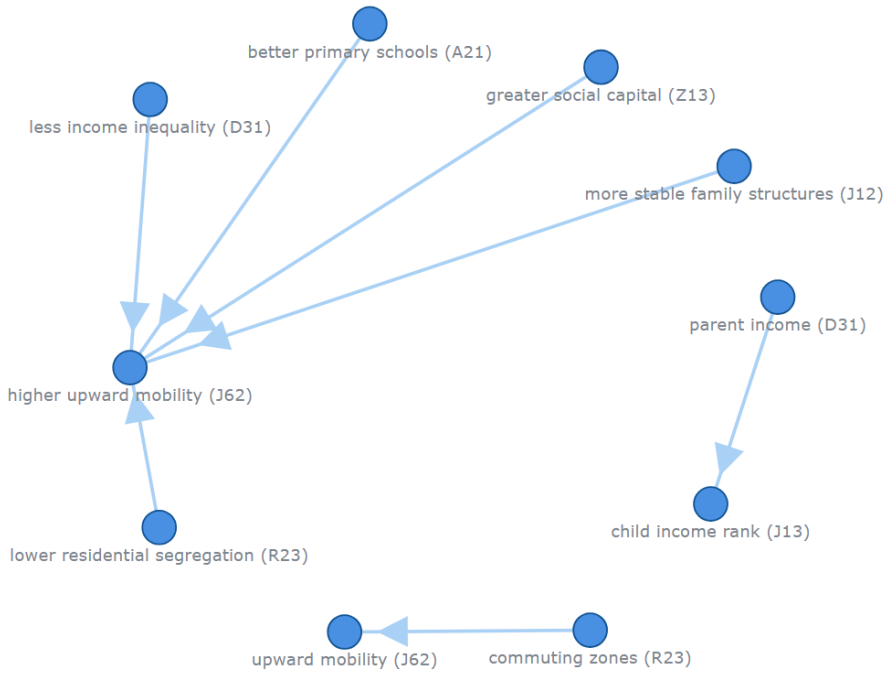
Figure 2: Mapping Causal Linkages Between JEL Codes Using AI



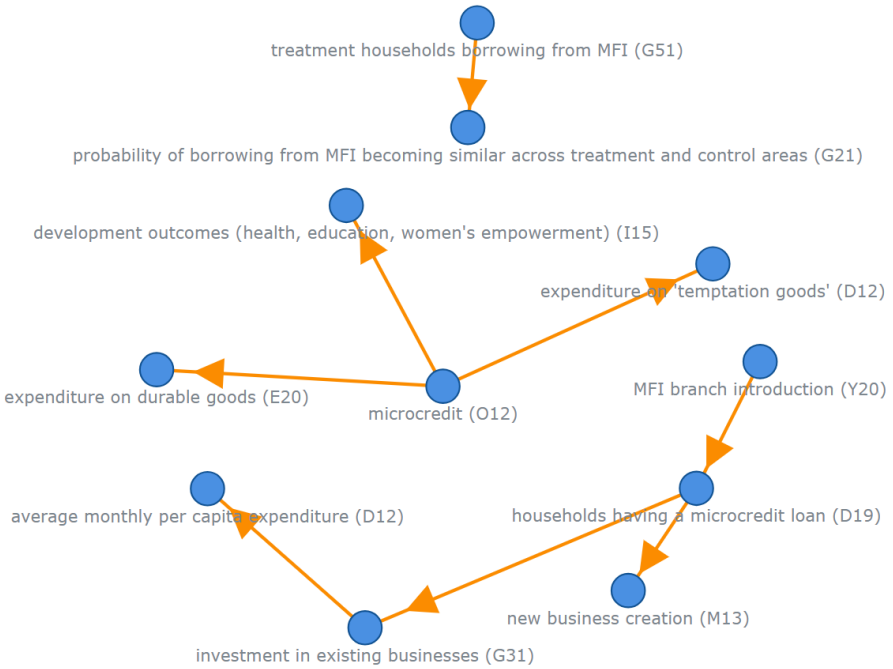
Note: This diagram illustrates our AI-driven methodology for analyzing and mapping causal and non-causal linkages between economic concepts, represented by JEL (Journal of Economic Literature) codes. Starting with a corpus of working papers, we use a custom prompt and pre-trained language model to extract causal relationships, identifying source (or cause) and sink (or effect) variables within the text. The extracted edge are parsed to generate directed linkages between JEL codes, forming a knowledge graph that aggregates these relationships across the corpus. We employ OpenAI's vector embeddings to numerically represent descriptions of JEL codes and utilize cosine similarity with sources and sinks, assigning the most similar JEL code to each of the source and sink nodes. This approach enables us to construct a structured representation of evidence in economics over time, facilitating the exploration of interconnected economic concepts and the evolution of empirical research frontiers.

Figure 3: Knowledge Graphs of Two Landmark Economic Papers (Part 1)

(a) Chetty et al. (2014)



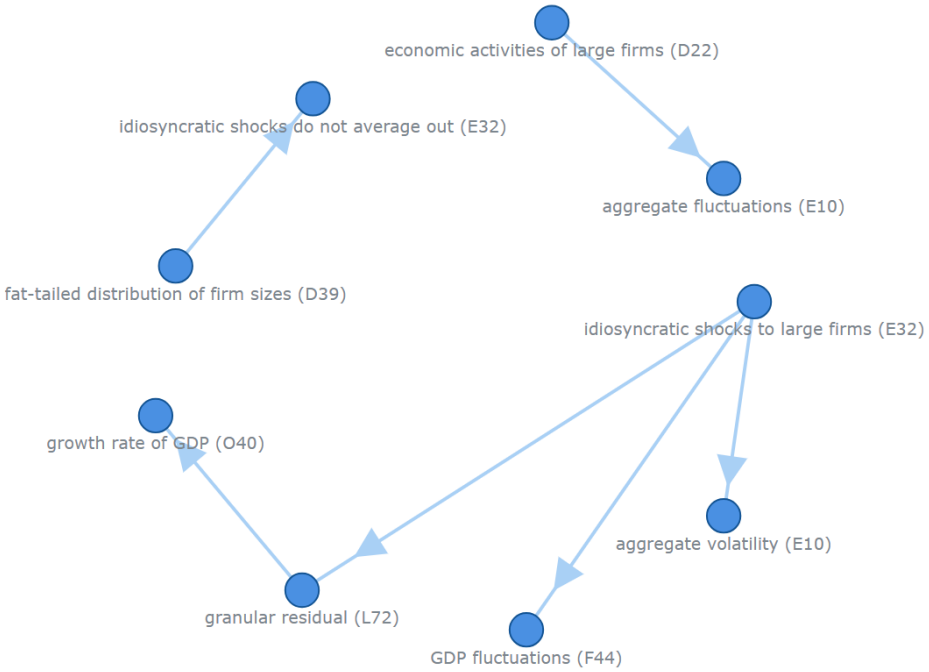
(b) Banerjee et al. (2015)



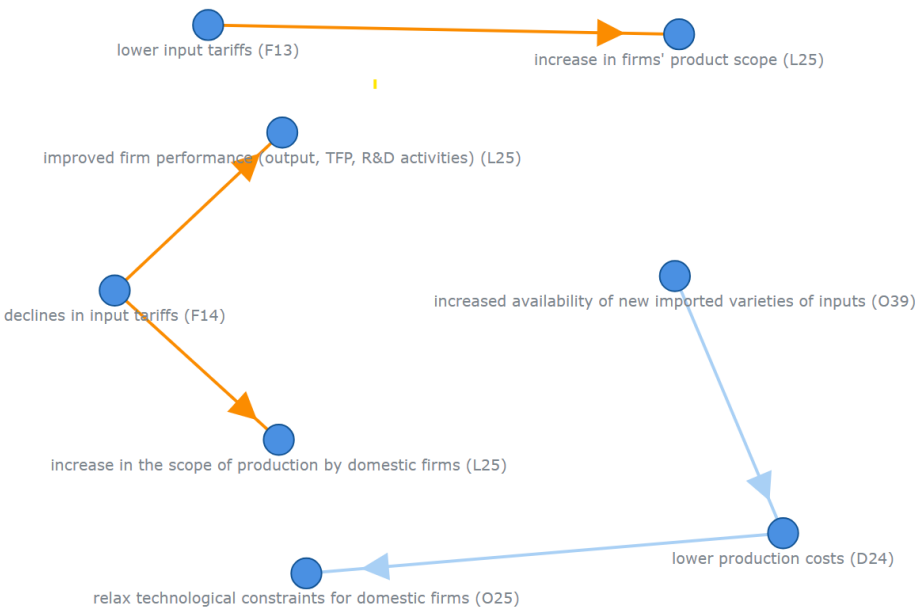
Note: This figure presents the knowledge graphs of two landmark economic papers. Causal relationships are shown in orange, non-causal relationships in blue. Nodes represent economic concepts mapped to JEL codes; arrows indicate the direction of claims from source to sink. Panel (a) displays the graph for Chetty et al. (2014), showcasing multiple factors associated with upward mobility in the United States. The graph has $|E_p| = 7$ edges (all non-causal), $P_p = 6$ unique paths, and a longest path length of $L_p = 1$, indicating a broad but direct exploration of associations without extended causal chains. Panel (b) shows the graph for Banerjee et al. (2015), illustrating the causal impact of introducing microfinance in India. The graph has $|E_p| = 8$ edges (all causal), $P_p = 12$ unique paths, and a longest path length of $L_p = 3$, reflecting a complex causal narrative with multiple interconnected outcomes resulting from the intervention. **Color Coding**: Edges evidenced using causal inference methods are in orange; non-causal edges are in blue.

