What Should Economics Ask Next?

A graph-based screening benchmark for candidate questions in economics

Prashant Garg
Imperial College London

17 March 2026
Draft for comments. Please do not cite without permission.

Abstract

As AI makes drafting, coding, and reviewing cheaper, choosing the right research question may become a more important bottleneck. This paper studies that problem in economics by representing plausible next questions as missing links in a directed map of the literature and ranking them using patterns in the surrounding research neighborhood. I build this map from over 240,000 papers in core economics and adjacent journals published over the last 50 years (1976 to early 2026), and test whether the ranking predicts connections that later emerge in the literature. A simple baseline that favors already well-connected topics performs better at the very top of the ranking, but the graph-based score becomes more useful once researchers look beyond just the top few suggestions, especially in adjacent journals, design-based causal work, and areas of the literature with many nearby pathways. The contribution is practical rather than universal: the method helps surface plausible next questions and helps researchers see why each suggestion is being made.

Keywords: research allocation, economics of science, knowledge graphs, preferential attachment, AI-assisted discovery

1 Introduction

Choosing what to work on is one of the least formalized decisions in economics. We have disciplined frameworks for identification, estimation, and inference, but much less for the upstream choice of which question deserves scarce attention in the first place. Bloom et al. (2020) argue that ideas are getting harder to find, while Jones (2009) emphasizes the growing knowledge burden faced by new researchers. Those arguments point in the same direction: the frontier becomes harder to navigate even as the stock of published work keeps growing.

That problem becomes sharper, not weaker, when AI lowers the cost of adjacent research tasks such as drafting, review assistance, and iterative revision. If downstream paper-production tasks become cheaper,

the bottleneck shifts upstream toward question choice. The question is no longer only how to write or review a paper more cheaply. It is how to decide which question deserves attention next.

This paper studies that narrower empirical problem. It does not answer the welfare question of what economics ought to study in the abstract. It asks whether the structure of past research can help surface plausible next questions, and whether those suggestions can be made clear enough to inspect and use. I start by building a map of how topics connect across papers. In that map, possible next questions appear as connections that have not yet been made. Formally, I represent those as missing directed links in a literature map built from economics and adjacent journals. Suppose the literature already contains links such as public debt → public investment and public investment → CO_2 emissions, but the direct relation public debt → CO_2 emissions has not yet appeared. That missing direct connection is a concrete candidate question. The closest computational analogy is link prediction, but the object here is narrower and more interpretable than generic network completion. In this paper, "should" is therefore used in that narrower operational sense. A question deserves attention next when it is neglected enough to remain open, supported enough to be credible, concrete enough to become a paper, and worth reading under a realistic shortlist budget rather than only at a winner-take-all top rank. The website at frontiergraph.com lets readers inspect those surfaced questions and the nearby evidence behind them.

That map-based view is useful because it preserves more of the local logic of scientific development than keyword overlap or raw citation counts alone. It lets us see whether a putative question is supported by nearby chains of papers and topics, and only then name those features more formally as paths, motifs, and local graph structure. The framework is intentionally modest. It is a discovery aid, not proof of importance; a prospective ranking exercise, not a welfare theorem; and a graph of extracted claim relations, not a full adjudication of causal truth from complete papers.

The empirical design starts from a field-weighted citation impact selected corpus of top core and adjacent journals. The selected sample contains 242,595 papers spanning 1976 to early 2026, of which 230,929 contain at least one extracted edge and 230,479 survive into the normalized graph used in evaluation. I build that graph from the paper-level extraction framework in Garg and Fetzer (2025), then distinguish between directed causal links and undirected contextual support inside a single graph object. Missing directed links are ranked by a graph-based score built from path support, underexploration gaps, motif support, and hub penalties. I then freeze the graph at year $t - 1$, rank candidates, and test whether those links first appear over 3-, 5-, 10-, and 15-year horizons.

The headline result is mixed and therefore informative. The toughest benchmark is a simple rule that favors topics that are already well connected. In network terms, that rule is preferential attachment, and it still wins in the pooled rolling benchmark at very tight shortlists. In concrete terms, a 100-question shortlist built from that benchmark retrieves roughly 2.6, 3.3, 7.0, and 10.0 more realized directed links than the graph score at $h = 3, 5, 10, 15$. But that is not the end of the story. Once the shortlist widens, so that a researcher is willing to inspect more than just the top few suggestions, the graph-based score becomes more competitive. The newer heterogeneity results also suggest that pooled averages hide meaningful variation across journals, methods, and parts of the literature. A separate path-development exercise points to a

second pattern: research often builds mediating structure around existing direct claims more often than it closes a direct link already implied by local paths.

The paper makes three contributions. First, it proposes a way to rank plausible next questions using the nearby structure of the literature; I treat the resulting object as a benchmarked screening problem in a directed literature graph. Second, it shows where that nearby structure helps more, and how research often develops by adding mediator paths around existing direct claims rather than only closing missing direct links. Third, it makes the object inspectable through a public browser that exposes suggested questions, nearby topics, supporting paths, and starter papers.

These findings connect the paper to several literatures at once. They speak to the economics of ideas and scientific search (Bloom et al., 2020; Jones, 2009), to the science-of-science literature on novelty, impact, and frontier tracing (Uzzi et al., 2013; Fortunato et al., 2018; Wang and Barabasi, 2021), and to current work on AI-assisted scientific workflows and discovery systems (Zhang et al., 2025; Shao et al., 2025). They also speak to a practical question. The public system at frontiergraph.com is meant to help researchers inspect why one candidate question surfaced, what local paths support it, and which nearby literatures are doing the work.

2 Related Literature and Positioning

This paper sits at the intersection of four literatures.

First, it belongs to the economics of ideas and discovery. Bloom et al. (2020) document rising research effort alongside falling research productivity across several domains, while Jones (2009) studies how accumulating knowledge changes the organization of innovative activity. Those papers focus on the production of ideas and the cost of reaching the frontier. The present paper shifts attention to a narrower but operationally central problem: given a large existing literature, how should one screen candidate next questions?

Second, the paper draws on the science-of-science literature that uses large-scale scientific data to study novelty, impact, and frontier formation. Fortunato et al. (2018) provide a broad synthesis. Wang and Barabasi (2021) show how scientific frontiers can be studied quantitatively, while Uzzi et al. (2013) show how novelty often combines conventional structure with a limited number of atypical combinations. That literature is highly relevant in spirit, but my object differs. I do not measure novelty from citations or reference-pair combinations. I define candidate questions as missing links in a claim graph and evaluate them prospectively.

Third, the benchmark logic comes from network growth and cumulative advantage. Price (1976) and Barabasi and Albert (1999) show why already connected nodes tend to attract more links. For this project, that is not a decorative comparison. It is the main null. If the future of the literature is mostly a popularity process, then a rich-get-richer rule should perform well when the target is future edge appearance. Preferential attachment is therefore not a straw man. It is a serious benchmark because it encodes one plausible

model of how scientific attention moves.

Fourth, the paper enters a fast-moving discussion around AI-assisted scientific work. Recent systems already help with tasks such as hypothesis generation, literature synthesis, manuscript feedback, and reviewer assistance (Zhang et al., 2025; Shao et al., 2025; Refine, 2026; ICLR, 2026; Stanford Agentic Reviewer, 2026; Project APE, 2026). My contribution is not a new general-purpose AI assistant and not a new claim-extraction model. The paper uses AI-extracted paper-level structure as an enabling layer, then asks an economics question: can we convert that structure into an inspectable, prospectively testable research-allocation object?

That positioning also helps distinguish this project from generic link prediction. Generic link-prediction work often asks whether a missing edge will appear in a network. Here the edges are substantively interpreted claim-like relations in economics, the benchmark is explicitly framed around scarce reading budgets, the principal comparison is with cumulative advantage, and the public output is designed to be inspectable question by question. The paper is therefore best understood as economics-first metascience with a graph-based empirical object, rather than as a pure machine-learning exercise.

3 Corpus, Paper-Local Extraction, and Node Normalization

The paper starts from a published-journal corpus rather than a broad scrape of all economics-adjacent writing. The selected journal corpus contains 242,595 papers drawn from the top 150 core economics journals and the top 150 adjacent journals under the field-weighted citation impact selection rule. The sample spans 1976 to early 2026. Of those papers, 230,929 contain at least one extracted edge, yielding 1,443,407 raw extracted edges. After normalization and graph construction, the evaluation graph retains 230,479 papers, 6,752 concept codes, and 1,271,014 normalized links.

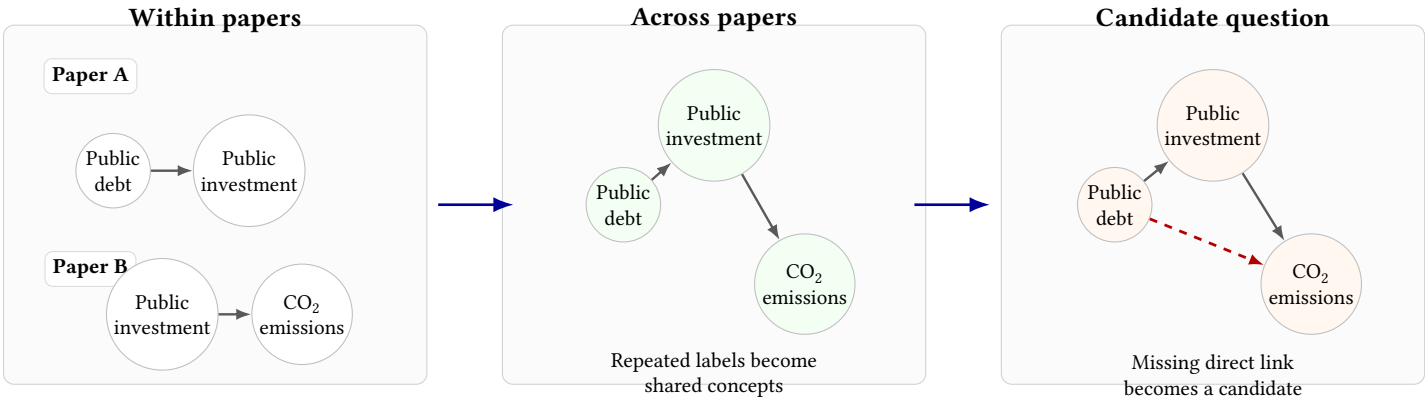
3.1 Corpus definition

The paper uses the published-journal corpus because the goal is to study realized scientific structure in an economics-facing literature, not to optimize coverage of working papers, drafts, or preprints. This choice is conservative. It sacrifices some freshness in exchange for clearer source control, more stable metadata, and a graph that is easier to interpret as realized economics research rather than a noisy mix of partially filtered text. Work-level metadata, source metadata, and journal assignments come from OpenAlex.[1]

The journal universe is deliberately narrow in two senses. First, it uses a selected journal universe rather than a universal bibliographic crawl. Second, it uses the FWCI-based selection rule so that the retained sample is tied to a published-literature core with credible economics or adjacent relevance. This is not meant to claim that the retained corpus exhausts the frontier. It is meant to define a reproducible, interpretable literature against which missing links can be backtested. Appendix C summarizes the resulting

[1] OpenAlex enters the paper as the bibliographic and journal-metadata layer. The extraction, normalization, and ranking steps are built on top of that source rather than inherited from it.

Figure 1: Paper text, local extraction, and concept matching



Notes. This figure shows the core measurement pipeline as a worked example. The unit of observation at extraction is the individual paper title and abstract. Each paper first produces a paper-local graph, repeated concept labels are then matched into shared concept identities across papers, and the resulting concept-level structure can surface a missing direct relation as a candidate question. Directed causal relations are stored for the design-based causal task, while noncausal contextual support remains in the same concept graph as undirected structure.

corpus waterfall and the retention numbers.

