

What Should Economics Ask Next?*

Prashant Garg
Imperial College London

12 April 2026

Draft for comments. Please do not cite without permission.

Abstract

How should an economist choose which open question to work on next? I build a directed literature graph from 242,595 published economics-facing papers spanning 1976 to 2026, rank still-open candidate questions using only the graph available at each historical date, and test those rankings against later publication. A learned reranker built on that graph achieves a 17.0% top-100 hit rate for mechanism-thickening questions and 10.7% for direct-closure questions at a five-year horizon, versus 1.7% for both under preferential attachment. The two families differ in what the screening recovers: mechanism-thickening yields denser shortlists, while direct-closure captures a larger share of all eventual realizations. These results establish that local literature structure contains useful upstream screening information—and that economics moves by deepening mechanisms around existing claims at roughly 3–12 times the rate at which it closes locally implied direct relations.

Keywords: research direction, economics of science, scientific discovery, question choice, literature graphs

JEL codes: O31, O33, C45, C81

1 Introduction

Choosing what to work on is one of the least formalized decisions in economics. We have disciplined tools for identification, estimation, and inference. We have much less guidance 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: as the stock of published work grows, it becomes harder to see which question is worth pursuing next. Foster, Rzhetsky, and Evans (2015) add a second concern. Scientists overwhelmingly choose conservative strategies even though riskier exploration disproportionately

*For helpful comments and discussions I thank Elliott Ash, Jasmin Baier, Thiemo Fetzer, Thomas Graeber, Paul Hünermund, Ralf Martin, Chris Roth, Zekai Shen, and seminar participants at the CEPR–MPWZ Text-as-Data workshop. All remaining errors are my own.

generates high-impact work. The problem is not only that question choice is hard. It is also tilted toward the familiar.

That problem becomes sharper when AI lowers the cost of downstream research tasks such as writing, background research, data analysis, and coding (Korinek, 2023). Agrawal, McHale, and Oettl (2024) make the complementary point in a model of prioritized search: once ranking hypotheses becomes cheaper and better, the value of choosing what to test shifts further upstream. The question here is narrower than “what should economics study?” in the philosophical sense. It is whether the structure of past research can help rank candidate next questions in a disciplined way.

This paper builds a directed literature graph from 242,595 published economics-facing papers and tests whether that graph can predict which still-open questions later become part of the literature. At each historical date, I freeze the graph at $t - 1$, rank open candidate questions using only information available at that date, and check which ones appear in later published work. Comparing a popularity baseline, a transparent graph score, and a learned reranker, the answer is yes: a learned reranker achieves a 17.0% top-100 hit rate for mechanism-thickening questions and 10.7% for direct-closure questions at a five-year horizon, versus 1.7% for both families under preferential attachment. The gains also differ in form. Mechanism-thickening yields denser shortlists (more later realizations per 100 suggestions), while direct-closure captures a larger share of all eventual realizations and ranks them earlier on average. The graph is therefore most useful as a tool for directing scarce reading time, not as a substitute for broad reading or field knowledge.

The paper studies two kinds of candidate question, which it calls *direct-to-path* and *path-to-direct*. In a direct-to-path case, the literature already contains a direct relation, but later work adds a clearer mechanism around it. In a path-to-direct case, the literature already contains nearby support for a relation, but the direct link itself is still missing and only appears later. These are different scientific moves. Direct-to-path asks where an accepted claim is still underexplained. Path-to-direct asks where nearby evidence already points toward a direct relation that the literature has not yet stated explicitly. An example: if the literature already links broadband access to business formation, later work adding a search-friction channel is a direct-to-path move. If the literature already connects trade liberalization to imported inputs and imported inputs to firm productivity, but has not yet stated a direct trade-to-productivity link, later work stating that link is a path-to-direct move.

The paper also documents the descriptive balance between these two motions. Economics more often adds mechanisms around existing claims than closes locally implied direct relations—at roughly 3–12 times the rate, depending on the horizon and era. That asymmetry is consistent with a scientific literature that deepens accepted lines rather than replaces them, and it bears on the interpretation of both shortlist families: the more abundant family is also the one for which the screening tool is most practically useful.

The roadmap is as follows. Section 2 positions the paper in four related literatures. Section 3 describes the corpus, extraction, and normalization pipeline. Section 4 defines the two question objects, the transparent scoring rule, and the learned reranker. Section 5 presents the paired shortlist comparison, the direct-to-

path/path-to-direct asymmetry, and the reranker decomposition. Section 6 interprets the findings and discusses implications. The appendix provides extraction details, full model comparisons, paired extensions showing the direct-to-path/path-to-direct trade-off across reading budgets, credibility audits, and supplementary usefulness checks.

2 Related Literature

Four literatures bear on this paper, and the contribution is clearest against all four.