Figure 4: Knowledge Graphs of Two Landmark Economic Papers (Part 2)

(a) Gabaix (2011)



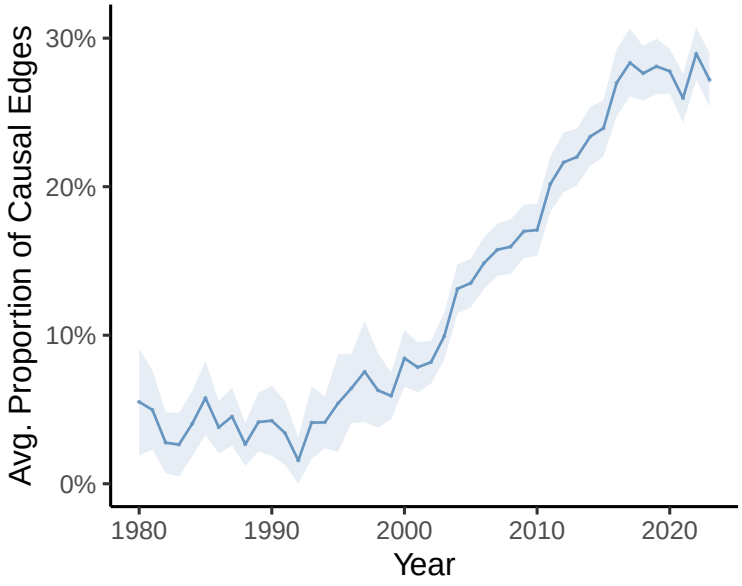
(b) Goldberg et al. (2010)



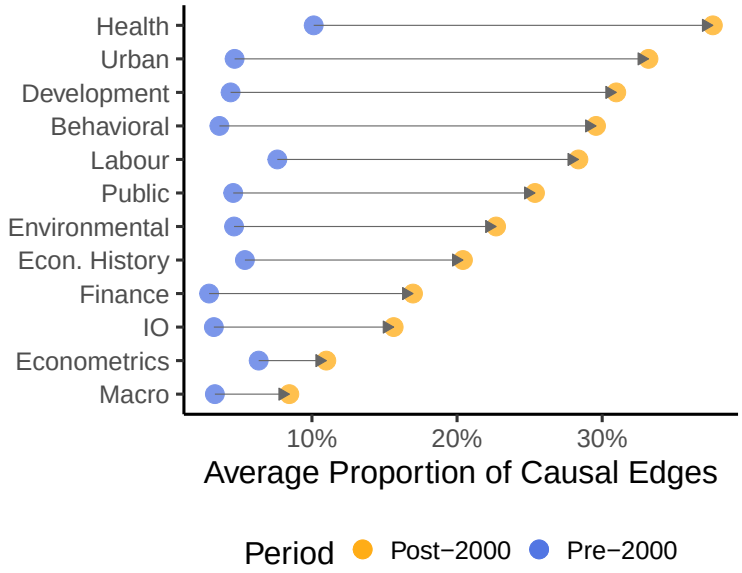
Note: This figure presents the knowledge graphs of two landmark economic papers. Causal relationships are shown in orange, non-causal relationships in blue. Nodes represent economic concepts mapped to JEL codes; arrows indicate the direction of claims from source to sink. Panel (a) presents the graph for Gabaix (2011), depicting theoretical relationships in macroeconomics concerning the impact of idiosyncratic firm-level shocks on aggregate economic fluctuations. The graph has $|E_p| = 6$ edges (all non-causal), $P_p = 11$ unique paths, and a longest path length of $L_p = 3$, indicating a complex theoretical narrative with deeper reasoning chains. Panel (b) displays the graph for Goldberg et al. (2010), focusing on the effects of input tariff reductions on Indian firms' product growth and performance. The graph has $|E_p| = 5$ edges (3 causal), $P_p = 5$ unique paths, and a longest path length of $L_p = 2$, reflecting a focused exploration of specific causal relationships supported by empirical methods. **Color Coding**: Edges evidenced using causal inference methods are in orange; non-causal edges are in blue.