3.2 Paper-local research graphs

The extraction layer builds on Garg and Fetzer (2025). Each title and abstract is converted into a paper-local graph in which nodes correspond to extracted concepts and edges summarize the relations the paper itself states, studies, or reports. The present paper inherits that idea but extends it in three ways that matter downstream. First, the schema is broader than explicit causal claims, because the benchmark also needs undirected contextual support. Second, the schema separates the paper's *causal presentation* from the *evidence method* used to support a claim. Third, the local graph stores contextual qualifiers in dedicated fields rather than forcing them into the node label. The goal is not simply to recover whether a paper makes a claim; it is to recover enough structured local information that the claim can later live inside a reusable concept graph. Code, prompt files, and release materials are available in the public repository at `https://github.com/prashgarg/frontiergraph`.

The design choice to work paper-locally first is deliberate. At extraction time, the task is not to solve global ontology matching inside the language model. It is to recover the paper's own internal concept reuse faithfully and to prevent the model from inventing relations that are only implied by transitivity. Exact prompts, the full schema, and the design logic are reported in Appendix A. **The code and prompt files used in this paper will be released at `https://github.com/prashgarg/frontiergraph`.**

Table 1: Core notation used in the paper

Symbol	Definition
$G_{t-1} = (V, E_{t-1})$	Claim graph assembled from papers observed through year $t-1$. It contains directed causal links and undirected contextual support inside one concept-level graph object.
u, v, w	Normalized concept nodes in the ontology-backed graph.
$u \to v$	Directed causal link or directed causal candidate.
$\{u, v\}$	Undirected noncausal pair.
h	Evaluation horizon in years.
K	Shortlist size in the fixed-budget retrieval problem.

3.3 Node normalization and concept identity

The normalization problem is central in this paper because candidate generation, path counts, gap measures, and missingness all depend on node identity. Paper-local concept strings vary in wording, scope, and granularity. "Inflation in Germany", "inflation", and "German inflation" should not automatically remain three separate graph nodes if the downstream object is a reusable concept graph. But they should not be merged carelessly either.

For that reason the paper builds a native concept ontology. The aim is to preserve concept identity at the level at which candidate questions are actually formed. The pipeline first resolves easy cases with deterministic lexical signatures, then uses a text-embedding matching step to rank plausible concept matches for harder cases, and finally stores mapping provenance and quality bands so weaker tail recoveries remain auditable.[2] The ontology build creates native concept codes, clusters a selected head pool into accepted concepts, applies hard and soft mappings for the tail, and then uses a force-mapped recovery layer so the fuller corpus remains in the graph rather than being dropped for lack of a clean early match. Appendix B gives the full algorithmic detail.

4 Candidate Questions and Evaluation Design

This section explains how the paper turns possible next questions into ranked suggestions and then tests them against the later literature. At year $t-1$, I start from the literature map assembled from papers observed through that date. A candidate question is represented as a connection that has not yet appeared in that map. Formally, let $G_{t-1} = (V, E_{t-1})$ denote the claim graph assembled through that date. For a directed causal candidate, $u \to v$ is eligible when that ordered directed link has not yet appeared in the historical graph. For an undirected noncausal candidate, $\{u, v\}$ is eligible when the pair has not yet appeared as undirected support. The headline object in this paper is the directed causal candidate.

[2] The embedding step is used as a retrieval and ranking device after exact and signature-based passes, not as an unconstrained merge rule. That ordering keeps obvious cases deterministic and makes softer matches inspectable.

4.1 Missing links as candidate questions

One way to read the novelty and frontier literatures is that many scientific advances come from combinations or connections that were not yet explicit in the recorded structure of a field (Uzzi et al., 2013; Fortunato et al., 2018; Wang and Barabasi, 2021). The representation used here takes that intuition in a narrow form. The point is not that every paper can be reduced to one edge. The point is that many research moves can be approximated as the appearance of a relation that was already plausible in the nearby literature before it became explicit. That lets us define a prospectively testable object. The approach is closer to frontier tracing than to open-ended ideation: it asks which direct relations the existing local graph seems to invite next.

For the directed causal task, the candidate universe at cutoff t is

$$\mathscr{C}_t^D = \{(u, v) \in V \times V \,:\, u \neq v,\ (u \to v) \notin E_{t-1}^D\},$$

where E_{t-1}^D denotes the directed causal subgraph observed through year $t-1$. For the undirected noncausal task, the candidate universe is

$$\mathscr{C}_t^U = \{\{u, v\} \subset V \,:\, u \neq v,\ \{u, v\} \notin E_{t-1}^U\},$$

where E_{t-1}^U denotes the undirected contextual support subgraph. A future realization for the directed task occurs when $u \to v$ first appears during $[t, t+h]$. A future realization for the undirected task occurs when $\{u, v\}$ first appears during $[t, t+h]$.

4.2 Gap and boundary questions

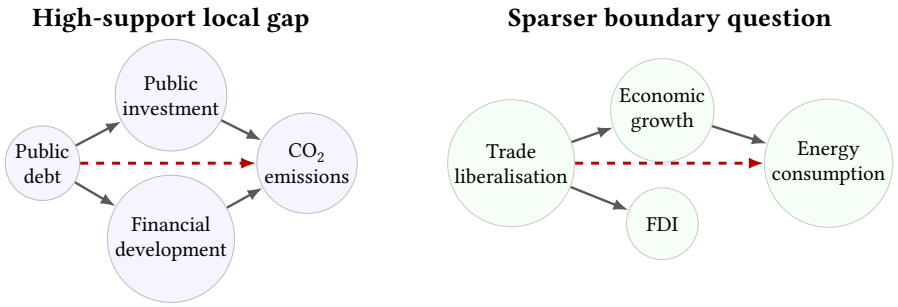
Two kinds of surfaced questions matter in practice. Gap questions already have rich nearby support but remain directly underworked. Boundary questions connect areas that still have little direct traffic between them. The paper's score is designed to separate these cases rather than collapse them into one novelty index. Gap questions are often easier to defend as plausible next questions; boundary questions are often more adventurous and potentially more fragile.

In the language of the graph, gap questions have high local support and low direct completion. Boundary questions have weaker local closure but connect parts of the graph that are still far apart. The distinction becomes substantively useful later, because the main model tends to surface a different gap-boundary mix than the popularity benchmark.

4.3 How the score reads the graph

The ranking rule combines four ingredients: path support, underexploration gap, motif support, and hub penalty. The easiest way to read them is in plain language. Path support asks whether the two endpoint

Figure 2: Gap questions and boundary questions are distinct graph patterns



Notes. The left panel shows a gap-like candidate: nearby support is already dense, but the direct relation remains missing. The right panel shows a more boundary-like candidate: the two end concepts are connected only by thinner bridges. The node labels are lightly cleaned versions of currently surfaced economics-facing examples. These diagrams are conceptual rather than exhaustive: they illustrate how the score distinguishes questions with rich local closure from questions that bridge sparser regions of the graph.

concepts are already connected by short routes through nearby mediators. The gap term asks whether those routes exist even though the direct relation itself is still absent or thin. Motif support asks whether the same endpoint pair keeps reappearing inside nearby structural patterns rather than in only one fragile corner of the graph. The hub penalty moves in the opposite direction: it reduces scores that are high only because both endpoints are extremely generic, heavily connected concepts.

In the implementation used for this paper, the score takes the form

$$s(u, v) = \alpha \tilde{P}(u, v) + \beta G(u, v) + \gamma \tilde{M}(u, v) - \delta \tilde{H}(u, v),$$

where $\tilde{P}(u, v)$ is normalized path support, $G(u, v)$ is the underexploration gap, $\tilde{M}(u, v)$ is normalized motif support, and $\tilde{H}(u, v)$ is the normalized hub penalty. The main specification uses $\alpha = 0.5$, $\beta = 0.2$, $\gamma = 0.3$, and $\delta = 0.2$, matching the scoring implementation used in the released outputs. Those weights are fixed as a transparent design choice rather than tuned to maximize forecasting performance. Put differently, a pair scores highly when several nearby routes keep pointing toward it, the direct relation still looks oddly absent relative to that support, and the pair is not simply another generic high-degree combination. Figures 2 and 4 are useful here because they show what the rule is actually rewarding: multiple short paths, repeated local support, and a direct link that still looks underfilled relative to its neighborhood.

That transparency choice matters for interpretation. The graph already stores stability, causal-presentation, evidence-type, and edge-role metadata, but the main score does not yet fully weight those signals. That is deliberate in this version. The released score is meant to be inspectable question by question rather than optimized as a forecasting black box, and richer credibility weighting is better understood as the next extension than as a hidden tuning layer.

Figure 3: Observed local paths can nominate a missing direct link, but later research can also move in the reverse direction

Notes. The left panel fixes the benchmark object used throughout the paper: a local path such as $u \to w \to v$ can nominate a missing direct link $u \to v$ as a candidate next paper. The right panel shows a different research move: later work can add a mediator path around an existing direct relation rather than closing a missing direct link. The main backtest focuses on the left-hand object. Section 5.5 returns to the right-hand pattern.

4.4 Prospective evaluation

The prospective design freezes the graph at year $t-1$, ranks candidates using only information available at that date, and then checks whether those links first appear over the evaluation horizon. This vintage design avoids using future edges, future degrees, and later realizations when scoring the historical cutoff year,[3] which keeps the interpretation close to the real ex ante decision problem. The benchmark is therefore not "how well does the score fit the historical graph?" but "how well would this score have surfaced links that later became realized work?" That distinction matters because retrospective fit can reward mechanical regularities that are not actually useful for research allocation.

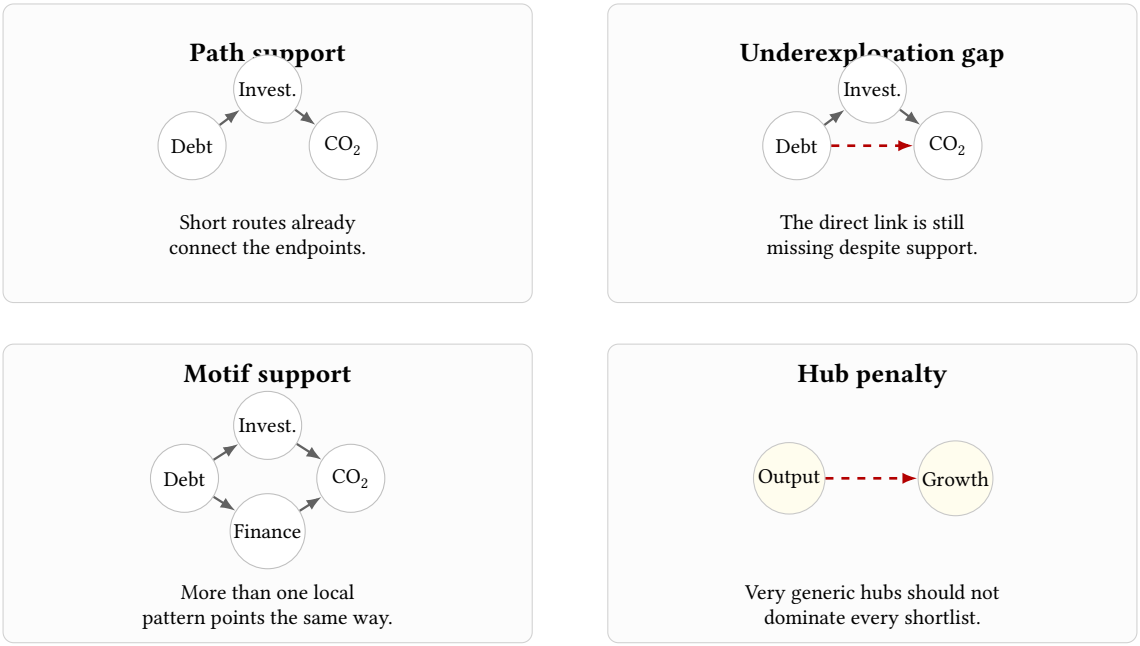
For each cutoff year, the graph is built from the historical stock only. Realizations are then defined from future papers over the chosen horizon. The benchmark is rolling rather than one-shot, so each horizon is evaluated across multiple cutoff dates. A cutoff is eligible for horizon h only if $t + h \leq 2026$, and the heterogeneity atlas applies the same rule on a five-year cutoff grid when it extends the exercise to $h = 20$. The headline benchmark focuses on directed causal candidates, but the fuller atlas also evaluates directed and undirected objects separately and then pools them by weighted aggregation rather than by constructing one mixed ranking universe.