The first is the economics of ideas, discovery, and research direction. Bloom et al. (2020) document rising research effort alongside falling research productivity, while Jones (2009) studies how accumulating knowledge changes the organization of innovative activity.¹ Those papers focus on the cost of reaching the frontier. However, a related literature asks whether incentives and norms steer researchers toward safe topics. Azoulay, Graff Zivin, and Manso (2011) show that tolerating early failure produces more novel and high-impact work. Foster, Rzhetsky, and Evans (2015) show that conservative strategies dominate in practice even though riskier exploration disproportionately generates breakthroughs.² Although those papers study incentives at the system level, they motivate a narrower, operationally tractable question: given a large existing literature, can its local structure help screen candidate next questions in a way that is less path-dependent than cumulative advantage alone?

The second literature is science-of-science work using large-scale scientific data to study novelty, impact, and frontier formation. Fortunato et al. (2018) and Wang and Barabasi (2021) provide overviews of this field. Two specific findings are relevant here. Uzzi et al. (2013) show that novelty often combines conventional structure with a small number of atypical combinations. Park, Leahey, and Funk (2023) document a broad decline in disruptive science since the mid-twentieth century.³ Although this literature is relevant in spirit, the empirical object here is different. I do not measure novelty from citations or reference-pair combinations. I define candidate next questions as missing links in a directed claim graph and evaluate them prospectively against later publication.

The third literature covers network growth, link prediction, and learning to rank. Price (1976) and Barabasi and Albert (1999) show why already connected nodes tend to attract more links—a finding that here becomes the main null rather than a decorative comparison. The broader link-prediction literature (Liben-Nowell and Kleinberg, 2007; Martinez et al., 2016) and the learning-to-rank tradition (Liu, 2009) show that learned models can improve on fixed heuristics. More directly, recent work treats missing links in scientific knowledge graphs as candidate research directions. Krenn and Zeilinger (2020) and Gu and Krenn (2025)

¹The “ideas are getting harder to find” framing is itself debated. Fort et al. (2025) argue that what is declining is the translation of ideas into measured growth, not the rate of idea generation itself.

²Bhattacharya and Packalen (2020) argue that citation-based incentives have further shifted scientists away from new ideas toward established topics, creating a self-reinforcing momentum that a screening tool could partially offset.

³Petersen, Arroyave, and Pammolli (2024) argue that part of the measured decline reflects citation inflation rather than a real shift in innovativeness. For this paper, the relevant implication is narrower: whether or not measured disruption is declining, the literature’s structure visibly thickens around existing claims over time, and that thickening is exactly what the path-development results in Section 5.2 measure.

propose knowledge-graph based discovery benchmarks; Sourati et al. (2023) and Rzhetsky et al. (2015) connect link prediction to scientific discovery. The present paper differs from all of these in three ways: it uses a directed economics claim graph, evaluates candidate questions in walk-forward vintages, and frames the problem as one of scarce reading time rather than generic link prediction. A nearby but distinct branch asks not which missing relation later appears, but which disconnected literatures or rising clusters are worth monitoring. Swanson’s literature-based discovery program (Swanson, 1986) and emerging-topic systems (Blei and Lafferty, 2006; Chen, 2006; Small et al., 2014) fall in this branch. The object here is more economist-facing: a ranked list of directed candidate questions evaluated against realized publication.

The fourth literature is AI-assisted scientific work. Recent systems help with hypothesis generation, literature synthesis, and manuscript support (Zhang et al., 2025; Shao et al., 2025; Korinek, 2023).⁴ Si, Yang, and Hashimoto (2024) find that LLM-generated research ideas are rated more novel, though slightly less feasible, than expert-generated ones. Agrawal, McHale, and Oettl (2024) formalize AI-assisted discovery as prioritized search over a combinatorial hypothesis space. d’Aquin (2025) argues that knowledge graphs are becoming scaffolds for AI-based scientific discovery.⁵ The closest social-science comparisons in this space are Tong et al. (2024), who extract causal relations from 43,312 psychology papers and use knowledge-graph link prediction to generate hypotheses, and Lee et al. (2025), who construct a 10,490-triplet benchmark from economics and finance studies. The present paper differs in scale, in using prospective walk-forward evaluation rather than expert judgment, and in explicitly separating the benchmarkable historical object from the surfaced question—the human-readable candidate prompt built from the local graph neighborhood.

The present paper uses AI-extracted paper-level structure as an enabling layer and then asks an economics question: can that structure be turned into an inspectable, prospectively testable object for deciding what to read and work on next? Table 3 summarizes how this paper differs from the closest comparable systems. The positioning is economics-first metascience with a graph-based empirical object, not a general-purpose AI assistant or a claim-extraction model.

3 Data, Extraction, and Normalization

The sample covers 242,595 published economics-facing papers from 300 journals spanning 1976 to early 2026, at the directed concept-pair level. The pipeline extracts a paper-level claim graph from each title and abstract, matches repeated concept labels across papers into a shared node inventory, and then ranks missing directed links for prospective evaluation. Of those papers, 230,929 contain at least one extracted edge, yielding 1,443,407 raw extracted edges. After normalization and graph construction, the benchmark

⁴Contemporary deployed examples include Refine (<https://www.refine.ink/>), the ICLR 2026 Reviewer Guide (<https://iclr.cc/Conferences/2026/ReviewerGuide>), Stanford Agentic Reviewer (<https://paperreview.ai/tech-overview>), and Project APE (<https://ape.socialcatalystlab.org/>). These are useful examples of the current tool landscape, but they are not treated as archival scholarly references in the bibliography.