Figure 5: Trends in the Proportion of Causal Edges Over Time and by Field

(a) Over Time

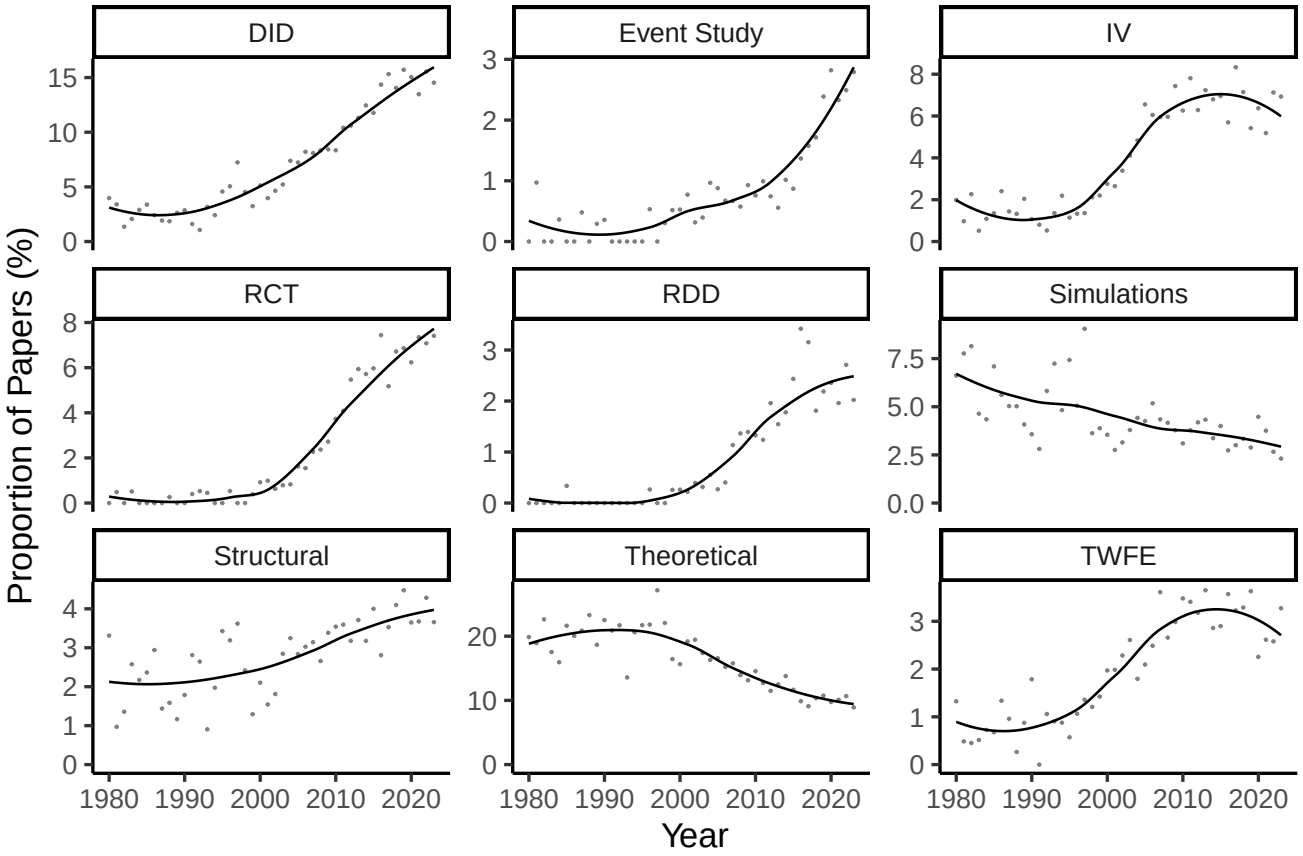


(b) By Field



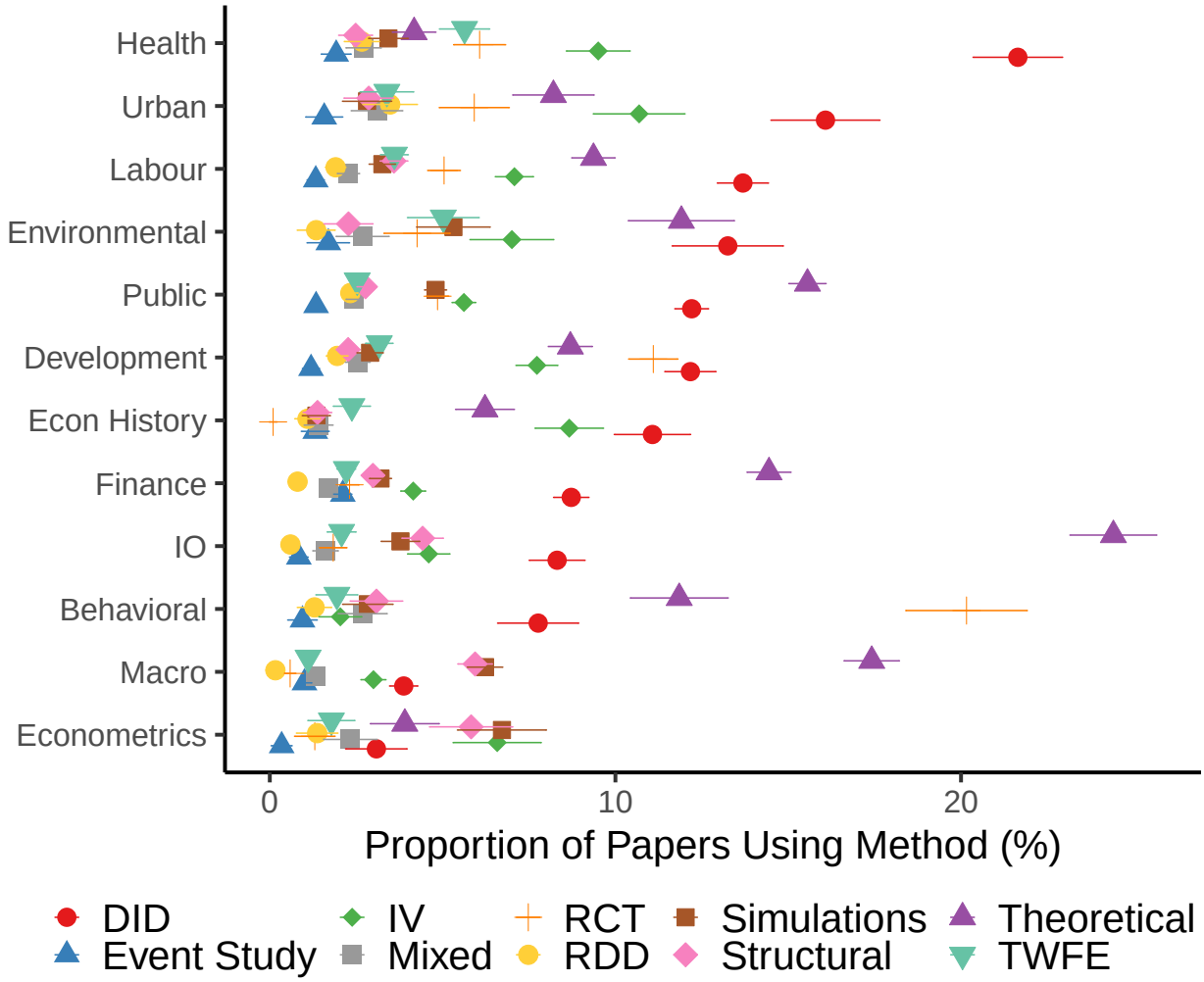
Note: This figure presents the trends and distribution of the average proportion of causal edges per paper in NBER and CEPR working papers across different dimensions. **Panel (a)** displays the average proportion of causal edges per paper from 1980 to 2023, showing a significant increase from approximately 4.2% in 1990 to around 27.8% in 2020. The solid blue line represents the average, and the shaded area indicates the 95% confidence interval. **Panel (b)** shows the average proportion of causal edges by field, comparing the pre-2000 (royal blue) and post-2000 (orange) periods. Most fields exhibit substantial increases in the average proportion of causal edges over time. Fields such as Health, Urban, Development, and Behavioral show the largest increases and highest post-2000 levels. These patterns suggest that the adoption of causal inference methods has become more widespread across various fields in economic research, reflecting the broader impact of the credibility revolution.

Figure 6: Proliferation of Empirical Methods Over Time in NBER and CEPR Working Papers



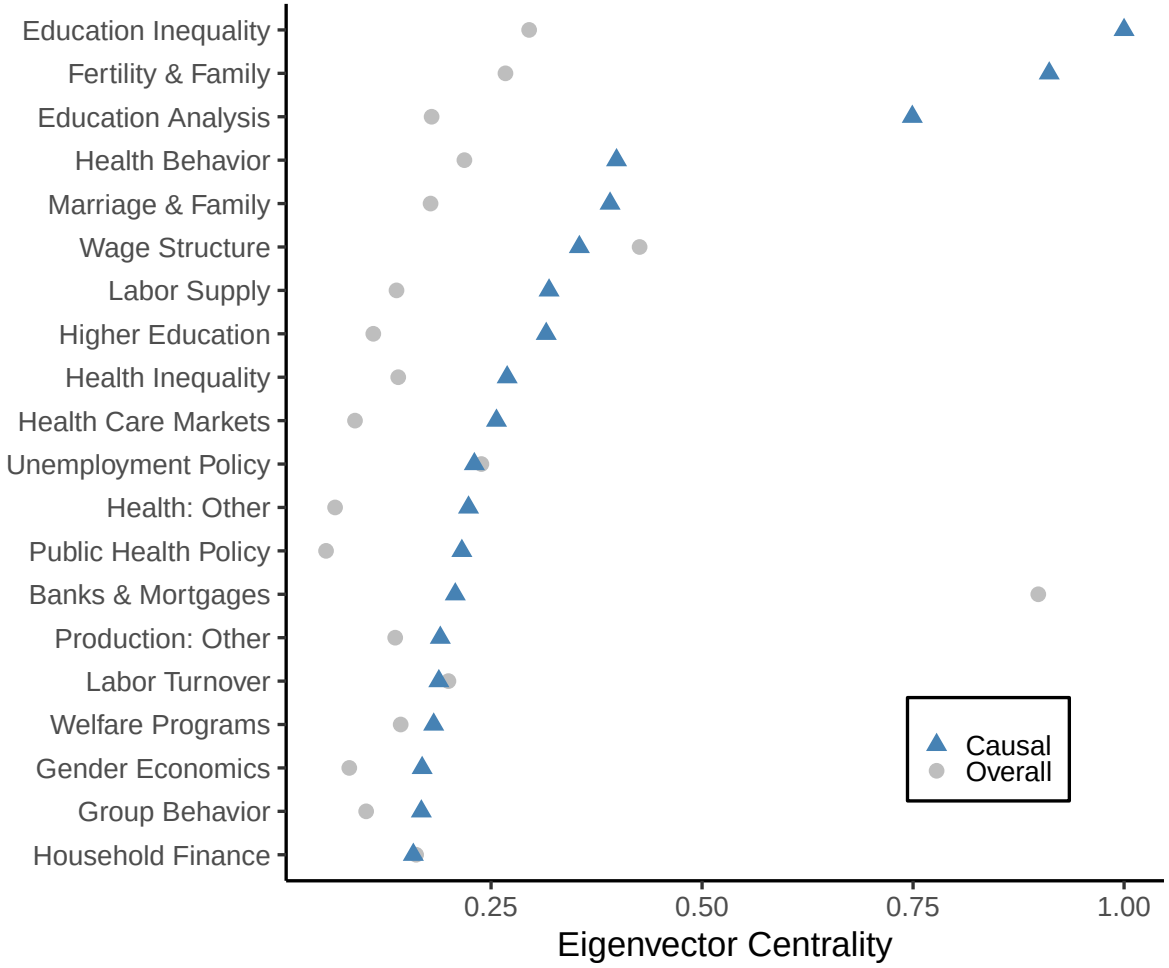
Note: This figure shows the proliferation of key empirical methods used in NBER and CEPR working papers over time: Difference-in-Differences (DiD), Instrumental Variables (IV), Randomized Controlled Trials (RCTs), Regression Discontinuity Design (RDD), Two-Way Fixed Effects (TWFE), Structural Estimation, Event Studies, Simulations, and Theoretical/Non-Empirical research. Each panel represents the proportion of papers utilizing one of these methods per year, with the y-axis showing the proportion of total papers and the x-axis indicating the year of publication. The data covers all NBER and CEPR working papers from 1980 to 2023. DiD has seen a significant increase since the 1980s, rising from around 4% to over 15% of papers in recent years, reflecting its growing importance in empirical research. IV methods have also increased steadily from approximately 2% to over 6% over the same period. RCTs and RDDs, while starting from near zero in the 1980s, have grown to over 7% and 2% respectively in recent years, indicating the rising feasibility and acceptance of experimental and quasi-experimental designs in economics. Conversely, the use of theoretical and non-empirical research has declined significantly, from around 20% in 1980 to under 10% in 2023, suggesting a shift towards empirical analysis in the discipline. The use of simulations has decreased from over 6% in 1980 to around 2–4% in recent years. These trends highlight the increasing emphasis on credible identification strategies and the evolution of empirical methods in economics.