Preferential attachment as benchmark. Preferential attachment scores a candidate ordered pair by source out-degree times target in-degree:

$$\mathrm{PA}(u, v) = d^{out}(u) \times d^{in}(v).$$

[3] In other words, the model at year $t-1$ is not allowed to borrow information from papers that appear in t or later. That is the sense in which the design avoids leakage here.

Figure 4: How the score reads a local neighborhood



Notes. This figure breaks the score into four local graph features. Path support asks whether short routes already connect the endpoints. The underexploration gap asks whether the direct relation is still missing despite that support. Motif support asks whether more than one nearby pattern points toward the same endpoint pair. The hub penalty discounts pairs that would rank highly only because both concepts are very generic and heavily connected. The diagrams are stylized, but the labels are adapted from live economics-facing examples.

In plain language, it is a rich-get-richer rule: already central concepts attract more future links. This is the right benchmark because the literature is not generated by neutral exploration alone. Topics with existing visibility, existing datasets, recognizable methods, and established readership often attract still more work. If a graph-based score cannot outperform that baseline in the benchmark, that is a substantive result rather than a disappointment.

Preferential attachment is also the main benchmark because cumulative advantage is the main economic null in this setting. Standard graph baselines such as common-neighbors, Katz-style scores, or embedding methods are useful future appendix comparisons, but they are not the primary benchmark here because the paper is not asking a generic network-completion question.

Fixed-budget retrieval. The evaluation problem is about how many suggestions a researcher is willing to inspect. That is why the paper emphasizes Recall@100, other fixed-budget shortlist measures, and frontier-style comparisons over larger K.

Horizon choice. The main horizons are 3, 5, 10, and 15 years because they correspond to distinct practical windows. Three years is a short-run scouting horizon. Five years is a natural publication and diffusion window in economics. Ten years captures slower movement in topics that take longer to propa-

gate through papers, methods, and field conventions. Fifteen years is still empirically useful in the fuller published-journal sample and is promoted into the main paper because many slower literatures are only partly visible at shorter horizons. Twenty years remains an appendix extension.

5 What the Benchmark Shows

I first ask whether the method beats a simple popularity-based rule at the very top of the ranking. It usually does not. I then ask whether the graph-based score becomes more useful once a researcher is willing to inspect a broader shortlist, weight later links by downstream reuse, and look at where in the literature local structure does more of the screening work.

How to read the benchmark

The most intuitive body metric is *future links per 100 suggested questions*: among 100 surfaced questions, how many later appear as realized directed links in the literature. Recall@100 asks what share of all future realized links are captured in a 100-question shortlist. Mean reciprocal rank rewards putting those realized links nearer the top of the same shortlist.

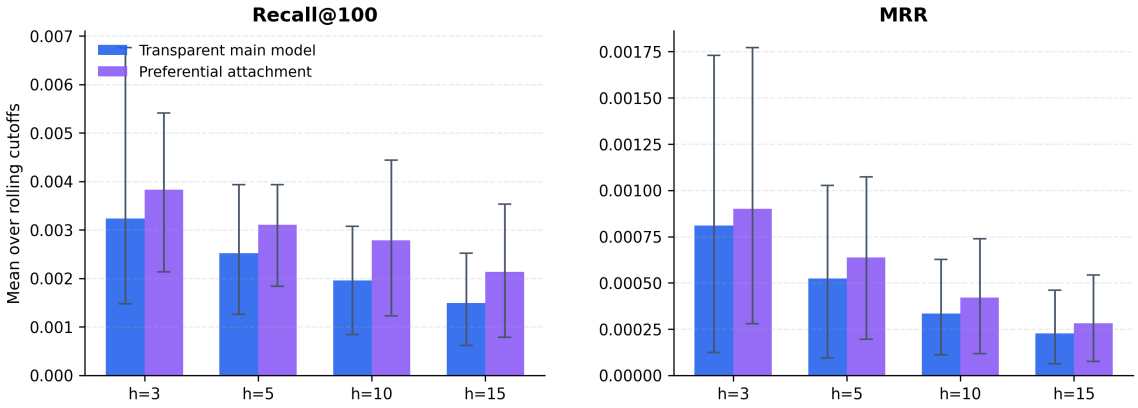
5.1 Popularity at the strict shortlist

At the strict top of the ranking, the simple popularity-based rule still wins. In network terms that rule is preferential attachment. The easiest way to read the magnitude is in future links captured inside a 100-question shortlist: preferential attachment places about 8.3, 12.0, 23.3, and 36.3 future directed links inside the top 100 at $h = 3, 5, 10, 15$, while the graph-based score places about 5.7, 8.7, 16.3, and 26.3. So the popularity benchmark buys roughly 2.6, 3.3, 7.0, and 10.0 extra realized directed links inside the same 100-candidate shortlist. Put differently, preferential attachment retrieves roughly 40 percent more realized directed links than the graph score, depending on the horizon. The normalized Recall@100 and MRR statistics tell the same story.

The small normalized values are real, but they are not trivial. This is a severe screening task over a very large candidate universe. On average, the future contains about 2,955 realized directed links at $h = 3$, 4,994 at $h = 5$, 13,221 at $h = 10$, and 29,809 at $h = 15$. So the top-100 shortlist is not being asked to recover a handful of outcomes. It is being asked to pull forward a small fraction of a very large future stock. That is why the benchmark should be read as a scarce-reading-time problem, not as a classifier accuracy problem.

The right conclusion from the strict headline is not "the graph score fails." It is narrower and more interesting. If the decision problem is to identify the single most likely next direct link under a very tight reading budget, cumulative advantage remains very hard to beat. The literature keeps returning to already central concepts. The next question is whether that headline survives once the screening frontier widens and we ask where the literature is more or less popularity-dominated.

Figure 5: Preferential attachment remains stronger at the strict shortlist margin



Notes. The left panel asks what share of later-realized links are captured inside a 100-question shortlist (Recall@100). The right panel asks how highly those later-realized links are placed within the same shortlist (MRR). Each bar is the mean across eligible rolling cutoffs for a given horizon, with bootstrap confidence intervals. For readers who prefer a more concrete scale, the corresponding mean hits inside the top-100 shortlist are about 5.7, 8.7, 16.3, and 26.3 for the graph score and 8.3, 12.0, 23.3, and 36.3 for preferential attachment across $h = 3, 5, 10, 15$.

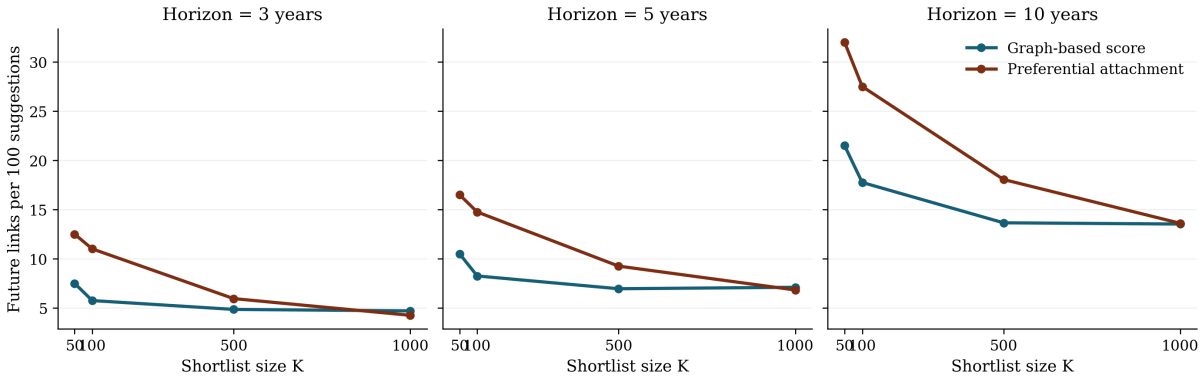
5.2 The attention-allocation frontier

The first way to move from prediction toward allocation is to relax the top-100 bottleneck. Economists rarely consume one candidate suggestion and stop; they inspect a shortlist. The attention-allocation outputs therefore ask what happens as that shortlist expands from $K = 50$ to $K = 1000$. I summarize that margin using "future links per 100 suggested questions," which is just the shortlist precision rescaled into a more readable unit.

The result is again mixed but informative. At $h = 3$, preferential attachment places about 11.0 future links per 100 suggestions at $K = 100$, compared with 5.75 for the graph score. By $K = 1000$, the two rules are essentially tied in practical terms: preferential attachment yields about 4.25 future links per 100 while the graph score yields about 4.70. The same pattern appears at $h = 5$: the gap is 14.75 versus 8.25 at $K = 100$, but 6.83 versus 7.10 by $K = 1000$. At $h = 10$, the tight-budget gap remains larger, yet even there the frontier narrows substantially, from 27.5 versus 17.75 at $K = 100$ to 13.6 versus 13.5 by $K = 1000$. The point is not that the graph score suddenly becomes the dominant forecaster. It is that a winner-take-all top-rank view overstates how far popularity dominates the broader reading lists that real researchers actually inspect.

That makes the current paper's answer to the title more precise. If "what should economics ask next?" is interpreted as "what is the single most likely next direct link?", preferential attachment wins. If it is interpreted as "which questions should a researcher read, scope, or test next under a realistic shortlist budget?", the graph-based object becomes more relevant. It remains weaker at the very top rank, but it moves materially closer once the screening problem looks more like actual research browsing.

Figure 6: The attention-allocation frontier softens the strict-shortlist headline



Notes. Each panel reports mean future links per 100 surfaced suggestions as the shortlist expands from $K = 50$ to $K = 1000$. In plain terms, the figure asks what happens as a researcher becomes willing to inspect more suggestions. Preferential attachment remains stronger at very small K, but the gap shrinks sharply as that shortlist widens.

5.3 What changes when future links are value-weighted

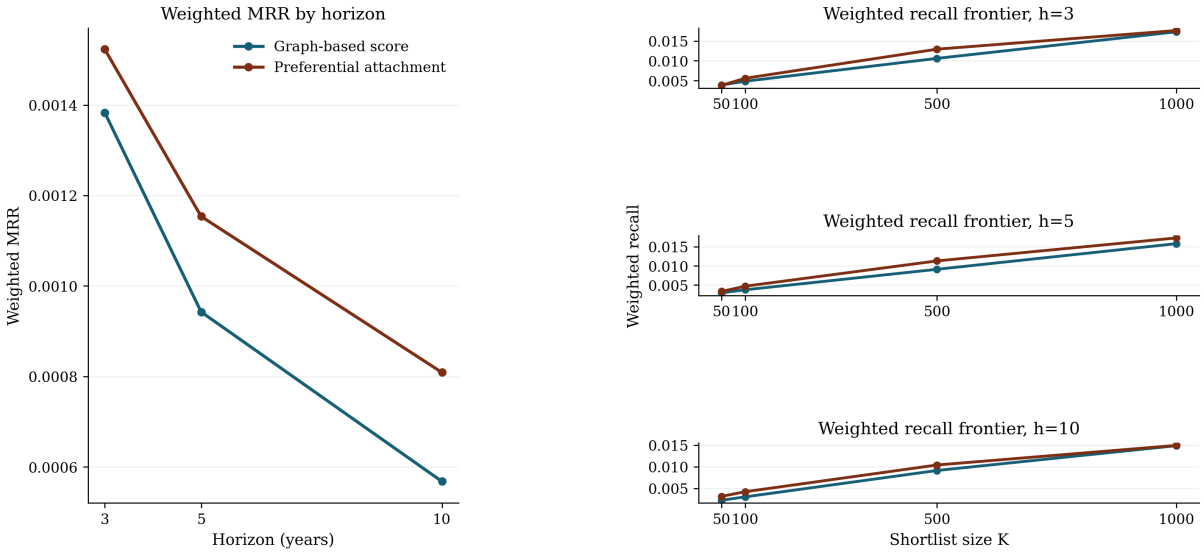
Future appearance is not the only margin that matters. A later realized link can also be weighted by downstream reuse, so that some realized links count more than others. The impact-weighted rerun therefore asks whether the graph score looks relatively better once the future is weighted by later reuse rather than treated as binary appearance alone.