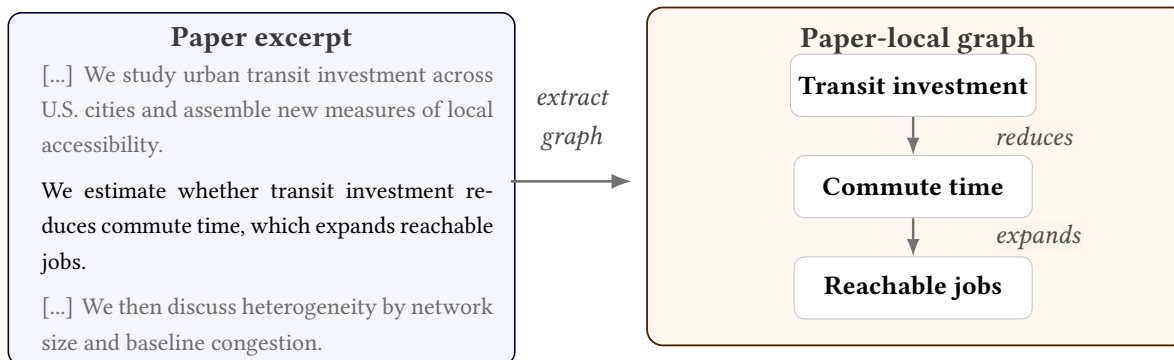
⁵A related concern is that data-driven tools may narrow exploration rather than broaden it. Hoelzemann et al. (2024) show that when data highlights attractive but suboptimal paths, it can suppress breakthrough discovery. A screening tool has to be honest about that risk. The attention-frontier and heterogeneity results in the appendix speak to it directly.

graph retains 230,479 papers and 1,271,014 normalized links. The frozen ontology—a fixed vocabulary of canonical concept labels used for cross-paper node matching and audit—contains 154,359 rows and is a separate object from the live benchmark graph. The next figures walk through the pipeline in order: one paper-level extraction, cross-paper matching into a shared graph, one live graph neighborhood, and the difference between the benchmark event and the richer question a reader inspects.

3.1 Corpus

The corpus covers the top 150 core economics journals and the top 150 adjacent journals, selected by field-weighted citation impact (FWCI). Work-level metadata, source metadata, and journal assignments come from OpenAlex.⁶ The sample uses published journal papers rather than working papers, drafts, or preprints, because the goal is to study realized scientific structure in an economics-facing literature. That choice sacrifices some freshness in exchange for clearer source control, more stable metadata, and a graph that is easier to interpret as realized research rather than a mix of partially filtered text. The FWCI-based selection rule ties the retained sample to a published-literature core with credible economics or adjacent relevance. The corpus is not meant to exhaust the frontier. It defines a reproducible, interpretable literature against which missing links can be backtested. Appendix G summarizes the corpus waterfall and retention numbers.

Figure 1: A paper excerpt and its local graph



Notes. This figure isolates the extraction unit. One paper becomes one paper-local graph. The point is to show what is recovered before any cross-paper matching or normalization. The example is lightly cleaned from a live economics-facing transport-labor question, but the layout is stylized to make the extraction step legible.

3.2 Paper-local graphs

The extraction layer builds on Garg and Fetzer (2025). Each title and abstract is converted into a paper-level graph in which nodes are extracted concepts and edges summarize the relations the paper itself states, studies, or reports (Figure 1). The present paper extends that idea in three ways. First, the schema is broader

⁶OpenAlex 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.

than explicit causal claims, because the benchmark also needs undirected contextual support.⁷ Second, the schema separates the paper’s *causal presentation* from the *evidence method* used to support a claim. Third, the graph stores contextual qualifiers in dedicated fields rather than forcing them into the node label. The goal is not just to recover whether a paper makes a claim. It is to recover enough local structure that the claim can later be used inside a reusable concept graph. Code, prompt files, and release materials are available in the public repository at <https://github.com/prashgarg/frontiergraph>.

The decision to work paper by paper first is deliberate. At extraction time, the task is not to solve global concept matching inside the language model. It is to recover each 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 D. **The code and prompt files used in this paper will be released at github.com/prashgarg/frontiergraph.**

3.3 Node normalization

Node normalization is central to the design 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 uses a fixed ontology and an auditable matching procedure rather than letting every paper-local label become its own global node or allowing unconstrained merging. The ontology combines JEL, an economics-filtered Wikidata pull, OpenAlex topics, OpenAlex keywords, and an economics-filtered Wikipedia crawl. The frozen baseline contains 154,359 rows. Matching then proceeds in stages. Exact lexical cases are resolved deterministically first. Harder cases use embedding retrieval, but the system keeps multiple candidates and reviewed overlays rather than forcing one irreversible merge.⁸ The practical aim is simple: merge obvious duplicates, avoid false precision, and keep uncertain cases visible for later audit rather than hiding them inside one aggressive normalization pass. Appendix E gives the full algorithm and the retention counts. Table 1 fixes the key notation used throughout the paper.