Figure 7: Cross-Sectional Breakdown of Empirical Methods by Field in NBER and CEPR Working Papers



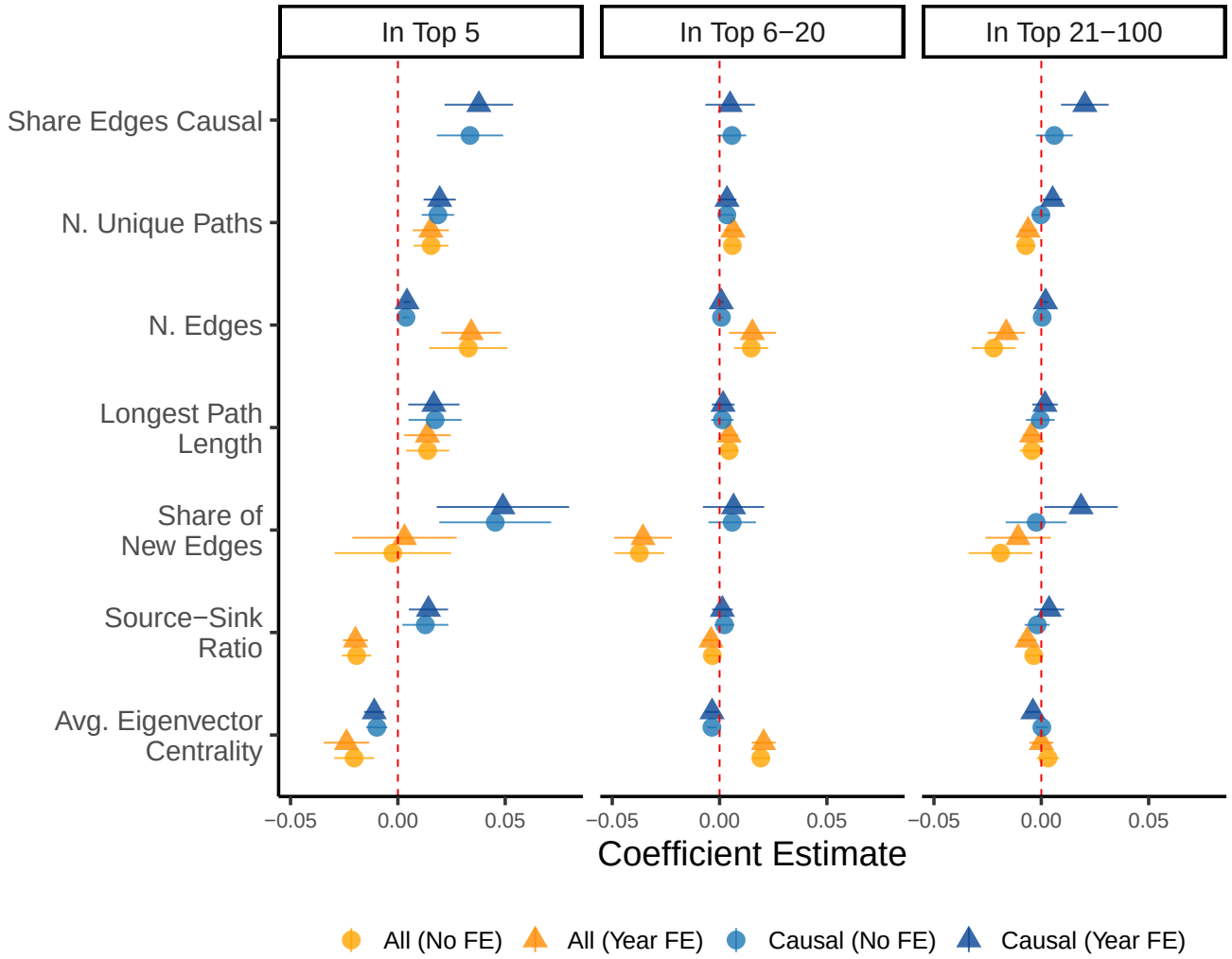
Note: This figure displays the cross-sectional distribution of nine empirical methods—Difference-in-Differences (DiD), Instrumental Variables (IV), Randomized Controlled Trials (RCTs), Regression Discontinuity Design (RDD), Event Studies, Simulations, Structural Estimation, Two-Way Fixed Effects (TWFE), and Theoretical/Non-Empirical research—across twelve fields in NBER and CEPR working papers. Each point represents the proportion of papers within a specific field that utilize a given method, with 95% confidence intervals depicted by error bars. The fields include Finance, Development, Labour, Public, Urban, Macroeconomics, Behavioral, Economic History, Econometrics, IO, Environmental, and Health. The plot highlights considerable variation in the adoption of empirical methods across fields. DiD is most commonly used in Health, Urban, and Labour, with over 21% of papers in Health, over 16% in Urban, and over 13% in Labour utilizing this method. RCTs are particularly prominent in Behavioral and Development, where they are used in over 20% and 11% of papers respectively, reflecting the feasibility of experimental interventions in these areas. Simulations and Structural methods are more prevalent in Macroeconomics and Econometrics, reflecting the need for complex theoretical modeling in these fields. Simulations account for over 6% of papers in Macroeconomics and over 6% in Econometrics. Structural methods are used in approximately 6% of papers in Macroeconomics and over 5% in Econometrics. Fields like Macroeconomics and Finance rely more on IV methods and simulations, with Macroeconomics having around 3% of papers using IV methods and over 6% using simulations. Theoretical and non-empirical research remains significant in fields like Industrial Organization and Macroeconomics, with over 24% and 17% of papers respectively. These cross-sectional patterns reflect the methodological preferences specific to the research questions and data availability in each field, underscoring how different areas of economics adopt various empirical strategies to address their unique challenges.

Figure 8: Top 20 JEL Codes by Eigenvector Centrality in Overall and Causal Knowledge Graphs



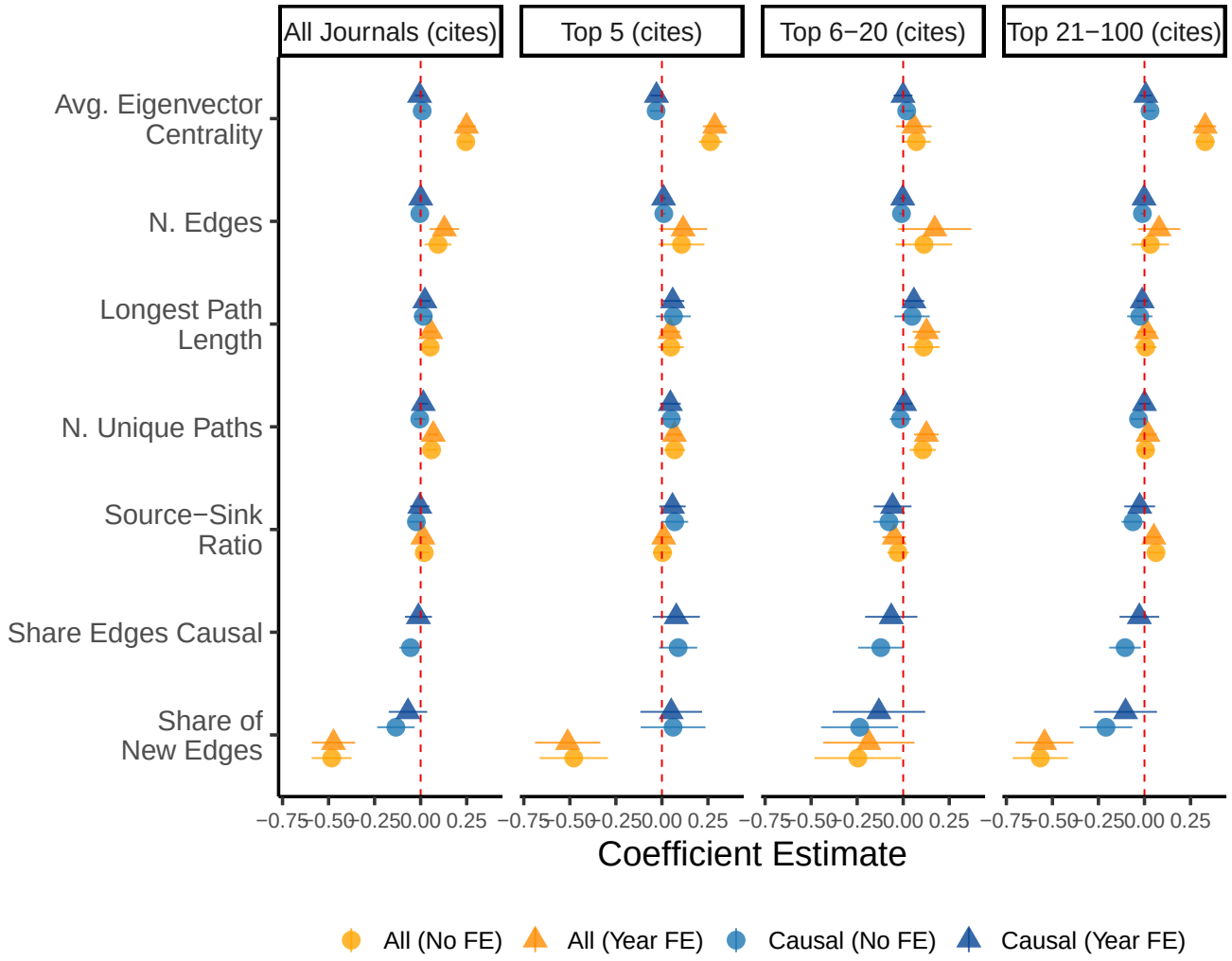
Note: This figure displays the top 20 JEL codes ranked by their eigenvector centrality within both the overall knowledge graph (gray points) and the causal knowledge graph (blue points) constructed from our dataset. Eigenvector centrality measures the influence of a node in a network, with higher scores indicating more central or connected concepts. The comparison reveals that while nodes like **G21** (Banks and Mortgages) and **J31** (Wage Structure) have high centrality in the overall graph, topics such as **I24** (Education and Inequality), **J13** (Fertility and Family), and **I21** (Analysis of Education) are more central in the causal graph. This suggests a shift in focus towards these areas when using causal inference methods in economic research.