The answer is again disciplined rather than triumphant. Weighted MRR still favors preferential attachment at each of the main horizons: about 0.001523 versus 0.001383 at $h = 3$, 0.001154 versus 0.000943 at $h = 5$, and 0.000809 versus 0.000568 at $h = 10$. So the strict-headline result is not only about low-value fills. Central concepts still capture more of the heavily reused future links. But the broader weighted frontier is less one-sided than the weighted MRR line alone suggests. At $K = 1000$, weighted recall is nearly tied at $h = 3$ and $h = 10$: preferential attachment reaches about 0.01762 and 0.01495, while the graph score reaches about 0.01729 and 0.01488. The gap is still larger at $h = 5$, but even there it is far smaller than the tight-rank headline would suggest.

That result matters for how the title should be read. It shows that the paper is not merely reclassifying trivial future links as success. Weighting by downstream reuse leaves the top-rank popularity story intact. The more favorable reading for the graph score enters instead through broader attention frontiers and through the kinds of literatures in which local structure does more screening work.

This is also where the paper's credibility story enters. The graph is not built from raw co-occurrence. It already carries stability, causal-presentation, evidence-type, and edge-role metadata from the paper-local extraction layer. Appendix E shows that directed causal rows have mean stability around 0.93, compared with about 0.87 for undirected contextual rows, and that over 90 percent of directed causal rows fall into the high-stability band. So the method-family heterogeneity results below are not just subfield color. They are part of the paper's broader claim that some local graph neighborhoods are more credible terrain for

Figure 7: Value-weighting changes the scale of the benchmark but does not overturn the headline



Notes. The left panel reports weighted MRR by horizon, where future realized links are weighted by later reuse. The right panels report weighted recall frontiers over shortlist size K. Preferential attachment still dominates the tighter top ranks, but the weighted frontier narrows materially at broad lists.

screening than others, even though the current main score does not yet fully weight those signals.

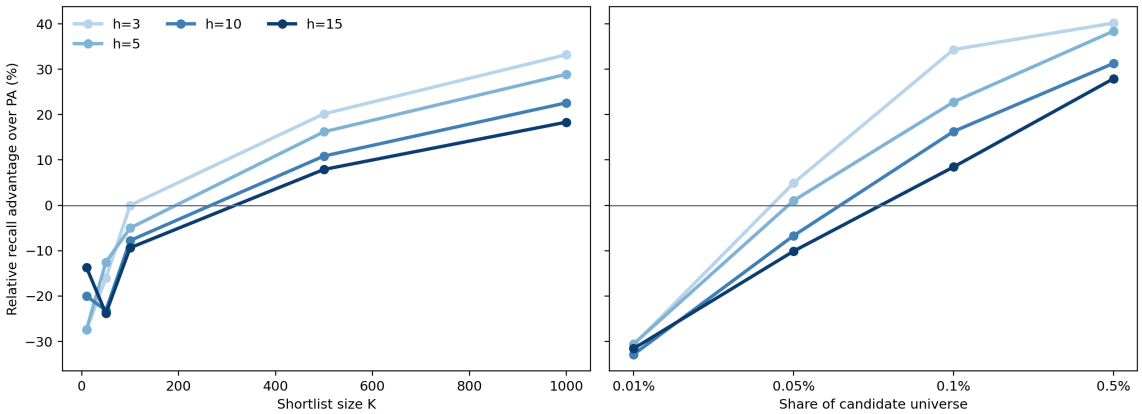
5.4 Where structure helps more

The pooled top-100 comparison hides meaningful variation. The most useful way to read the atlas is not as a search for one subgroup in which the graph score cleanly "wins." It is a map of where cumulative advantage is more dominant and where the nearby structure of the literature adds more screening value. Once the frontier is evaluated over broader fixed-K and percentile-K shortlists, the graph score becomes substantially more competitive than the strict top-100 headline suggests. In the pooled frontier view, its percentile-K advantage is slightly positive at each of $h = 3, 5, 10, 15$, even though the top-100 delta is near zero or negative. That already changes the interpretation of the exercise: the model looks weaker as a winner-take-all forecaster than it does as a broader screening rule.

The subgroup results sharpen this interpretation. Journal tier matters. Adjacent journals are more favorable terrain for the graph score than the core: the pooled percentile-frontier advantage is about 0.000320, 0.000275, 0.000201, and 0.000155 at $h = 3, 5, 10, 15$ in adjacent journals, compared with 0.000134, 0.000109, 0.000058, and 0.000038 in core journals. Method family matters as well. Design-based causal slices are much more favorable than panel- or time-series-heavy slices: the pooled percentile-frontier delta for the design-based group is about 0.000717, 0.001029, 0.000354, and 0.000063 at $h = 3, 5, 10, 15$, while the panel/time-series group is negative at every one of those horizons.

Funding adds nuance rather than a single clean pattern. In the coarse funded-versus-unfunded split, the

Figure 8: The pooled frontier view is more favorable to the graph score than the strict top-100 headline



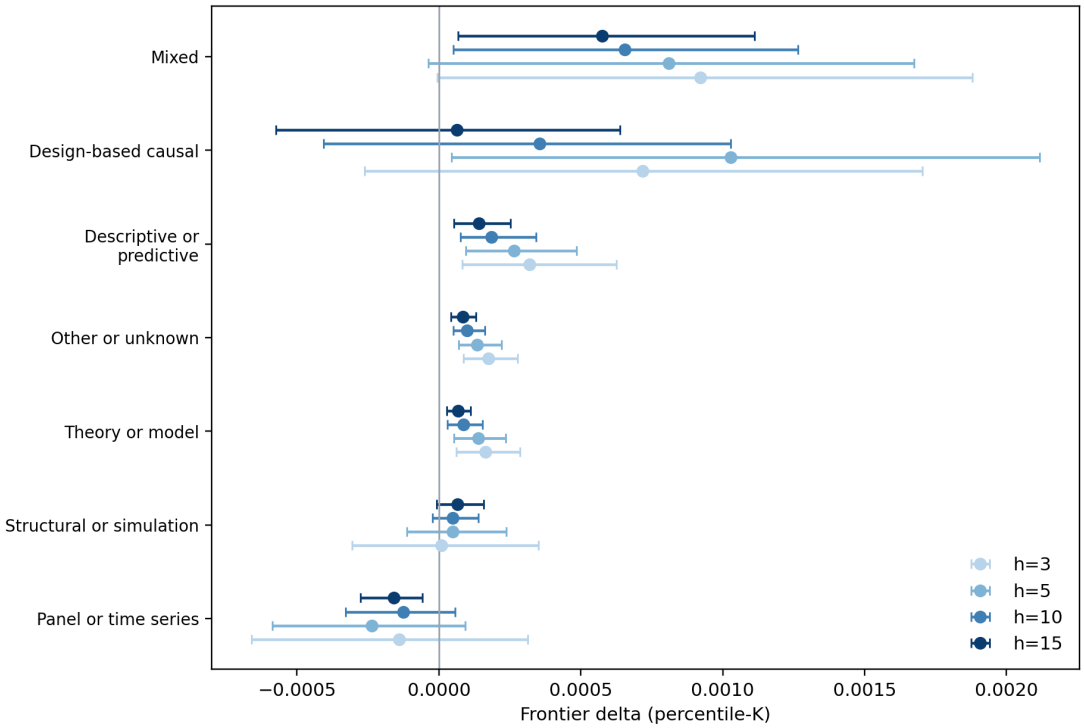
Notes. The pooled frontier figure reports the graph score's relative recall advantage over preferential attachment. Positive values favor the graph score. The lighter horizon lines correspond to shorter horizons and the darker lines to longer horizons. The key interpretation is that the graph score becomes more competitive once the shortlist expands beyond the strict top-100 comparison.

funded literature is actually less favorable to the graph score: the pooled percentile-frontier delta is negative at all main horizons for funded work and positive at all main horizons for unfunded work. That does not mean funding suppresses good ideas. It means that, in this benchmark, funded realizations look more popularity-dominated on average. The appendix therefore treats funding as suggestive rather than central, and Appendix Figure 17 is useful mainly because it shows that the funded pattern is not uniform. Among the stable high-support funders, the Economic and Social Research Council is the clearest positive outlier, while the large China and Germany groups are closer to zero and the U.S. National Science Foundation is roughly neutral.

The topic split makes the same point in more familiar language. The graph score looks strongest in healthcare systems and quality-of-life topics, in labor-market-and-inequality questions, and in several banking, housing, and macro-policy clusters. It looks weaker in a smaller set of environmental-policy and discrete-choice clusters. That should not be over-read as a ranking of subfields. It is a map of where the pooled average hides concrete heterogeneity.

The robust main-text message is therefore restrained but substantive. Broader frontier shortlists soften the pooled headline. Adjacent journals look better than the core. Design-based slices look better than panel or time series. Several concrete economics topics look better than the pooled average. Funding seems to matter, but mostly as a secondary institutional layer on top of the more basic popularity-versus-structure comparison. If the title is read as a question about where a structural screen is most useful, this subsection gives the clearest answer: not everywhere equally, but especially in adjacent, design-based, and several concrete economics-facing topic clusters.

Figure 9: Method-family heterogeneity is economically interpretable



Notes. This figure compares broad method families. The practical reading is simple: design-based causal work is materially more favorable terrain for the graph score than panel- or time-series-heavy work.

5.5 Path development beyond direct-link closure

5.5.1 Aggregate transition patterns

The direct-link framing is not the only way research can evolve. A literature can also move in the reverse direction: starting from a direct claim, later work can add mediating paths around it rather than closing a previously missing direct relation. I therefore distinguish two simple transition types on length-2 structure. "Path to direct" means that a supporting $u \to w \to v$ path already exists at $t-1$, the direct $u \to v$ link does not, and that direct link then appears by $t+h$. "Direct to path" means the direct link exists first, but a supporting mediator path appears only later.

The aggregate comparison yields a striking result. Direct-to-path transitions dominate path-to-direct transitions in every cutoff-period block and at every horizon currently studied. At $h = 10$, for example, the direct-to-path share rises from roughly 0.049 in the 1980s to 0.089 in the 1990s, 0.178 in the 2000s, and 0.355 in the 2010s. The corresponding path-to-direct shares are much smaller: about 0.023, 0.014, 0.015, and 0.020. So the literature often elaborates mechanisms around claims it already treats as direct rather than closing a missing direct link implied by nearby paths.

Table 2: Selected path-rich candidate questions

Candidate question	Supporting paths	Example mediators
Investment → carbon emissions	38	economic growth; technological innovation; economic development
Public debt → CO_2 emissions	23	economic growth; financial development; renewable energy consumption
Monetary policy → energy consumption	23	income; output; income inequality
Trade liberalisation → energy consumption	5	economic growth; foreign direct investment; trade liberalization
Urbanization → output growth	17	CO_2 emissions; energy consumption; energy use

5.5.2 Where path closure is more common

The journal split is especially revealing. At $h = 3, 5, 10, 15$, the share of realized path-related transitions that take the path-to-direct form is about 0.571, 0.579, 0.529, and 0.471 in adjacent journals, but only about 0.442, 0.443, 0.400, and 0.360 in the core. So adjacent journals are much more path-closure heavy. The core is more likely to elaborate around existing direct links.

The broad subfield split points in the same direction. Economics and Econometrics is relatively balanced at short horizons, with path-to-direct shares of about 0.528 and 0.532 at $h = 3$ and $h = 5$, before turning more direct-to-path heavy by $h = 10$ and longer horizons. Finance is more direct-to-path heavy throughout: the path-to-direct share is only about 0.418, 0.425, 0.398, and 0.366 at $h = 3, 5, 10, 15$. So the path result is not just a pooled artifact. It varies in economically interpretable ways across the literature.