⁷This is a deliberate departure from Garg and Fetzer (2025). Their object is the rise of credible causal language and design in economics, so the stricter identified-causal layer is the natural headline object there. Here the downstream task is broader research allocation over candidate questions, so the main target is a wider causal-claim layer, while the stricter identified-causal layer is retained as a nested credibility-oriented benchmark.

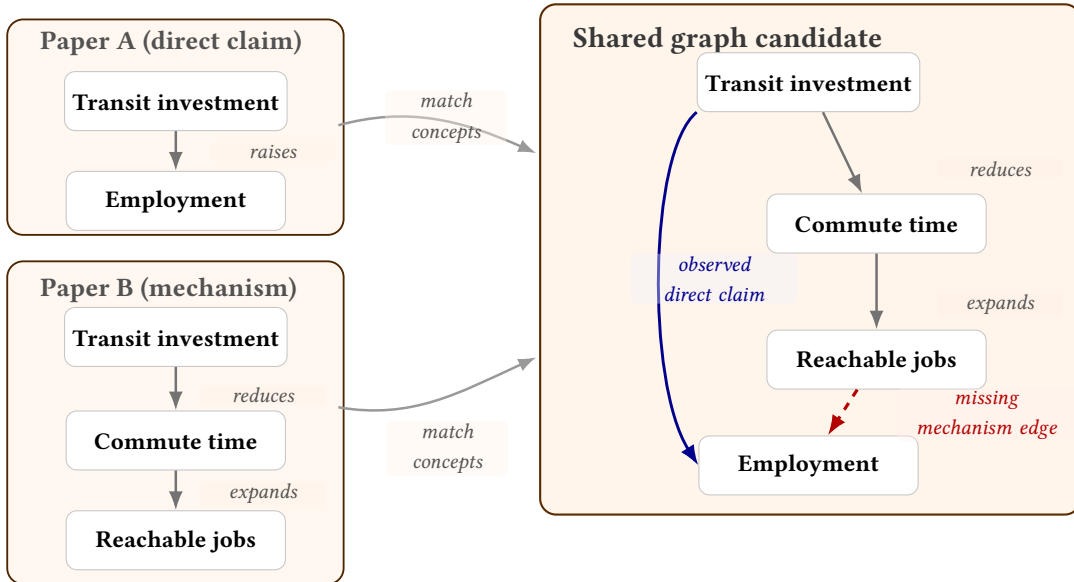
⁸The embedding step is used as a retrieval and ranking device after exact lexical passes, not as an unconstrained merge rule. That ordering keeps obvious cases deterministic and makes softer matches auditable.

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.
$u \rightarrow v$	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.

Notes. This table collects the symbols used repeatedly in the benchmark and scoring sections. The dated graph G_{t-1} is always the graph observed through year $t - 1$, so every score, feature, and candidate is constructed without seeing papers from year t or later. h is the forecast window in years, and K is the number of candidate links or surfaced questions a reader is willing to inspect.