Figure 9: Knowledge Graph Measures and Publication Outcomes



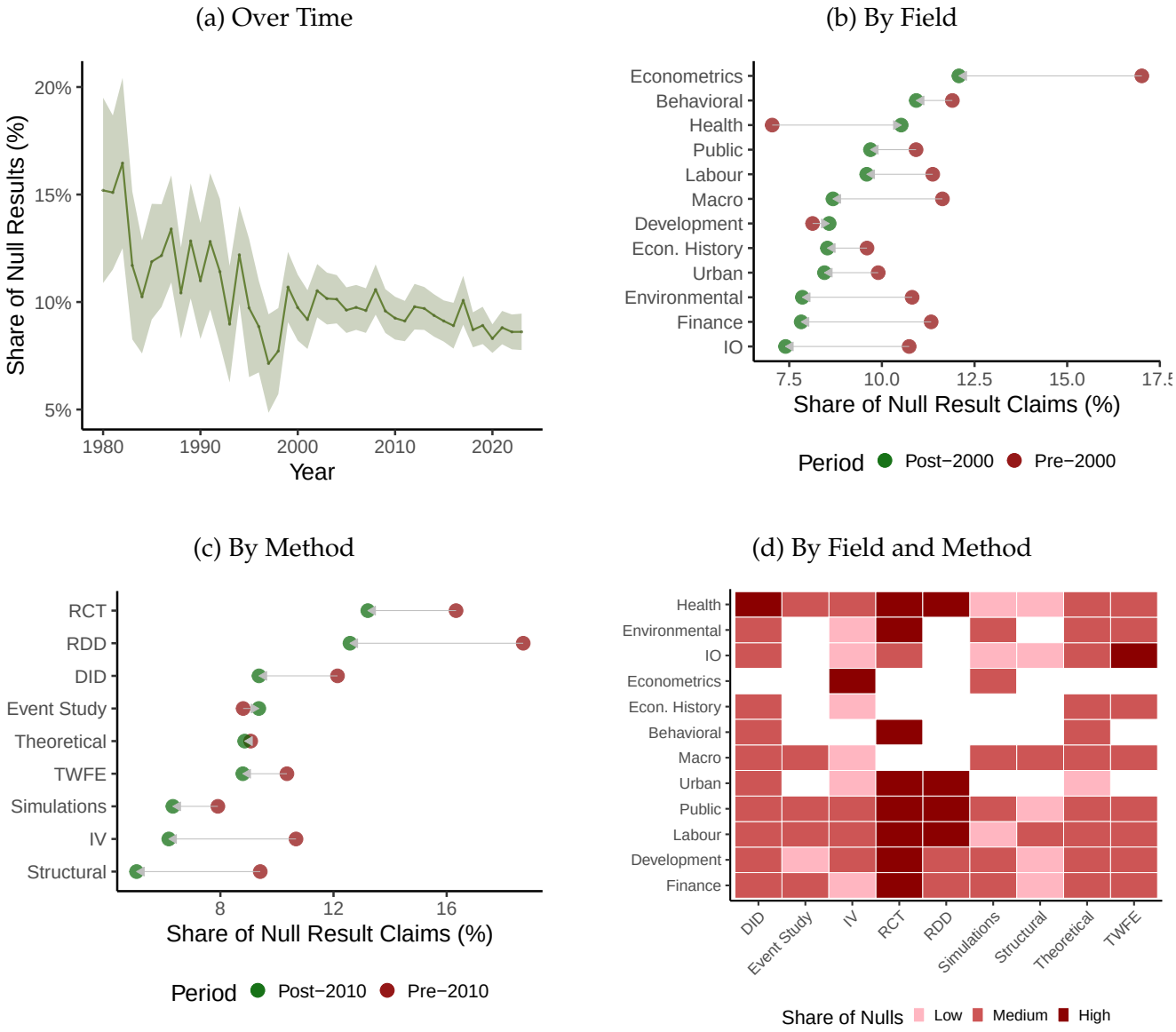
Note: This figure displays coefficient estimates from regression models where the dependent variables are indicators of publication outcomes: publication in a top 5 journal, publication in a top 6–20 journal, and publication in a top 21–100 journal. Each point represents the estimated effect of a knowledge graph measure on the publication outcome, with horizontal lines indicating 95% confidence intervals. The models include specifications with and without year fixed effects. Results are shown for both *All* measures (orange and dark orange points) and *Causal* measures (light blue and dark blue points). Positive coefficients indicate that higher values of the knowledge graph measure are associated with higher likelihood of publication in the respective journal tiers.

Figure 10: Knowledge Graph Measures and Citation Counts by Journal Category



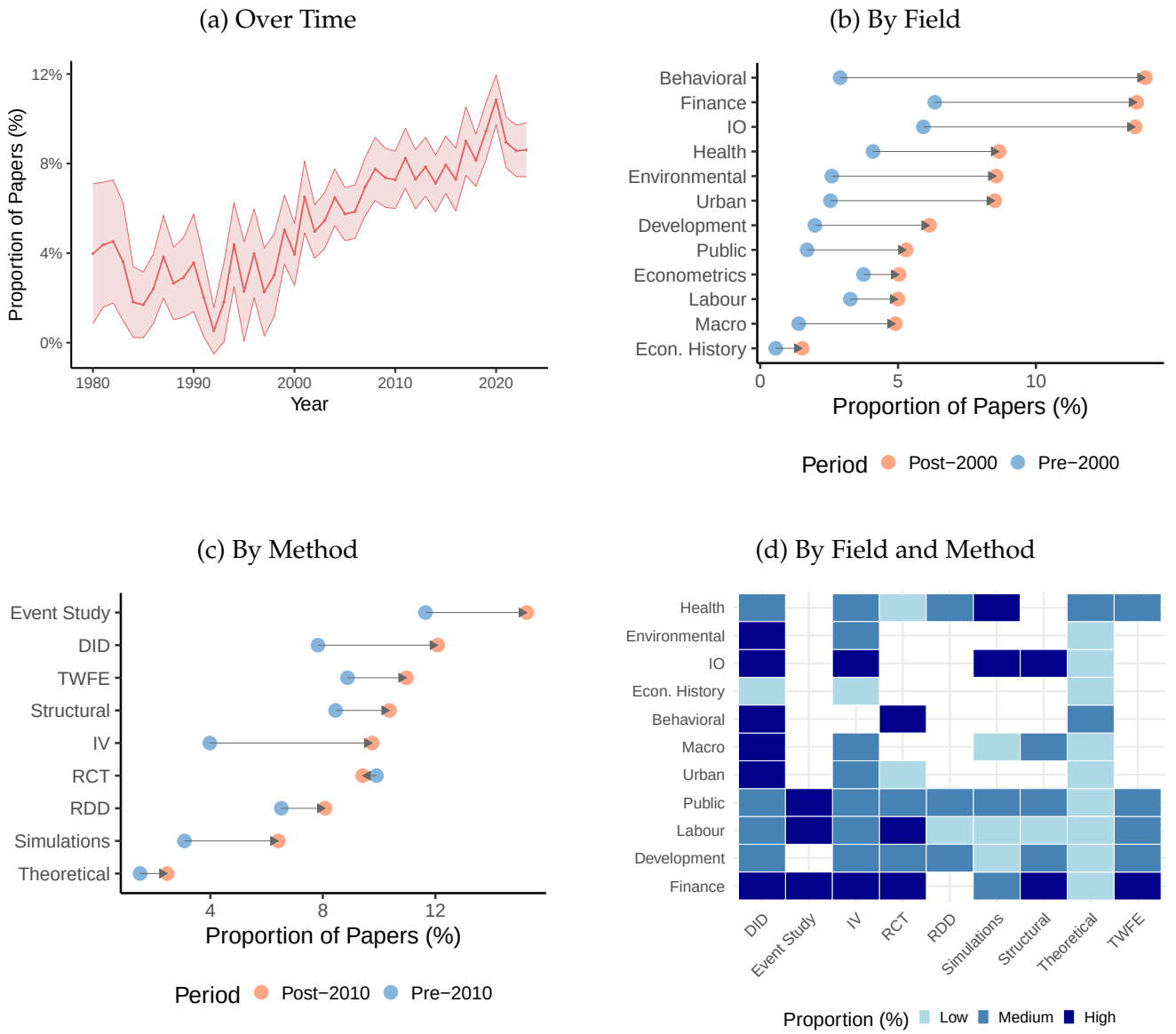
Note: This figure displays coefficient estimates from regression models where the dependent variable is the logarithm of citation counts ($\log(citations + 1)$). Each panel corresponds to a journal category: all journals, Top 5 journals, Top 6–20 journals, and Top 21–100 journals. Points represent estimated effects of knowledge graph measures on citation counts, with horizontal lines indicating 95% confidence intervals. Models include specifications with and without year fixed effects. Results are shown for both *All* measures (orange and dark orange points) and *Causal* measures (light blue and dark blue points). Positive coefficients indicate that higher values of the knowledge graph measure are associated with higher citation counts within the respective journal category.