5.5.3 Current path-rich examples

The recommendation layer already hints at what path-rich questions look like. By path-rich, I mean questions supported by many nearby chains of connections rather than by one isolated bridge. Investment → carbon emissions is supported by 38 observed paths through concepts such as economic growth, technological innovation, and economic development. Public debt → CO_2 emissions has 23 supporting paths through growth, financial development, and renewable energy consumption. Monetary policy → energy consumption has 23 supporting paths through income, output, and income inequality. These examples are not historical validation evidence, but they do show why the path-based object is concrete enough to inspect in the public interface rather than treat as an abstract graph statistic.

Taken together, Sections 5.1 to 5.5 imply a cumulative reading of the evidence. The strict top-100 benchmark is harsh and popularity-dominated. Broader attention frontiers soften that headline. Value-weighting changes the scale of the comparison without reversing it. Heterogeneity shows where structural screening is actually more useful. The path audit then explains why even that richer reading still does not exhaust

the graph's value: a good share of scientific development takes the form of mechanism-deepening around existing direct claims, not only direct-link closure itself. In that sense, the most useful questions to ask next are often better understood as path-rich research programs than as single isolated missing edges.

6 Discussion and Conclusion

Several limits matter for interpretation. A future realized link is not the same thing as truth, importance, or policy value. The benchmark is about future appearance in the literature, not about a complete normative theory of which questions economists should pursue. If cumulative advantage dominates the future, preferential attachment can outperform even when the graph score is surfacing more genuinely underexplored questions. That is why I treat the prospective benchmark as informative but not exhaustive.

Direction in the graph records ordered claim relations rather than final causal adjudication. The main score still treats the existence of an edge more seriously than the strength or credibility of the underlying evidence, even though method and stability metadata now exist in the pipeline. That is one reason the next methodological step should probably incorporate stronger credibility weighting and perhaps separate tasks for directed causal emergence, undirected contextual emergence, and path emergence.

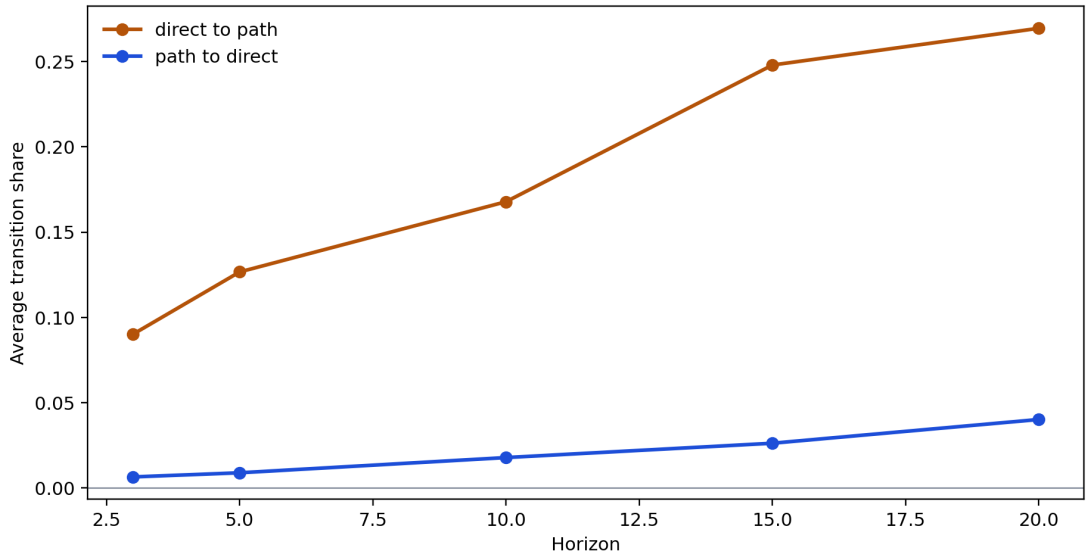
The published-journal corpus is a further deliberate restriction rather than a universal map of all economics research. It gives the project a cleaner realized literature, but it underweights working-paper traffic, seminar diffusion, and some forms of genuinely new frontier language.

These limits do not make the exercise empty. They define its scope. The paper's answer to its own title is narrower than a welfare theorem but still substantive: economics should not decide what to ask next only through cumulative advantage. The most useful surfaced questions are neglected enough to remain open, supported enough to be credible, concrete enough to become papers, and best read at realistic attention frontiers rather than at winner-take-all top ranks. Empirically, that means the strict top-100 shortlist still favors preferential attachment, but broader attention frontiers, value-weighted outcomes, heterogeneity, and path development all make more room for structural screening than the pooled headline alone suggests. A next iteration should add stronger credibility weighting and richer path-based objects, and could also compare explanation or reranking layers across LLMs as a bounded appendix-style extension without changing the current paper's observational core.

Figure 10: Economics-facing topic heterogeneity

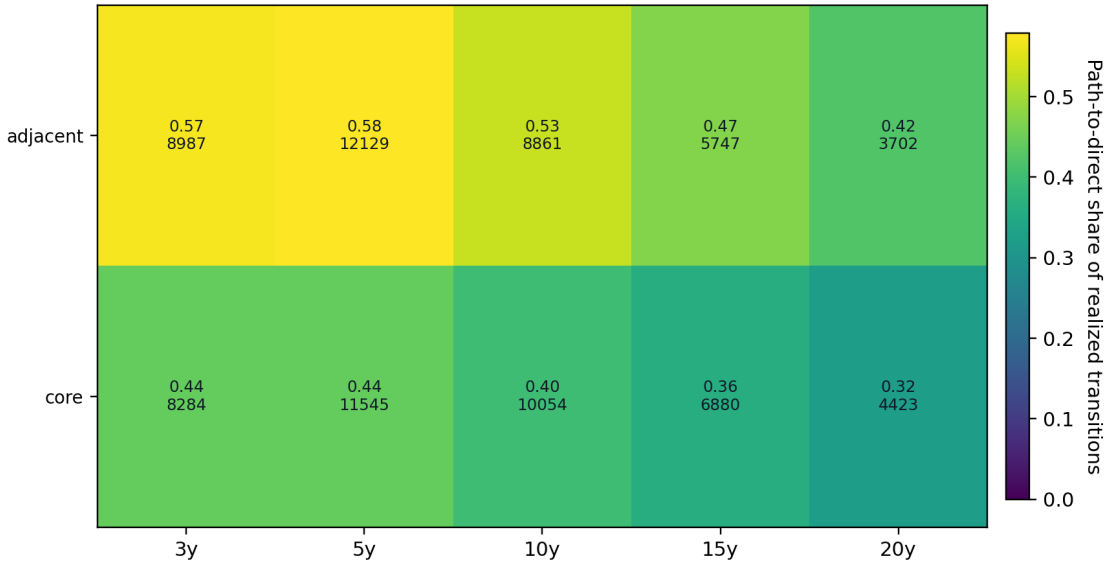
Notes. The main topic heatmap prioritizes the most populous economics-facing topic groups rather than all broad adjacent categories. Cell color reports the pooled percentile-frontier advantage of the graph score over preferential attachment, while the annotations report the top-100 hit delta in basis points.

Figure 11: Research often adds mediating path structure around existing direct links more often than it closes path-implied direct links



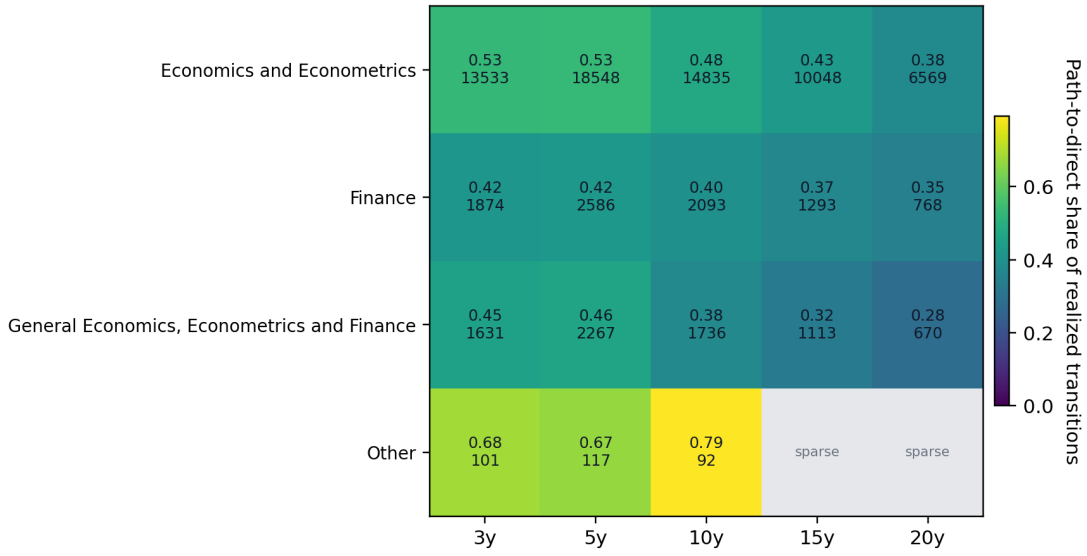
Notes. This figure compares two transition types on length-2 graph structure. "Path to direct" means a local path exists first and the missing direct edge later appears. "Direct to path" means a direct edge exists first and a supporting mediator path appears only later. The unit of analysis is the eligible concept pair at each cutoff-period and horizon cell. Shares are computed relative to eligible pair stocks in the corresponding transition class. The figure should be read as a graph-evolution result rather than a forecast metric: it shows that mechanism-deepening around existing direct claims is often more common than direct-link closure.

Figure 12: Path closure is relatively more common in adjacent journals than in the core



Notes. The figure reports the mix of realized path-related transitions by journal tier. The reported quantity is the share of realized transitions that take the path-to-direct form rather than the direct-to-path form. A higher value therefore means more direct-link closure relative to mechanism-deepening around existing direct edges. The unit of analysis is the realized transition within each journal-tier-by-horizon cell. Adjacent journals are consistently more path-to-direct heavy than core journals, although direct-to-path remains important in both.

Figure 13: Path transition mix by broad subfield



Notes. This figure reports the share of realized path-related transitions that take the path-to-direct form by broad subfield and horizon. Economics and Econometrics is more balanced than Finance at short horizons, but both become more direct-to-path heavy at longer horizons. The result supports the main interpretation that graph evolution often proceeds by mechanism-deepening around known direct claims.

A Paper-local graph extraction

This appendix documents the paper-local extraction layer used in the paper's evaluation. Garg and Fetzer (2025) show that economics papers can be converted into paper-level claim graphs by prompting a language model to recover nodes, directional relations, and claim metadata from title-and-abstract text. The present paper inherits that paper-local view of extraction but extends it for a different downstream object. Here the graph must support missing-link construction, gap-versus-boundary distinctions, causal versus noncausal splitting, and later node normalization into a reusable concept ontology. That requires a somewhat richer paper-local schema than a simple causal-claim inventory.

Three extensions matter most. First, the extraction schema includes undirected relations, because contextual noncausal support is substantively useful even when the headline task is directed causal emergence. Second, the schema separates the paper's *causal presentation* from the *evidence method* used to support a relation. A paper can speak causally while using a weak empirical design, or it can use a strong design while describing the finding more cautiously. Third, the local schema stores contextual qualifiers separately from the node label wherever possible, so that later normalization can target concept identity rather than local sample wording.

A.1 Prompts

The benchmark uses a fixed system prompt and a minimal user prompt template. They are reproduced exactly below.