Figure 2: Cross-paper matching reveals a shared candidate neighborhood

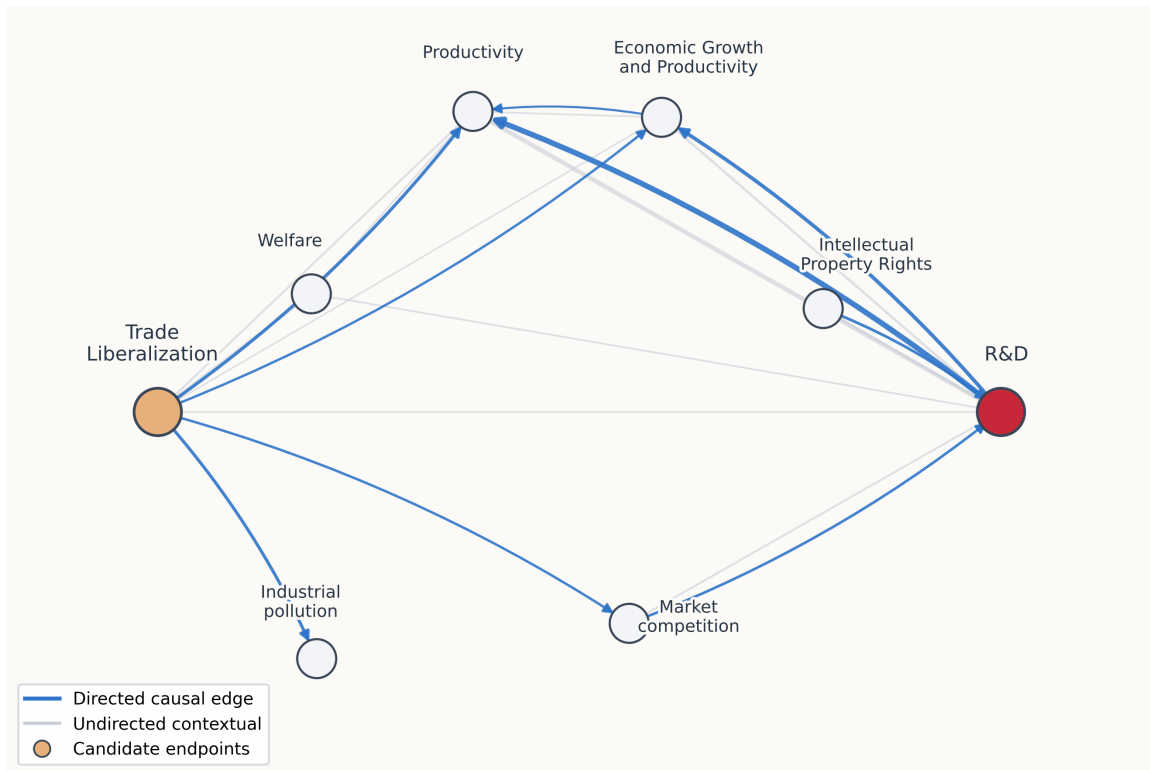


Notes. This figure isolates the cross-paper step. The two left panels are separate paper-local graphs. Only after repeated labels are matched into shared concepts does the right panel exist. The example is a direct-to-path case: one paper contributes a direct claim, another contributes a mechanism fragment, and the shared graph reveals a missing mechanism edge that can be benchmarked prospectively.

4 Question Objects and Evaluation

A natural first instinct is to ask a large language model directly: list the concept pairs likely to be studied next, or judge whether a given pair will be connected. The difficulty is one of evaluation rather than

Figure 3: A local neighborhood from the live graph



Notes. The previous figures are conceptual examples. This figure is a real neighborhood from the live graph, built around the candidate *Trade Liberalization* \rightarrow *R&D*. Blue arrows are directed causal edges; gray lines are undirected contextual support. The figure is not a result in itself. Its job is to show that the benchmark ultimately operates on graph neighborhoods of this kind rather than on hand-drawn concept diagrams.

prediction. A general-purpose model has been trained on much of the period one would hope to predict, so any accuracy figure is confounded by leakage from the test window into the training corpus (Kapoor and Narayanan, 2023; Carlini et al., 2023). It has no calibrated base rate at which pairs become linked, so shortlists cannot be compared against a top- K realization count. And it cannot be re-run at strict pre-training cutoffs, which rules out walk-forward evaluation. Semantic-similarity baselines over concept embeddings inherit the same problems. What follows in this section is designed around those constraints: a dated support graph, a transparent structural score, and a learned reranker evaluated walk-forward against realized future links.

The historical backtest requires separating two objects: the dated graph event that can be tested prospectively, and the reader-facing question built from the same neighborhood. Figure 3 shows one real neighborhood of the underlying concept graph. The remainder of this section makes the distinction precise.

4.1 Two question objects

The graph supports two historical families of candidate question. In a *direct-to-path* case, the literature already contains a direct relation at date $t - 1$, but the surrounding mechanism is still thin. Later work adds a mediating path around that relation. In a *path-to-direct* case, the literature already contains a local path at date $t - 1$, but the direct relation itself is still missing. Later work states that direct relation explicitly.

The distinction matters because the two families ask different things of a literature. Direct-to-path asks where an accepted relation is still underexplained. Path-to-direct asks where nearby evidence already points to a direct claim the literature has not yet stated. Direct-to-path is the more immediately readable family for most economists, since much empirical work moves from reduced-form relations to channels, heterogeneity, and scope conditions. But path-to-direct matters too: some later papers close a missing direct relation rather than only elaborate an accepted claim.

The headline benchmark is directed. At cutoff t , the candidate universe is the set of directed causal links that are still missing from the graph observed through year $t - 1$. Undirected contextual support remains important: it scores and interprets the directed candidates without itself being ranked.