Figure 11: Null Results in Economic Research Over Time, by Field, and Method



Note: This figure presents the trends and distribution of null results in NBER and CEPR working papers across different dimensions. **Panel (a)** displays the average share of null result claims per paper from 1980 to 2023, showing a decrease from approximately 15% in 1980 to around 8.6% in 2023. The solid blue line represents the average share, and the shaded area indicates the 95% confidence interval. **Panel (b)** shows the average share of null results by field, comparing the pre-2000 (red) and post-2000 (green) periods. Most fields exhibit a decrease in null result reporting over time, with fields like Econometrics and Behavioral maintaining higher shares of null results. **Panel (c)** presents the average share of null results by empirical method, comparing pre-2000 and post-2000 periods. Methods such as RCTs and RDDs report higher shares of null results, while Structural and Theoretical methods show lower shares. **Panel (d)** is a heatmap illustrating the average share of null results by field and method combinations. Darker shades represent higher shares of null results (Low is under 7%, Medium is 7-14%, and High is above 14%. Certain combinations, such as RCTs in Labour and Development, and RDDs in Health, are associated with higher shares of null results. These patterns suggest that both the field and the empirical method influence the likelihood of reporting null findings in economic research.

Figure 12: Trends in Private Data Usage Over Time, by Field, and Method



Note: This figure presents the trends and distribution of private company data usage in NBER and CEPR working papers across different dimensions. **Panel (a)** displays the proportion of papers using private company data from 1980 to 2023, showing an increase from approximately 3.97% in 1980 to around 8.61% in 2023. The solid line represents the average proportion, and the shaded area indicates the 95% confidence interval. **Panel (b)** shows the average proportion of private data usage by field, comparing the pre-2000 (coral) and post-2000 (dark blue) periods. Fields like Finance, IO, and Behavioral exhibit significant increases in private data usage over time. **Panel (c)** presents the average proportion of private data usage by empirical method, comparing the periods pre-2010 (coral) and post-2010 (dark blue). Methods such as Event Studies, DiD, and IV are associated with higher private data usage. **Panel (d)** is a heatmap illustrating the average proportion of private data usage by field and method combinations, categorized into Low (below 4%), Medium (4% to 10%), and High (above 10%). Darker shades represent higher proportions of private data usage. Certain combinations, such as DiD in Behavioral and Finance, and Structural in IO, are associated with higher private data usage. These patterns suggest that both the field and the empirical method influence the reliance on private data in economic research.

Appendix

A Details on LLM-based Information Retrieval

Background on Large Language Models Large Language Models (LLMs) like GPT-4o-mini have significantly advanced natural language processing by enabling machines to understand and generate human-like text. Pre-trained on extensive datasets, including academic papers, books, websites, and other textual sources, these models capture the complexities of language, semantics, and context. This extensive pre-training allows LLMs to perform a variety of tasks, such as text summarization, translation, question answering, and information retrieval (**??**Fetzer et al. 2024, **?**).

In our study, we leverage the LLM's ability to comprehend and extract structured information from unstructured text. By processing the first 30 pages of each economics working paper, the LLM identifies and extracts key metadata, methodological details, and causal claims. Unlike traditional NLP methods that rely on keyword matching or rule-based extraction, LLMs can understand complex language (often found in academic papers) and infer relationships between concepts, making them particularly effective for analyzing complex academic texts.

A.1 Prompt Design and Multi-Stage Extraction Process

Our information retrieval process involves a multi-stage approach, utilizing the LLM to extract and structure the necessary data efficiently.

A.1.1 Prompt Design

To effectively guide the LLM in extracting the required information, we developed comprehensive prompts that included detailed system and user instructions for each stage of the extraction process. The prompts were designed to specify the assistant's role, define the task, provide definitions where relevant, enforce a structured response format, and set guidelines for accuracy and consistency.

Assistant's Role and Task Definition We explicitly defined the assistant's role as an expert in economics paper analysis, specializing in interpreting complex academic content and extracting nuanced information. This role specification was intended to align the LLM's outputs with the expectations of an expert-level analysis. We provided detailed task definitions, instructing the assistant to analyze the text content of a provided economics paper and extract specific information related to metadata, causal claims, research methods, variables, data measurements, identification strategies, and data usage.

Inclusion of Canonical Definitions To ensure accurate and consistent classification of research methods, we included canonical definitions of key methodologies commonly used in economics, such as Randomized Controlled Trials (RCTs), Difference-in-Differences (DiD), Instrumental Variables (IV), Regression Discontinuity Designs (RDD), and others. By providing these standard definitions, we aimed to minimize ambiguity and enhance the reliability of method classification.

Structured Response Format We specified that the assistant's response should adhere to a structured JSON schema with predefined fields and data types. This structured format was crucial for facilitating subsequent data processing, aggregation, and analysis across the large corpus of papers. The JSON schema included detailed definitions and instructions for each field to ensure accurate and consistent extraction. One of the instances where it works the best is to provide an array of causal edges in a paper. Other useful application is specifying the data type: boolean (e.g. whether paper uses any data), categorical (e.g., source of data: private or public sector), numeric (e.g., number of authors) or string response (e.g., name of the data provider).

Guidelines for Accuracy and Consistency We emphasized strict adherence to the schema, accuracy, and clarity in the assistant's responses. We instructed the assistant to use the canonical definitions when classifying methods and to ensure that all required fields were accurately and completely filled out. These guidelines were essential to maintain the quality and consistency of the extracted data.

A.1.2 Overview of Prompts and Instructions

The following provides an overview of the system instructions provided to the LLM for each stage of the extraction process. Due to space constraints, we present summarized versions of the prompts. The complete system prompts and JSON schemas will be included in the replication package accompanying this paper.

Stage 1: Curated Summary In Stage 1, the assistant was instructed to analyze the first 30 pages of the paper and extract a curated summary of key elements. The system instructions included:

- **Assistant's Role**: An expert assistant specializing in analyzing economics papers.
- **Task**: Extract specific information related to research questions, causal identification strategies, data usage, data accessibility, acknowledgements, and metadata.
- **Guidelines**: Provide clear, detailed, and information-rich responses for each field. Focus exclusively on specified sections. Adhere strictly to specified formats and instructions.
- **Fields to Extract**: Research questions from the abstract, introduction, and full text; causal identification information; causal claims; framing and policy implications; data and units of analysis; data accessibility; institutional and author-level information; acknowledgements.

Stage 2: Causal Graph Retrieval In Stage 2, the assistant was tasked with extracting detailed causal relationships from the summaries provided in Stage 1. The system instructions included:

- **Assistant's Role**: An expert assistant specializing in analyzing economics papers.
- **Task**: Extract an exhaustive list of causal relationships from provided texts, capturing all intermediate steps, mediators, confounders, and other relevant nodes in the causal graph (DAG).
- **Guidelines**: Be exhaustive in extraction. Use exact terms from the authors. Provide each causal relationship in a structured format with specified fields.

- **Fields to Extract**: For each causal relationship, include the causal claim, cause, effect, type of causal relationship, whether evidence is provided, sign of impact, effect size, statistical significance, causal inference method, sources of exogenous variation, level of tentativeness.

Stage 3: Data and Accessibility In Stage 3, the assistant was instructed to extract structured information regarding data sources and accessibility from the data-related summaries in Stage 1. The system instructions included:

- **Assistant's Role**: An expert assistant specializing in analyzing economics papers.
- **Task**: Extract specific pieces of information related to data and units of analysis, and data accessibility.
- **Guidelines**: Carefully extract information for each field, adhering strictly to definitions and instructions. Use exact wording from the text when possible.
- **Fields to Extract**: Data usage indicators, total number of observations, units of analysis, data granularity, temporal and geographical context, data ownership, data accessibility, ethical considerations.

B Matching LLM Output to JEL Codes and OpenAlex Topics

After the LLM extracted the causal claims and provided free-text descriptions of the source and sink variables, we needed to standardize these variables to facilitate aggregation and systematic analysis across the corpus. This standardization was achieved by mapping the variable descriptions to official Journal of Economic Literature (JEL) codes and OpenAlex Topics.

B.0.1 Choice of Standardization Methods

We considered several options for standardizing the terms used in the source and sink variables, including JEL codes, OpenAlex Topics, Concepts, and the OECD Glossary of Statistical

Terms. Each option had its advantages and limitations. JEL codes are well-understood within the economics community and facilitate interpretability but are relatively broad in classification. OpenAlex Topics offer more granularity with around 4,500 topics but are less familiar to economists. Concepts provide even more detail with approximately 60,000 terms but are being deprecated by OpenAlex and are not widely recognized in economics. The OECD Glossary contains about 6,700 economics-related terms but may be biased towards statistical concepts.