System prompt.

```
You extract a paper-local research graph from a paper title and abstract.

Return only structured output that matches the supplied JSON schema.

Task:
- Read the paper title and abstract.
- Build a paper-local graph with `nodes` and `edges`.
- Reuse the same node when the same concept genuinely recurs within the same paper.
- Do not use outside knowledge.
- Do not infer relationships that are not supported by the title or abstract.

Purpose:
- The output will later be turned into a larger deterministic research graph.
- Downstream systems depend on consistent paper-local node reuse.
- If the abstract contains a chain like `A -> B`, `B -> C`, and `X -> B`, the shared concept `B` should be
↪  represented by the same paper-local node if it is genuinely the same concept.
- However, do not merge distinct concepts just because they seem related.

Critical rules:
- Do not create transitive closure.
- If the abstract states `A -> B` and `B -> C`, do not create `A -> C` unless the paper explicitly states `A -> C`.
```

```
- Do not create both `A -> B` and `B -> A` for one undirected claim.
- If the title or abstract says two variables are associated or correlated without directional language, encode
↪  one edge with `directionality = undirected`.
- For undirected edges, use the first-mentioned concept as `source_node_id` and the second-mentioned concept as
↪  `target_node_id` only as a storage convention.

How to represent nodes:
- Use concise noun phrases grounded in the paper text.
- Keep node labels concept-level when possible.
- Do not bake country or year into the node label unless it is essential to the concept itself.
- Put local scope information into `study_context` or `condition_or_scope_text`.
- Use `surface_forms` for distinct mentions that refer to the same paper-local concept.
- Use `study_context` only for context explicitly stated in the title or abstract.
- If no context is stated, use:
  - `unit_of_analysis: []`
  - `start_year: []`
  - `end_year: []`
  - `countries: []`
  - `context_note: "NA"`

How to represent edges:
- Extract only relations that the title or abstract states, studies, or reports.
- Keep background or prior-literature claims only if they are explicitly stated in the title or abstract, and mark
↪  them with `edge_role = background`.
- Use `claim_text` as a short normalized relation string.
- Use `evidence_text` as a short supporting excerpt or close paraphrase from the title/abstract only.

Directionality:
- Use `directionality = directed` when the paper frames one concept as affecting, predicting, changing,
↪  increasing, decreasing, explaining, or determining another.
- Use `directionality = undirected` when the paper frames the relation as association, correlation, co-movement,
↪  similarity, or linkage without directional commitment.
- Prediction is directional, even if it is not causal.

Causal presentation:
- `explicit_causal`: the paper explicitly uses causal language such as affects, causes, leads to, increases,
↪  reduces, impact of, effect of.
- `implicit_causal`: the paper strongly frames the relation as an effect or treatment relation without fully
↪  explicit causal wording.
- `noncausal`: the paper frames the relation as association, correlation, prediction, linkage, or descriptive
↪  relation.
- `unclear`: the wording is too ambiguous to classify confidently.
- This field is about how the paper presents the relation, not whether the method truly justifies causality.

Relationship type:
- `effect`: one concept is presented as affecting another.
- `association`: correlation, co-movement, linkage, or association.
- `prediction`: one concept predicts or forecasts another.
- `difference`: one concept differs across groups, places, times, or conditions.
- `other`: only if none of the above fit.

Edge role:
- `main_effect`: central edge or main result in the abstract.
- `mechanism`: pathway or channel relation.
```

```
- `heterogeneity`: subgroup or conditional variation in a relation.
- `descriptive_pattern`: stylized fact or descriptive empirical pattern.
- `background`: motivating or prior-literature relation stated in the abstract.
- `robustness`: supporting or validating relation rather than the main contribution.
- `other`: only if needed.

Claim status:
- `effect_present`: the abstract reports that the relation is present.
- `no_effect`: the abstract reports no effect or no relation.
- `mixed_or_ambiguous`: the abstract reports mixed, inconsistent, or ambiguous results.
- `conditional_effect`: the relation holds only for some subgroup, time period, or condition.
- `question_only`: the abstract raises or studies the relation but does not report a result.
- `other`: only if needed.

Explicitness:
- `result_only`: the relation is presented as a result.
- `question_only`: the relation is posed as a question or objective only.
- `question_and_result`: the abstract both frames the question and reports a result on the same relation.
- `background_claim`: the relation appears as background motivation or prior literature.
- `implied`: the relation is clearly implied by the abstract wording but not directly phrased as a standalone
↪  claim.

Condition or scope:
- Use `condition_or_scope_text` for subgroup, timing, geographic, or sample qualifiers on the edge.
- Examples: `among older workers`, `during recessions`, `in rural counties`, `for low-income households`.
- Use `NA` if not needed.

Sign:
- `increase`, `decrease`, `no_effect`, `ambiguous`, `NA`
- Use `NA` if sign is not applicable or not stated.

Statistical significance:
- `significant`: the abstract clearly says the result is statistically significant.
- `not_significant`: the abstract clearly says it is not statistically significant.
- `mixed_or_ambiguous`: significance differs across findings or is ambiguously stated.
- `not_reported`: no significance statement is provided.
- `NA`: only if not applicable.

Evidence method:
- Choose the best supported option from the schema.
- `experiment`: field, lab, survey, or randomized experiment.
- `DiD`: difference-in-differences or closely related staggered-treatment treatment-control design.
- `IV`: instrumental variables or closely related design based on an instrument.
- `RDD`: regression discontinuity or closely related cutoff-based design.
- `event_study`: dynamic pre/post treatment-event design.
- `panel_FE_or_TWFE`: panel fixed-effects or two-way fixed-effects empirical design without a clearer method
↪  family being the main identification label.
- `time_series_econometrics`: VAR, VECM, ARDL, cointegration, error-correction, Granger-causality, GARCH, or
↪  similar time-series econometric design.
- `structural_model`: estimated structural economic model.
- `simulation`: simulation or computational experiment.
- `theory_or_model`: formal theory, conceptual model, or analytical model without direct empirical estimation.
- `qualitative_or_case_study`: interview, ethnographic, archival qualitative work, or case study.
- `descriptive_observational`: nonexperimental empirical analysis without a clearer identified design.
```

```
- `prediction_or_forecasting`: predictive or forecasting model where the emphasis is forecast performance rather
↪  than causal identification.
- Use `do_not_know` if the abstract does not reveal enough.
- Use `other` only if a method is clearly stated but does not fit the listed categories.

Nature of evidence:
- Choose the broad evidence type used for that edge.

Uses data:
- `true` if the edge is supported by data use described in the title/abstract.
- `false` for theory-only, conceptual, simulation-only, commentary, or clearly non-data papers.

Sources of exogenous variation:
- Record only if explicitly stated in the title or abstract.
- Otherwise use `NA`.

Tentativeness:
- `certain`: strong assertive language.
- `tentative`: cautious or suggestive language.
- `mixed_or_qualified`: strong claim with explicit qualification or limits.
- `unclear`: cannot tell.

What not to do:
- Do not label edges as collider, confounder, mediator, instrument, or any other downstream graph-structural role.
- Do not globally canonicalize concepts across papers.
- Do not create edges from general world knowledge.
- Do not invent countries, years, samples, or methods.

If the title/abstract contains no extractable graph:
- return `nodes: []` and `edges: []`
```

User prompt template.

```
Extract a paper-local research graph from the following title and abstract.

Use only the information in the title and abstract.
Return only the structured output that matches the supplied JSON schema.

Title:
{{paper_title}}

Abstract:
{{paper_abstract}}
```

A.2 Schema

The model returns a paper-local graph with `nodes` and `edges`. The underlying output is JSON. Tables 3 and 4 present the same fields in reader-facing labels for ease of inspection.

Table 3: Paper-local node schema

Field	Allowed values / type	Meaning	Why it exists downstream
Node identifier (`node_id`)	string	Paper-local identifier such as `n1`.	Needed so edges can reuse the same local concept deterministically within a paper.
Concept label (`label`)	short string	Concise concept label grounded in the title/abstract.	Becomes the base string passed into normalization and ontology mapping.
Surface forms (`surface_forms`)	array of strings	Distinct surface mentions in the title/abstract that refer to the same local concept.	Preserves within-paper synonymy without forcing global canonicalization at extraction time.
Unit of analysis (`study_context.unit_of_analysis`)	array of enumerated strings	Explicit unit of analysis linked to the node, if stated.	Keeps sample and scope off the concept label while preserving paper-local context.
Start year (`study_context.start_year`)	array of integers	Explicit start years if stated.	Preserves local scope without creating separate concept nodes for years.
End year (`study_context.end_year`)	array of integers	Explicit end years if stated.	Same reason as start year.
Countries (`study_context.countries`)	array of strings	Explicit countries if stated.	Preserves local setting without baking geography into concept identity unless essential.
Context note (`study_context.context_note`)	string	Residual local scope text such as "older workers" or "rural counties".	Retains paper-local nuance for audit and display while leaving the node label concept-level.

Table 4: Paper-local edge schema

Field	Allowed values / type	Meaning	Why it exists downstream
Edge identifier (`edge_id`)	string	Paper-local identifier such as `e1`.	Stable local relation key.
Source and target nodes (`source_node_id`, `target_node_id`)	strings	Local source and target node references.	Connect extracted relations back to the paper-local node inventory.
Directionality (`directionality`)	directed / undirected	Whether the text presents the relation directionally.	Determines whether the downstream graph stores an ordered link or an undirected contextual pair.
Relationship type (`relationship_type`)	effect / association / prediction / difference / other	Coarse semantic type of the relation.	Supports later filtering and audit of what kind of relation is being surfaced.
Causal presentation (`causal_presentation`)	explicit_causal / implicit_causal / noncausal / unclear	How the paper *describes* the relation.	Separates language from design quality; useful for credibility audits and task splitting.
Argument role (`edge_role`)	main_effect / mechanism / heterogeneity / descriptive_pattern / background / robustness / other	Role of the relation inside the paper's argument.	Distinguishes central claims from channels, subgroup effects, and background statements.
Claim status (`claim_status`)	effect_present / no_effect / mixed_or_ambiguous / conditional_effect / question_only / other	What result the paper reports for the relation.	Prevents question-only edges from being treated as equivalent to reported positive results.

Field	Allowed values / type	Meaning	Why it exists downstream
Explicitness (`explicitness`)	result_only / question_only / question_and_result / background_claim / implied	How explicitly the relation is framed in text.	Useful for separating central reported findings from implied or motivating claims.
Condition or scope (`condition_or_scope_text`)	string	Edge-level scope or subgroup qualifier.	Keeps conditional language attached to the relation rather than the node.
Claim text (`claim_text`)	string	Short normalized relation text.	Audit-friendly summary of the extracted edge.
Evidence text (`evidence_text`)	string	Short supporting excerpt or close paraphrase from the title/abstract.	Makes the edge inspectable in the public tool and in manual audits.
Direction of effect (`sign`)	increase / decrease / no_effect / ambiguous / NA	Reported sign of the relation when stated.	Supports later descriptive summaries and credibility splits.
Effect size (`effect_size`)	string	Reported magnitude if stated.	Retained for completeness and future extensions.
Statistical significance (`statistical_significance`)	significant / not_significant / mixed_or_ambiguous / not_reported / NA	Significance status stated in text.	Keeps reported evidence strength distinct from sign or design.
Evidence method (`evidence_method`)	enumerated method family	Main method family named or implied in the abstract.	Determines whether a relation enters the graph as directed causal or undirected contextual support.
Other method description (`evidence_method_other_description`)	string	Free-text description when method is `other`.	Preserves method specificity without exploding the method taxonomy.
Nature of evidence (`nature_of_evidence`)	quantitative / qualitative / mixed_methods / theoretical_or_conceptual / simulation / review_or_commentary / NA	Broad evidence type.	Helps later distinguish theory, empirical, and simulation-heavy slices.
Uses data (`uses_data`)	boolean	Whether the edge is supported by data use.	Simple empirical-versus-theory indicator.
Exogenous variation source (`sources_of_exogenous_variation`)	string	Explicit source of exogenous variation if named.	Reserved for future credibility-weighting extensions.
Tentativeness (`tentativeness`)	certain / tentative / mixed_or_qualified / unclear	How assertive the language is.	Keeps cautious claims distinct from strong declarative ones.

A.3 Design choices

Paper-local node reuse. The model is asked to reuse the same local node whenever the same concept genuinely recurs within a paper. That choice matters because downstream graph construction depends on whether a path inside one paper reuses a local concept consistently. If each mention were given a fresh node, the later concept graph would inherit spurious fragmentation before normalization even begins.

No transitive closure. The prompt explicitly forbids the model from creating $A \rightarrow C$ when the text only states $A \rightarrow B$ and $B \rightarrow C$. This is crucial for the benchmark. Missing direct links are the object of interest. If extraction itself created transitive closure, the benchmark would mechanically erase many of the very candidates it later wants to rank.