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

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

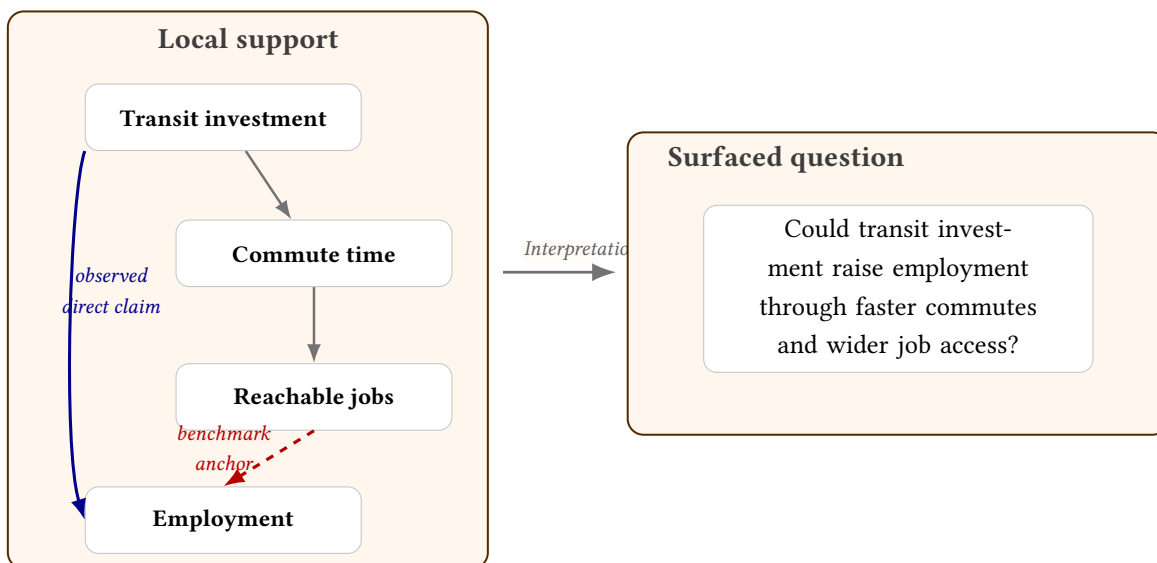
where E_{t-1}^D denotes the directed causal subgraph observed through year $t - 1$. A future realization occurs when $u \rightarrow v$ first appears during $[t, t + h]$. At each cutoff, take the active concepts, form all ordered source-target pairs, and drop any pair whose directed causal edge is already in the graph. What remains is the set of still-missing direct claims the paper can rank at date t . A candidate counts as realized only when that direct claim first appears within the horizon window, not when the concepts merely keep co-appearing.

The historical event and the surfaced question are not the same. The historical event is whether the relevant direct edge or path appears later in the graph. The surfaced question is the human-readable prompt built around the local neighborhood. That separation matters because economists do not browse bare graph events. They browse questions that could plausibly become papers (Figure 4).

The surfaced question stays centered on the endpoints—or on the endpoints plus one mediator—because the benchmark requires one dated event and the reader needs one inspectable question. A fuller local neighborhood may contain several paths and overlapping readings, but those richer motifs enter as supporting evidence rather than as the event itself.

The main backtest also holds the active concept set fixed at the cutoff. A distinct extension is *node activation*: dormant ontology concepts, weakly grounded phrases, or genuinely new concepts entering the active graph later. That object matters, but it requires a different dated event than missing-link appearance and is therefore better treated as a separate frontier problem.

Figure 4: The benchmark anchor is narrower than the surfaced question



Notes. The dashed edge inside the left panel is the benchmark anchor that can be backtested. The reader-facing question is richer: it keeps that same anchor but presents the surrounding support graph as a readable question. The surfaced question is therefore a compression of the neighborhood evidence, not the benchmark event itself.

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 gap-boundary split matters for interpreting the screening results: the reranker surfaces a different gap-boundary mix than the popularity benchmark (Section 5).

4.3 Transparent scoring rule

The ranking rule combines four ingredients: path support, underexploration gap, motif support, and hub penalty. Using the same transit-investment example that runs through the earlier method figures, path support asks whether the two endpoint 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.

Formally, the score for a candidate pair (u, v) is a weighted sum of four components:

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

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 $\delta = 0.2$. Those weights are fixed by design rather than tuned to maximize forecasting performance. 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 5 and 6 show directly what the rule rewards: multiple short paths, repeated local support, and a direct link that still looks underfilled relative to its neighborhood. The notation is easier to read if each term is treated as answering one question. Is there already short local support connecting the endpoints? That is \tilde{P} . Does the direct claim still look missing relative to that support? That is G . Is the signal repeated in more than one nearby motif rather than hanging on one fragile path? That is \tilde{M} . Are the endpoints simply very generic hubs that would connect to almost anything? That is \tilde{H} , which is subtracted rather than added. So the score is not an opaque index. It is a transparent compromise between support, missingness, repeated local confirmation, and a penalty for generic popularity.

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 choice is deliberate. The 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.

4.4 Learned reranker

The transparent score is fixed-weight and inspectable. Its weights are chosen by design rather than tuned to historical outcomes, which keeps the score readable question by question. The cost is that fixed weights may miss combinations of graph features that matter for screening. The learned reranker asks a narrower question: if the candidate set and time discipline stay fixed, does a model that learns how to weight the same graph information improve the shortlist? The reranker uses the same missing-link candidates, only information available at the historical cutoff, and no text, author, or institutional features.

Feature families. The reranker uses five nested feature families. The base is the transparent graph score itself. Two further families add local graph structure—path support, motif counts, mediator counts, co-occurrence signals, endpoint degree products, and a same-field indicator—and timing signals: support age, recency of the most recent supporting edge, and recent-window degree and incident counts for each endpoint. The final two families cover evidence composition (mean edge stability, evidence-type diversity, venue diversity, source diversity, and mean field-weighted citation impact at each endpoint) and boundary

and gap indicators (whether the endpoints sit in different field groups with no co-occurrence, whether the pair looks gap-like, and the local closure density around the pair). The richest specification uses 34 graph-derived features; Table 15 in Appendix H lists the full inventory. The ordering is cumulative so each nested specification’s marginal contribution can be isolated.