After careful consideration, we opted to focus on JEL codes for standardization. JEL codes were chosen as our primary method due to their familiarity and acceptance within the economics discipline, facilitating interpretation and communication of results.[26]

B.0.2 Embedding-Based Matching Methodology

To map the free-text variable descriptions to the standardized codes, we employed an embedding-based matching approach using vector representations of the texts. We utilized the OpenAI text embedding model `text-embedding-3 large`, which generates 1,024-dimensional vector embeddings that capture the semantic meaning of the text. Embeddings were generated for: (i) the free-text descriptions of the source and sink variables extracted by the LLM, (ii) the official descriptions of JEL codes, enhanced by concatenating descriptions from higher-level codes to provide richer context, and (iii) the summaries of OpenAlex Topics.

We calculated the cosine similarity between the embeddings of the variable descriptions and the embeddings of the JEL codes and OpenAlex Topics.[27] We assigned the variable to the codes/topics with the highest similarity scores.

B.0.3 Advantages and Limitations of the Embedding-Based Approach

By leveraging embeddings, we moved beyond simple keyword matching, which can be limited by variations in terminology and phrasing. The embedding-based approach captures the

[26] OpenAlex Topics could also be used as a supplementary method, providing additional granularity and capturing interdisciplinary aspects not covered by JEL codes. OpenAlex Topics will be useful when scaling to research beyond economics.

[27] Cosine similarity measures the cosine of the angle between two vectors, providing a value between -1 and 1, where higher values indicate greater similarity.

semantic meaning of the texts, allowing us to match variable descriptions to standardized concepts even when different terms are used to describe similar ideas (e.g., unemployment rate" vs. joblessness"). This method enhanced the robustness of our matching process, reducing the impact of typos, synonyms, and variations in language. It allowed us to systematically standardize a large number of variable descriptions across the corpus, facilitating cross-paper comparisons and aggregations.

While the embedding-based matching approach offers significant advantages, there are limitations to consider. The quality of the matches depends on the accuracy of the embeddings and the chosen similarity threshold, which in our case is the one with highest similarity. By focusing on only one matching concept, we are imposing a structure on the latent causal graph. It may well be that a source or a sink could be captured by multiple JEL codes. However, for simplicity and consistency across papers, we decided to capture only the best match. Future research can consider setting a threshold, e.g., 0.85, beyond which multiple JEL codes can be accepted. However, this comes with its own set of hyper-parameter selection: there is a trade-off between precision and recall; a higher threshold increases precision but may miss relevant matches, while a lower threshold increases recall but may include irrelevant matches.[28]

By using both JEL codes and OpenAlex Topics, we leveraged the strengths of each system, with JEL codes providing familiarity and OpenAlex Topics offering granularity. Overall, the use of embeddings was instrumental in standardizing and operationalizing our source and sink variables.

B.1 Validation and Quality Assurance

To ensure the reliability of the extracted data, we implemented validation checks at each stage. The structured outputs were parsed and checked for compliance with the specified schemas. In cases where inconsistencies or missing data were detected, the prompts were refined, and the extraction was repeated.

We also conducted manual reviews of a sample of the extracted data to assess accuracy. This

[28] One application of having one-to-many-mappings as a result of a similarity threshold is a study of growing interdisciplinary nature of economics research.

included checking the correctness of the causal claims extracted, the appropriateness of the mapped JEL codes, and the consistency of metadata. The feedback from these reviews informed further refinements to the prompts and extraction process.

A large scale validation of our approach will follow in future drafts. This includes contacting corresponding authors to validate the causal graphs in their own papers.

B.2 Limitations and Considerations

While the use of LLMs provides significant advantages in processing large volumes of complex text, there are limitations to consider. The LLM's extraction is dependent on the quality and clarity of the original text; ambiguities or omissions in the papers may lead to incomplete extraction. The LLM may occasionally misclassify or misinterpret information, particularly with nuanced methodological details. While we collected additional attributes such as effect sizes and statistical significance, these features were experimental and not used in the main analysis due to variation in reporting standards.

Additionally, the embedding-based matching approach for standardizing variables may not capture all nuances of the economic concepts involved. There is a risk of misclassification if the variable descriptions are ambiguous or if the embeddings do not accurately represent the semantic content. Despite these limitations, we believe that the overall methodology provides a robust framework for large-scale analysis of economic research.

B.3 Replication Package

The full system prompts, JSON schemas, and code used in the extraction process will be included in the replication package accompanying this paper. Researchers interested in replicating or extending our analysis can refer to these materials for detailed guidance.[29]

[29] For early access, please contact the authors at prashant.garg@imperial.ac.uk.

C Validating information retrieval

C.1 Validation with the Brodeur et al. (2024) Dataset

To validate the accuracy of our AI-based information retrieval methods, we compared our dataset with an external dataset from Brodeur et al. (2024), which examines p-hacking, data type, and data-sharing policies in economics. Their dataset includes 1,106 articles published in leading economics journals between 2002 and 2020, with detailed annotations on the empirical methods used and the fields of study.

We matched our dataset with theirs using paper titles, employing both direct matches and fuzzy matching techniques to maximize the number of matched papers. In total, we matched 307 papers between the two datasets. For these matched papers, we compared our classifications of empirical methods—Difference-in-Differences (DiD), Randomized Controlled Trials (RCT), Regression Discontinuity Design (RDD), Instrumental Variables (IV)—and fields—Urban Economics, Finance, Macroeconomics, Development—with those from Brodeur et al. (2024) to assess the accuracy of our information retrieval.

We calculated two evaluation metrics: *Accuracy* and the *F1 Score*, where the ground truth labels are those from Brodeur et al. (2024). The results are presented in Table A2.

The results indicate that our information retrieval methods achieve high levels of accuracy and F1 scores in identifying both empirical methods and fields of study. In particular, the identification of RDD methods and classification of papers in Macroeconomics and Urban Economics show excellent performance, with accuracy and F1 scores exceeding 0.90. These findings provide confidence in the reliability of our AI-based extraction and classification processes in at least two dimensions of our analysis: fields and methods.

C.2 Validation with the Plausibly Exogenous Galore Dataset

To further validate our information retrieval methods, we compared our data on causal claims with an external dataset, the Plausibly Exogenous Galore dataset[30]. This dataset includes entries

[30]The Plausibly Exogenous Galore dataset is a curated list of plausibly exogenous variations in the empirical economics literature, maintained by Sangmin S. Oh. Available at https://www.notion.so/

for 1,435 papers (as of August 2024), each with the main Left-Hand Side (LHS) and Right-Hand Side (RHS) variables, as well as the primary source of exogenous variation, as identified by authors and contributors. Unlike our dataset, which captures causal claims at the paper-claim level, the Plausibly Exogenous Galore dataset records information at the paper level, often focusing on key variables rather than the complete knowledge graph of causes, effects, and sources.

We matched 485 papers from our dataset to entries in the Plausibly Exogenous Galore dataset using exact and fuzzy title matching. To facilitate a meaningful comparison, we aggregated our data at the paper level by concatenating all causes, effects, and sources of exogenous variation. This resulted in structured phrases: "The causes are: <cause>; <cause>; ..." for causes, "The effects are: <effect>; <effect>; ..." for effects, and "The source(s) of exogenous variation are: <source>; <source>; ..." for sources. Similarly, the Plausibly Exogenous Galore data was formatted as "The cause is: <RHS>" for the RHS (cause), "The effect is: <LHS>" for the LHS (effect), and "The source(s) of exogenous variation are: <source of exogenous variation>."

Using OpenAI's `text-embedding-3-large` embeddings model, we generated embeddings for each component separately (causes, effects, and sources of exogenous variation) and calculated cosine similarity scores for each component between our data and the Plausibly Exogenous Galore dataset. We computed the average, minimum, maximum, median, and standard deviation of these similarities.

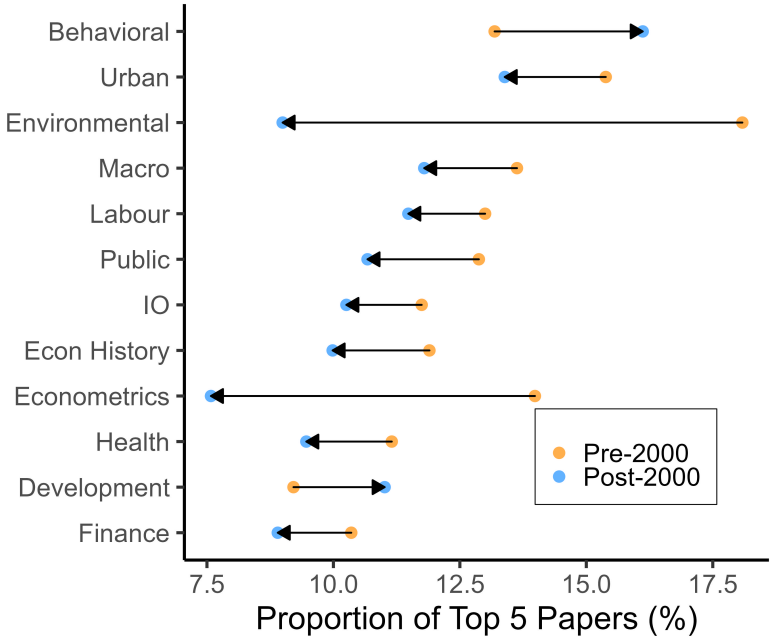
The results (Table A3) show that while cause and effect similarities were moderate, reflecting variability in how these are represented across claims within a paper, the source of exogenous variation achieved a higher similarity score. This result supports the reliability of our method for capturing consistent exogenous variation sources, as these tend to vary less across claims. These findings reinforce our information retrieval system's ability to align with external validations, particularly for core causal identifiers.