Directed versus undirected storage. The extraction schema stores one undirected relation using the first-mentioned concept as a storage convention, rather than duplicating it as two directed edges. This keeps contextual support distinct from directionally stated claims and avoids turning association language into artificial causal direction.

Keeping scope off the node label. Country, year, subgroup, and sample qualifiers are stored in dedicated context fields whenever possible. This is a normalization choice made early. A benchmark that ranks missing links between concepts needs concept identity to remain as stable as possible across papers. Local scope still matters for audit and interpretation, but it should not automatically become part of the canonical node label.

Separating `causal_presentation` from `evidence_method`. A paper can talk causally without using a strong design, and it can use a stronger design while describing the result more cautiously. Keeping these separate is what later allows the benchmark to define directed causal candidates by method while still auditing how papers describe those relations.

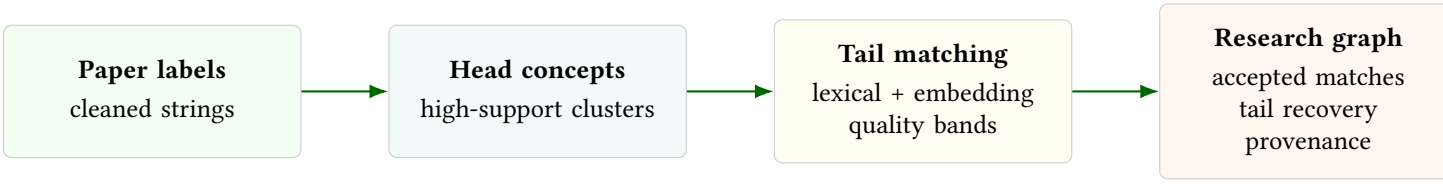
Separating `claim_status`, `explicitness`, `tentativeness`, and `edge_role`. These fields overlap in plain language but do different jobs downstream. `claim_status` records whether a finding is present, absent, mixed, or only posed as a question. `explicitness` records whether the relation is stated as a result, a question, background, or only implied. `tentativeness` records how strongly the paper speaks. `edge_role` records whether the relation is the main effect, a mechanism, heterogeneity, or something else. The benchmark and public tool both become much less legible if these distinctions are collapsed into one generic confidence flag.

Reproducibility. The prompts, schema files, extraction scripts, ontology pipeline, and paper-generation code used in this draft will be released at `https://github.com/prashgarg/frontiergraph`. The manuscript uses placeholder repository text because the repository will be made public at release time.

B Node normalization and ontology construction

Node normalization is a major measurement problem in its own right. Paper-local concept strings vary in wording, scope, and granularity. Candidate generation, path counts, and missingness all depend on node

Figure 14: Node normalization and concept matching



Notes. This figure summarizes the ontology pipeline used in the benchmark. High-support labels are clustered into head concepts, lower-support labels are matched with lexical rules and embedding-based ranking, and unresolved tail labels are recovered with stored provenance and quality bands. The key substantive point is that the fuller force-mapped corpus remains in the graph while mapping source and confidence are preserved.

identity. So unlike a field-level descriptive exercise, this paper cannot treat node definition as secondary.

B.1 Why a native ontology is needed here

Garg and Fetzer (2025) use the extracted paper-level objects for a different downstream task. The present paper asks a more node-sensitive question. Here the downstream object is a reusable concept graph in which a candidate next paper is a missing link between *specific concepts*. In that setting, concept identity cannot be handled at a broad field level. The distinction between public debt and public investment, or between monetary policy and energy consumption, is exactly the distinction the benchmark needs to preserve. The native ontology is therefore best understood as an extension of the earlier paper's extraction logic to a setting in which node identity does much more of the empirical work.

B.2 The implemented pipeline

The ontology build proceeds in stages.

Head-pool construction. The pipeline first constructs a head pool from the raw normalized label inventory using coverage and support logic. Labels can enter the head pool because they exceed support thresholds in distinct papers and journals, or because they are needed to cover enough of the observed mass of node instances. The code scores candidate head labels using support in papers, journals, instances, distinct partners, and distinct edge papers.

Accepted head concepts. Selected head labels are then clustered into accepted native concepts. Exact and reviewed same-label constraints are combined with blocked-pair and isolate constraints so that obviously compatible head labels cluster while problematic merges can be prevented. The accepted clusters become concept IDs of the form `FG3C...`, each with a preferred label and alias set.

Table 5: Mapping stages in the ontology pipeline

Stage	What it does	Why it is needed
Head pool	Selects high-support labels by coverage and support thresholds.	Defines a stable candidate set from which reusable native concepts can be built.
Accepted heads	Clusters compatible head labels into native concept IDs.	Creates the ontology's concept inventory.
Hard mapping	Uses exact, lexical-signature, and reviewed embedding rules.	Resolves easy cases conservatively before any softer inference.
Soft mapping	Uses shortlist or global embedding matching, plus unique lexical shortlists.	Maps the middle tail without forcing every label immediately.
Pending labels	Stores unresolved labels and why they failed to map.	Makes the missing tail auditable instead of silently dropping it.
Force-mapped tail recovery	Assigns unresolved labels to existing head concepts with stored score, margin, and quality band.	Expands benchmark coverage while preserving mapping provenance and confidence.

Hard mappings. Labels that are not already accepted heads are next mapped with conservative lexical and embedding-based rules. Exact accepted labels map directly. Otherwise the pipeline checks no-parenthesis signatures, acronym and punctuation signatures, singularized signatures, and a small reviewed embedding layer. These mappings are high-confidence enough to write into the hard mapping tables.

Soft mappings. The remaining labels move to the soft stage. If a label has an embedding and a shortlist of plausible head labels, the pipeline uses shortlist embedding matching. If a shortlist is ambiguous but the label still has an embedding, the pipeline uses a global embedding search over the embedded head inventory. Labels with a unique lexical shortlist but no embedding can also be soft-mapped lexically. Labels that still do not map are written into the pending table.

Pending labels and tail recovery. The long unresolved tail matters in this project because dropping unmapped labels changes the candidate universe and can mechanically over-concentrate the benchmark on well-mapped central concepts. The canonical benchmark therefore adds a force-mapped tail recovery stage. Unresolved labels are embedded in batches, matched to the existing head concepts, and written back as `force_embedding_backoff` mappings with stored cosine similarity, runner-up margin, and quality band. This stage is what lets the fuller benchmark retain 230,479 mapped papers rather than only the stricter mapped-core subset.

B.3 Why the force-mapped fuller corpus is now canonical

The fuller force-mapped corpus is the canonical benchmark because the alternative was to let mapping quality mechanically determine which literatures count as candidate-generating. That would have over-selected the cleanest, most repetitive, and most central concept strings. The force-mapped layer does introduce weaker mappings, which is why provenance and confidence are kept. But that is preferable to treating the well-mapped core as if it were the whole literature. The benchmark therefore uses the fuller graph while keeping mapping source and confidence available for later sensitivity checks.

C Benchmark construction and significance

Table 6 collects the main corpus counts that define the empirical universe. The two most important transitions are from the selected journal papers to papers with extracted edges, and then from raw extracted edges to normalized evaluation links. The latter step is what makes the missing-link benchmark coherent across papers.

Table 6: Corpus and normalization summary

Quantity	Count
Selected journal papers	242,595
Papers with extracted edges	230,929
Raw extracted edges	1,443,407
Normalized benchmark papers	230,479
Unique concepts in evaluation graph	6,752
Directed causal rows	89,737
Undirected contextual rows	1,181,277
Total normalized links	1,271,014

The mainline evaluation table is reported in the refreshed benchmark outputs and reproduced in the manuscript build. The important comparative fact is stable even before the longer appendix tables are read: preferential attachment remains stronger at the strict top-100 margin, while the graph score catches up materially as the shortlist grows. The significance tests are paired over common cutoff-year cells, with bootstrap confidence intervals reported for Recall@100 and mean reciprocal rank.

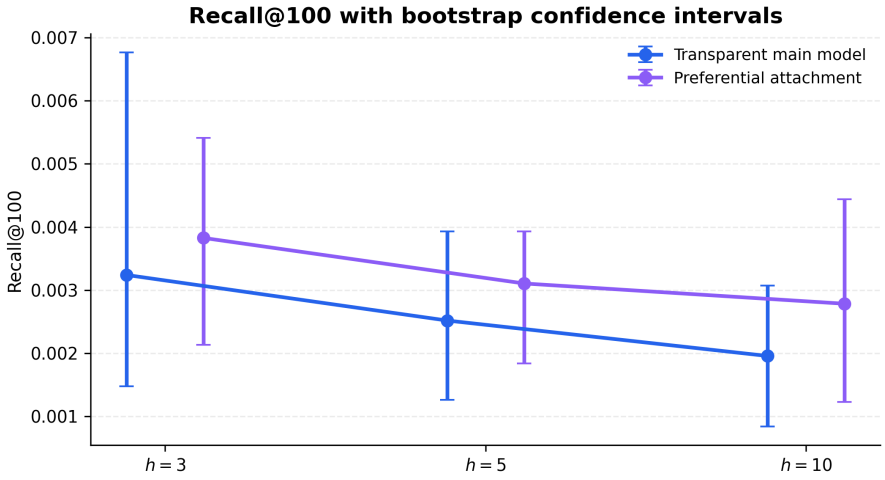
Table 7: Strict shortlist benchmark across the four main horizons

Metric	$h = 3$	$h = 5$	$h = 10$	$h = 15$
Recall@100, graph-based score	0.003239	0.002518	0.001956	0.001494
Recall@100, preferential attachment	0.003826	0.003105	0.002784	0.002138
MRR, graph-based score	0.000811	0.000524	0.000334	0.000227
MRR, preferential attachment	0.000901	0.000637	0.000420	0.000281

Table 8: Paired bootstrap comparison: graph-based score minus preferential attachment

Quantity	$h = 3$	$h = 5$	$h = 10$	$h = 15$
Δ Recall@100	-0.000587	-0.000588	-0.000828	-0.000644
p-value for Δ Recall@100	0.740	0.064	< 0.001	< 0.001
Δ MRR	-0.000090	-0.000113	-0.000086	-0.000054
p-value for Δ MRR	< 0.001	< 0.001	< 0.001	< 0.001

Figure 15: Main horizon comparison with bootstrap confidence intervals



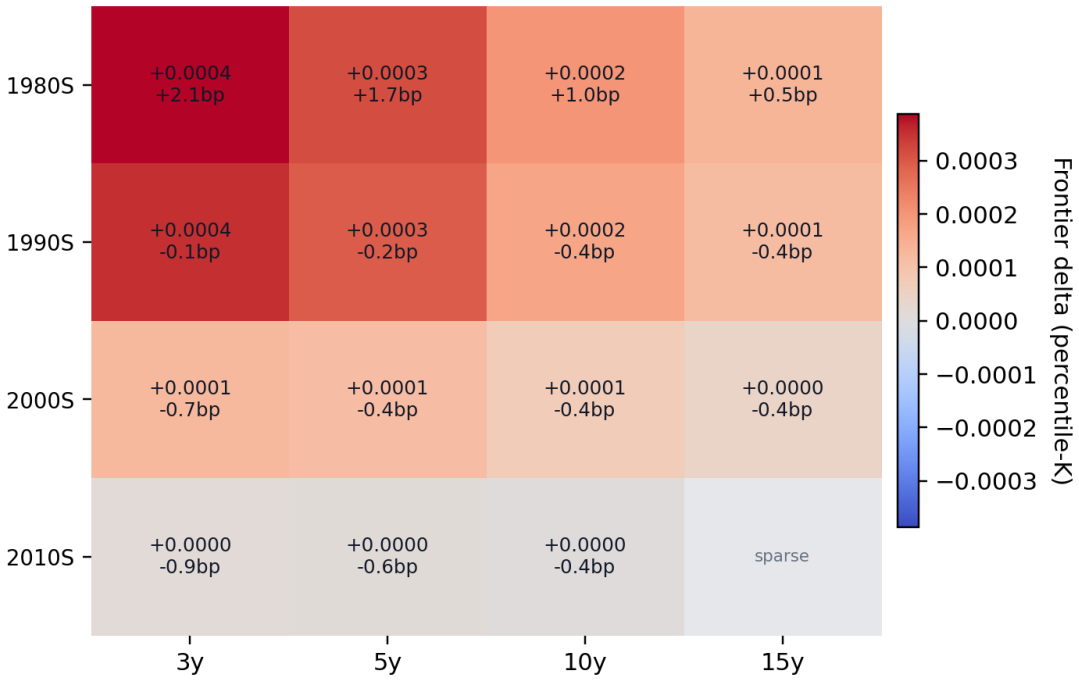
Notes. This appendix figure reports bootstrap confidence intervals for Recall@100 by model and horizon using the canonical evaluation table. The unit of analysis is the rolling cutoff-year benchmark cell. The confidence bands quantify uncertainty from the small number of pooled cutoffs, not uncertainty over the full universe of possible research questions.