Training design. Training follows the same walk-forward discipline as the main benchmark. At each evaluation cutoff t , the model is trained only on earlier cutoff-year cells and is evaluated at t . Features are computed from the historical corpus through year $t - 1$. The positive label is whether the candidate edge first appears within the evaluation horizon. I test two model families: logistic regression with class-balanced weights, and a pairwise ranking model trained on positive-versus-negative feature differences. Both use L_2 regularization. Because the feature families are nested, the paper can ask a simple question at each step: does adding more graph information improve screening beyond topology alone? Appendix H gives the full training design, regularization grid, best configurations by horizon, and grouped feature decompositions.

Scope. The reranker stays inside the graph-screening framing: every feature comes from the literature graph (no text, no author identity, no external data), the candidate set is the same set of missing links, and the temporal discipline is the same walk-forward design.⁹

4.5 Walk-forward evaluation

The prospective design freezes the graph at year $t - 1$, ranks candidates using only information available at that date, and checks whether those links first appear over the evaluation horizon.¹⁰ This keeps the benchmark close to the ex ante decision problem rather than retrospective fit. The question is not “how well does the score fit the historical graph?” but “how well would this score have surfaced links that later became realized work?” Retrospective fit can reward mechanical regularities that carry no information about research allocation; the walk-forward design cannot.

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 $t + h \leq 2026$; the appendix horizon extension applies the same rule on a five-year cutoff grid when it extends the transparent benchmark to $h = 20$. The headline benchmark focuses on directed causal candidates. The fuller atlas also

⁹The difference is only in how the graph features are combined: by fixed design judgment in the transparent score, or by weights learned from earlier cutoffs in the reranker. A separate design question is how far out in the support graph to look when generating candidates. The main benchmark uses `max_path_len = 3`, allowing paths of length two (one mediator) and three (two mediators). Appendix I documents the sensitivity of the results to that choice, comparing path lengths 2 and 3 on the same tuned pipeline. Appendix I.5 makes that comparison concrete with four curated endpoint pairs (one per cluster), and Appendix I.5 extends the illustration to full per-source top-20 target sweeps for eight anchor concepts. Appendix H includes a Shapley-value decomposition showing which features drive the combined model’s predictions.

¹⁰The model at year $t - 1$ cannot use information from papers appearing in t or later—that is the sense in which the design avoids leakage.

evaluates directed and undirected objects separately and pools them by weighted aggregation rather than constructing one mixed ranking universe. Figure 7 makes the temporal discipline visual.

Benchmark objects and metrics. The paper keeps two dated historical objects side by side. In path-to-direct, a local path already exists and the later event is the first appearance of the direct link. In direct-to-path, a direct relation already exists and the later event is the first appearance of a supporting path. *Future links per 100 suggestions* is the shortlist hit rate: Precision@K, expressed per 100 suggestions so it stays comparable across reading budgets. Recall@100 asks what share of all later realizations in that family appear in the top 100. Mean reciprocal rank rewards putting later realizations nearer the top. These measures should be compared within family first, and then across families with the denominator difference in mind.

Preferential attachment as benchmark. The first benchmark is a cumulative-advantage rule. It asks whether one can do well simply by following existing traffic in the graph. A candidate looks better when it joins two endpoints that already attract many links. The standard network name for that logic is *preferential attachment*. In this paper it scores a candidate ordered pair by source out-degree times target in-degree:

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

It is a serious benchmark here, not a decorative one. Scientific attention is not allocated by neutral exploration alone. Visible topics, familiar data, established methods, and a larger installed literature often attract still more work. If a graph-based screen cannot beat that rule, then future question choice is mostly a popularity process. The appendix reports stronger transparent variants. The main text asks the simpler question of whether graph structure adds screening value beyond cumulative advantage.¹¹

Co-occurrence as a benchmark. The second benchmark is a co-mention rule. It asks whether one really needs directed claim structure at all. Perhaps it is enough to know that two endpoint concepts appear together repeatedly in the same papers, even if one ignores direction, claim type, and local path structure. That is also a serious null in this setting. Many literatures are organized less by explicit causal direction than by repeated joint attention: topics, variables, and mechanisms that are often discussed together keep reappearing together. If that is the main force in the data, then a simple co-occurrence score should already recover much of what later gets studied. Much of the recent work on predicting future research connections takes exactly that route. Krenn and Zeilinger (2020) and Gu and Krenn (2025), for example, work with undirected co-occurrence, while related scientific-discovery systems often begin from co-mentioned entities or undirected knowledge links (Rzhetsky et al., 2015; Sourati et al., 2023). That

¹¹The label “preferential attachment” comes from the network-growth literature. In that literature, nodes that already have many links are more likely to receive new ones (Price, 1976; Barabasi and Albert, 1999). Readers who do not use the term can think of it more simply as the graph analogue of cumulative advantage or a rich-get-richer process. In a science setting, the same logic says that already visible papers, concepts, or topics tend to attract still more connections. That is exactly why it is the right null in this paper.

approach has clear advantages: it requires no extraction of causal direction or claim metadata, and it scales easily. The question here is whether the harder extraction of directed claim structure adds screening value beyond what co-mention alone delivers. I therefore include a co-occurrence baseline that scores each candidate pair by how many papers mention both endpoints, ignoring direction, claim type, and path structure.¹²

Preferential attachment is the cumulative-advantage null. Co-occurrence asks whether undirected co-mention is enough. The transparent graph score is the readable graph baseline. The learned reranker is the strongest graph-based screen on the same candidate set.