`1a897b8106ca44eeaf31dcd5ae5a61b1?v=ff7dc75862c6427eb4243e91836e077e`.

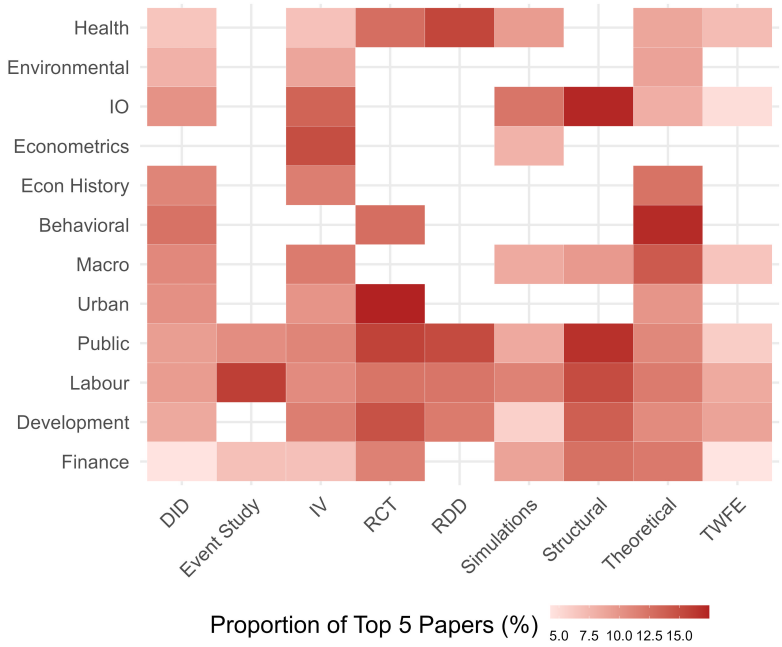
D Appendix Tables and Figures

Figure A1: Proportion of Papers Published in Top 5 Journals, Pre- and Post-2000.

(a) By Field

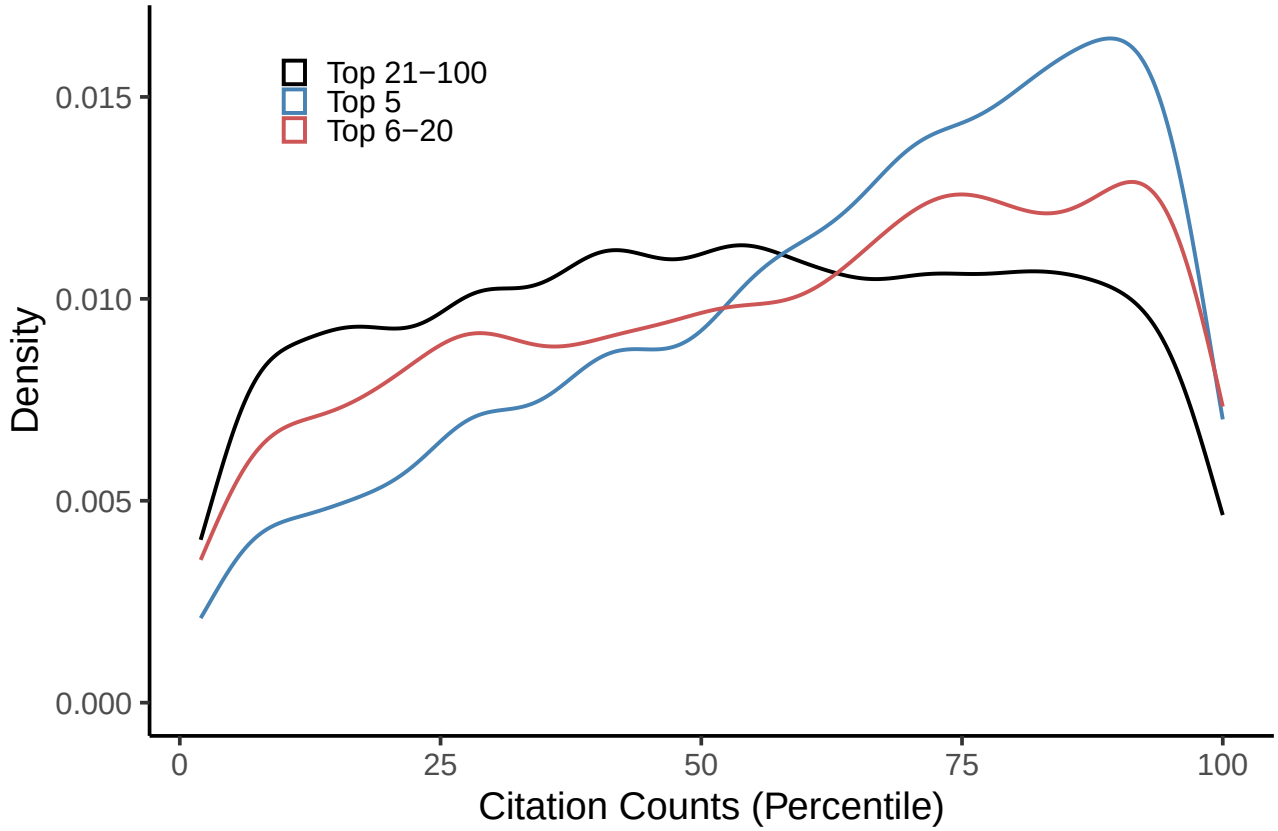


(b) By Field and Method



Note: This figure displays the proportion of working papers that were eventually published in the top five economics journals, broken down by (a) field and (b) field by empirical method. Data are derived from our matched publication dataset. Arrows in panel (a) indicate changes between pre-2000 and post-2000 periods. The heatmap in panel (b) shows that certain fields and methods have higher publication rates in top journals, with field-method combinations such as Theoretical methods in Behavioural, Structural in IO or RCTs in Urban.

Figure A2: Distribution of Citation Percentiles by Journal Category



Note: This figure displays kernel density plots of the citation percentiles for papers published in Top 5, Top 6–20, and Top 21–100 journals. The citation percentiles are calculated based on the entire sample, with higher values indicating higher citation counts relative to other papers. The plot shows that while papers published in higher-ranked journals tend to receive more citations on average, the most highly cited papers are more evenly distributed across journal categories. This suggests that exceptionally influential papers can emerge from a wide range of journals.

Table A1: Summary Statistics of Citation Percentiles by Journal Category

Journal Category	Mean Percentile	Median Percentile	Standard Deviation
Top 5	62.18	68	25.86
Top 6–20	57.06	61	27.97
Top 21–100	52.12	53	27.37

Note: This table summarizes the citation percentiles for papers published in different journal categories. The mean and median percentiles indicate that papers published in Top 5 journals have higher citation impact on average compared to those in lower-ranked journals. However, the overlap in distributions suggests that highly cited papers can also be found in lower-ranked journals.

Table A2: Validation Results of Information Retrieval

Variable	Accuracy	F1 Score
Method: DiD	0.7762	0.8447
Method: RCT	0.7063	0.8269
Method: RDD	0.9371	0.9644
Method: IV	0.7413	0.8183
Field: Urban Economics	0.9138	0.9545
Field: Finance	0.8788	0.9295
Field: Macroeconomics	0.9744	0.9870
Field: Development	0.6643	0.7937

Note: This table presents the validation results of our information retrieval methods compared to the ground truth provided by Brodeur et al. (2024). **Variable** indicates either the empirical method (DiD, RCT, RDD, IV) or the field of study (Urban Economics, Finance, Macroeconomics, Development). **Accuracy** measures the proportion of correctly classified instances, while **F1 Score** represents the harmonic mean of precision and recall. Higher values indicate better performance.

Table A3: Validation Results with Plausibly Exogenous Galore Dataset

Variable	Mean Similarity	Median	Std Dev	Min	Max
Cause Similarity	0.6140	0.6245	0.1127	0.2285	0.8798
Effect Similarity	0.6386	0.6467	0.0893	0.3795	0.8452
Exogenous Variation Similarity	0.8014	0.8142	0.0975	0.4920	0.9863

Note: This table presents the cosine similarity results between our dataset and the Plausibly Exogenous Galore dataset, which records plausibly exogenous sources, main causal factors, and outcomes in the economics literature. **Mean Similarity** indicates the average cosine similarity score, while **Std Dev** shows the standard deviation of these scores, capturing the variability. **Min** and **Max** represent the range of similarity scores across matched papers.