The candidate-kind split also matters for interpretation. The direct benchmark is carried mainly by the directed causal task. The undirected contextual task is substantively useful for the fuller graph and for the heterogeneity atlas, but it contributes much less to strict top-100 shortlists because contextual relations are far more diffuse. That is why the main text can focus on directed causal emergence without pretending that the undirected structure is unimportant.

D Heterogeneity atlas extensions

The full heterogeneity atlas reports pooled and kind-split results over 5-year cutoffs, fixed-K and percentile-K frontiers, and a wider horizon set out to 20 years. The main text uses only the most interpretable cuts. The appendix records the rest so that the paper does not overclaim from one pooled average.

Figure 16: Time-period heterogeneity in the pooled frontier comparison



Notes. Rows correspond to cutoff-period bins and columns to horizons. Cell color reports the pooled percentile-frontier advantage of the graph score over preferential attachment. The annotations report the strict top-100 delta. The figure shows that the pooled frontier view remains more favorable than the top-100 view across much of the time dimension, but the advantage attenuates at longer horizons.

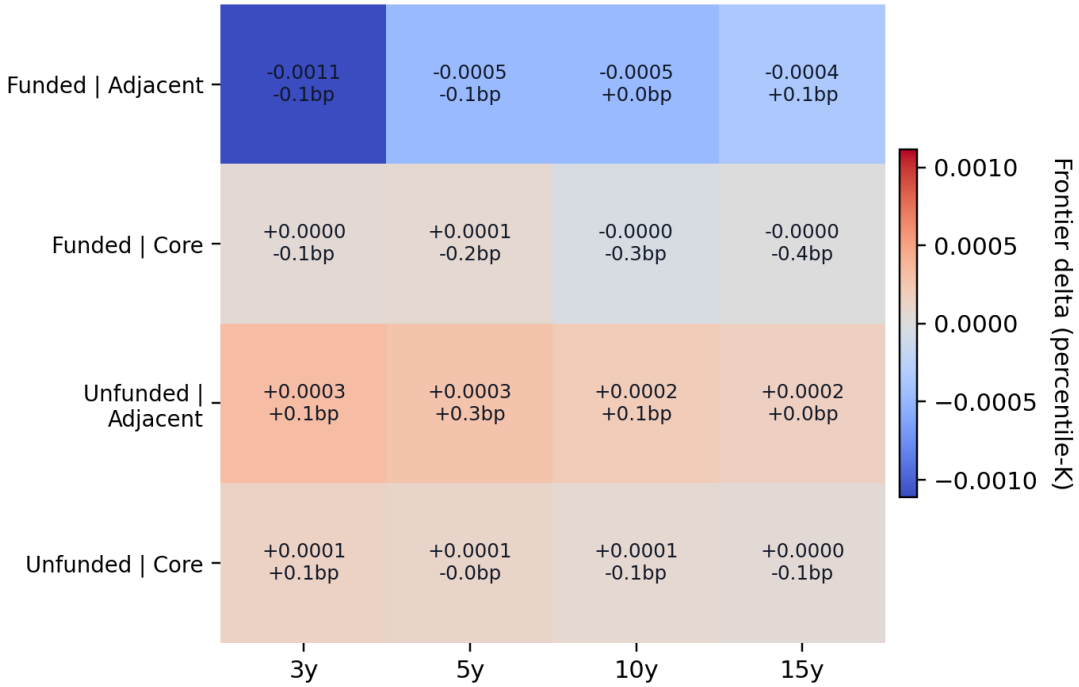
E Credibility audit summaries

The main score does not yet fully weight evidence quality, but the benchmark object is not blind to it either. The extraction layer already records stability, causal presentation, evidence type, and related claim metadata. The tables below should therefore be read as a quality audit of the empirical object rather than as a replacement ranking model. They show that directed causal rows are a smaller but relatively high-stability slice of the graph, that design-heavy method families remain well represented inside that slice, and that the current benchmark is not built from unstructured co-occurrence counts.

Table 9: Credibility audit by edge kind

Edge kind	Rows	Papers	Mean stability	Explicit-causal share
Directed causal	89,737	23,213	0.930	69.0%
Undirected contextual	1,181,277	221,192	0.868	45.3%

Figure 17: Funding and journal-tier interactions



Notes. This interaction view combines funding status and journal tier. The main text uses funding cautiously because coverage is uneven and because the resulting cells mix institutional composition with scientific behavior. The figure is still useful because it shows that several of the most popularity-dominated cells are funded-journal combinations rather than an undifferentiated funded literature.

F Path-evolution extensions

The main text focuses on the aggregate transition comparison, the journal split, the broad economics-versus-finance contrast, and a short table of path-rich examples. The appendix therefore only keeps the supplementary interpretation notes.

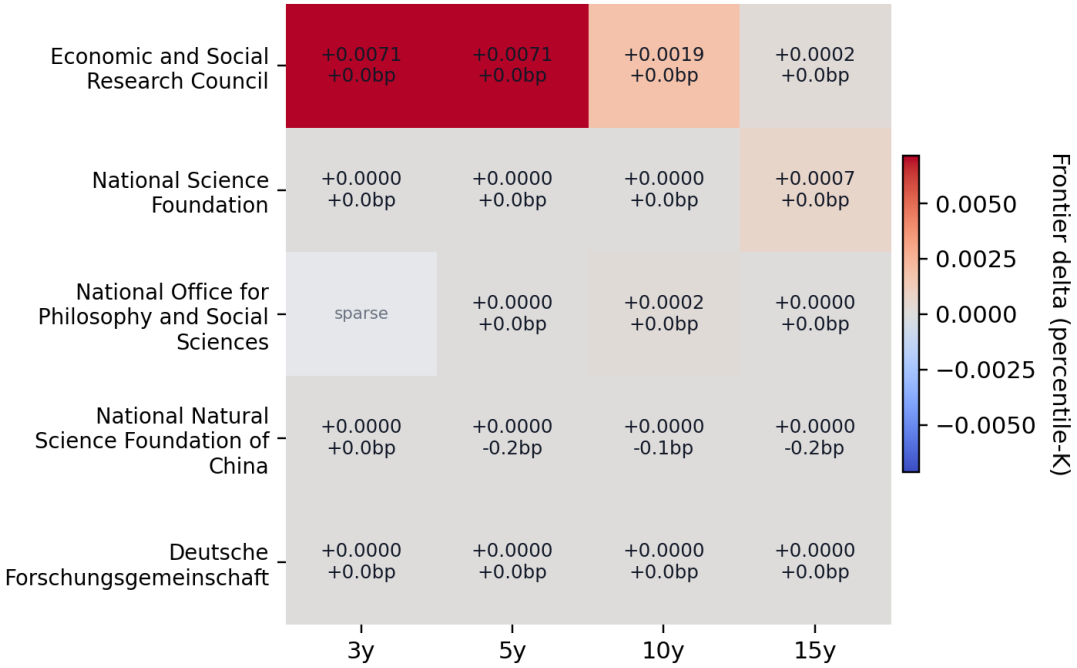
References

Albert-Laszlo Barabasi and Reka Albert. Emergence of scaling in random networks. *Science*, 286(5439): 509–512, 1999. doi: 10.1126/science.286.5439.509. URL https://www.science.org/doi/10.1126/science.286.5439.509.

Nicholas Bloom, Charles I. Jones, John Van Reenen, and Michael Webb. Are ideas getting harder to find? *American Economic Review*, 110(4):1104–1144, 2020. doi: 10.1257/aer.20180338. URL https://www.aeaweb.org/articles?id=10.1257/aer.20180338.

Santo Fortunato, Carl T. Bergstrom, Katy Borner, James A. Evans, Dirk Helbing, Stasa Milojevic, Alexander M. Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, Alessandro Vespignani, Ludo Waltman,

Figure 18: Stable top-funder extensions



Notes. Only high-support funders are shown. The stable set requires at least 800 future edges and at least three eligible cutoff cells across the main horizons. In the present atlas that keeps the Economic and Social Research Council, the National Natural Science Foundation of China, the Deutsche Forschungsgemeinschaft, and the U.S. National Science Foundation. The figure is an appendix result because funder-level interpretation mixes institution, geography, topic composition, and metadata coverage.

Dashun Wang, and Albert-Laszlo Barabasi. Science of science. *Science*, 359(6379):eaao0185, 2018. doi: 10.1126/science.aao0185. URL https://doi.org/10.1126/science.aao0185.

Prashant Garg and Thiemo Fetzer. Causal claims in economics. *arXiv preprint arXiv:2501.06873*, 2025. URL https://arxiv.org/abs/2501.06873.

ICLR. Iclr 2026 reviewer guide. https://iclr.cc/Conferences/2026/ReviewerGuide, 2026. Accessed March 2026.

Benjamin F. Jones. The burden of knowledge and the "death of the renaissance man": Is innovation getting harder? *Review of Economic Studies*, 76(1):283–317, 2009. doi: 10.1111/j.1467-937X.2008.00531.x. URL https://academic.oup.com/restud/article/76/1/283/1577537.

Derek J. de Solla Price. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, 27(5):292–306, 1976. doi: 10.1002/asi.4630270505. URL https://doi.org/10.1002/asi.4630270505.

Project APE. Project ape autonomous policy evaluation. https://ape.socialcatalystlab.org/, 2026. Accessed March 2026.

Table 10: Directed-causal credibility audit by evidence type

Evidence type	Rows	Mean stability	Explicit-causal share
Panel FE / TWFE	44,106	0.938	68.8%
Difference-in-differences	16,658	0.933	75.1%
Experiment	15,616	0.900	56.7%
Event study	6,247	0.940	73.5%
Instrumental variables	5,601	0.933	78.3%
Regression discontinuity	1,509	0.922	80.3%

Table 11: Stability-band shares by edge kind

Edge kind	High stability	Mid stability	Low stability
Directed causal	91.8%	5.6%	2.6%
Undirected contextual	85.4%	5.1%	9.5%

Refine. Refine – ai-powered research assistant. https://www.refine.ink/, 2026. Accessed March 2026.

Erzhuo Shao, Yifang Wang, Yifan Qian, Zhenyu Pan, Han Liu, and Dashun Wang. Sciscigpt: advancing human–ai collaboration in the science of science. *Nature Computational Science*, 2025. URL https://www.nature.com/articles/s43588-025-00906-6.

Stanford Agentic Reviewer. Tech overview. https://paperreview.ai/tech-overview, 2026. Accessed March 2026.

Brian Uzzi, Satyam Mukherjee, Michael Stringer, and Ben Jones. Atypical combinations and scientific impact. *Science*, 342(6157):468–472, 2013. doi: 10.1126/science.1240474. URL https://pubmed.ncbi.nlm.nih.gov/24159044/.

Dashun Wang and Albert-Laszlo Barabasi. *The Science of Science*. Cambridge University Press, 2021. doi: 10.1017/9781108610834. URL https://www.cambridge.org/core/books/science-of-science/572A745A6F97B55A263F5E86225E3F70.

Yue Zhang, Xingyu Fan, Hong Wang, Muhan Zhang, and Jie Tang. Exploring the role of large language models in the scientific method: from hypothesis to discovery. *npj Artificial Intelligence*, 1:14, 2025. URL https://www.nature.com/articles/s44260-025-00159-7.