Fixed-budget retrieval. The evaluation problem is not whether a useful question appears somewhere deep in a long ranking. It is whether useful questions appear in the first set a reader can plausibly inspect. Researchers do not read 5,000 candidates. Editors do not review 5,000 possibilities for a special issue. Funders do not investigate an unlimited menu of exploratory ideas. The relevant object is therefore a fixed reading budget: what does the graph deliver in the first 50, 100, or 500 suggestions that scarce attention can actually reach? That is why the paper emphasizes Recall@100, other fixed-budget shortlist measures, and frontier-style comparisons over larger K . Those metrics are not a cosmetic presentation choice. They correspond to the real screening problem the paper is trying to inform.

Horizon choice. The main horizons are 5, 10, and 15 years. Five years is a natural first window in economics because publication and diffusion are slow relative to many neighboring fields (Hadavand et al., 2024). But a five-year window is not enough to see many slower adjustments in topic, method, and citation uptake. Economics papers often keep accumulating attention over long citation life cycles, which is one reason to look beyond only the short run (Hamermesh, 2018). Ten years therefore captures slower movement in questions that take time to propagate through papers, methods, and field conventions. Fifteen years remains empirically useful because many slower literatures are still only partly visible at shorter windows. Three years and twenty years are kept as appendix extensions, but the main benchmark comparison is now centered on 5, 10, and 15 years.

5 Results

5.1 Paired shortlist comparison

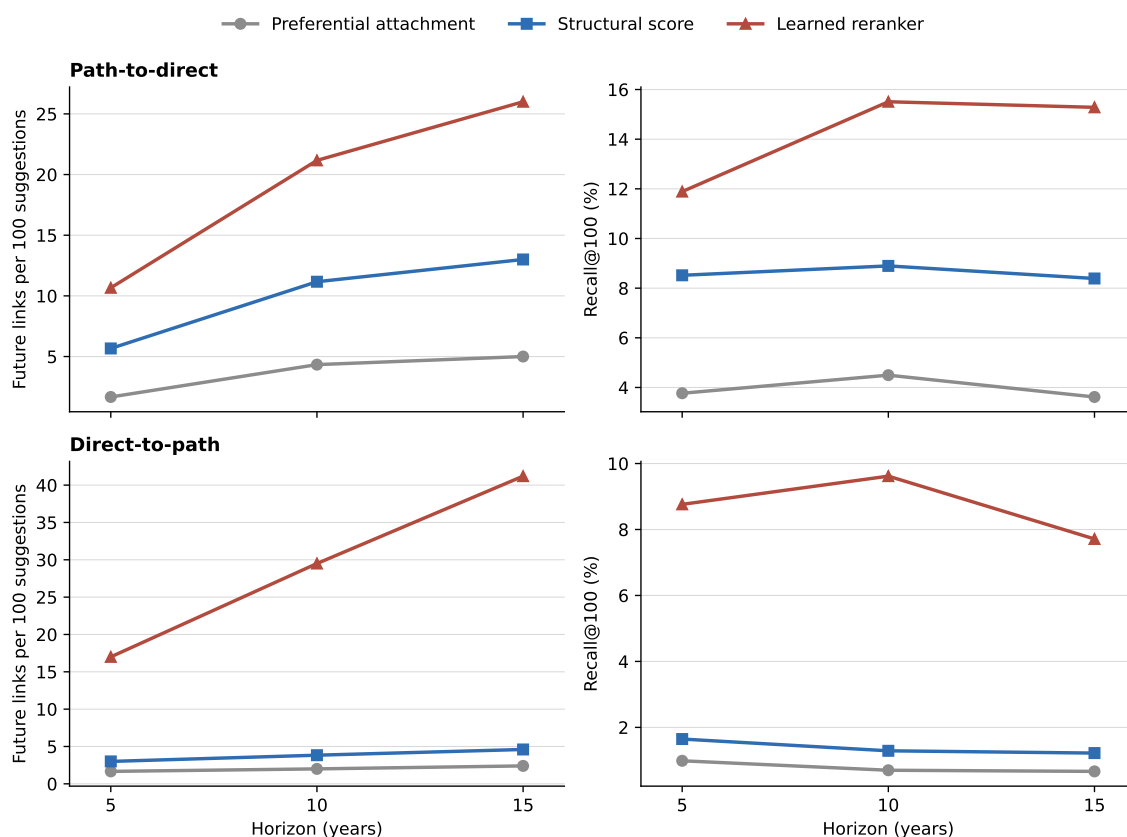
The first comparison asks whether local structure adds anything beyond popularity. That is a real test in this setting because research attention is path-dependent. Visible topics, familiar data, and established

¹²I do not know of a clean economics paper that uses this exact co-occurrence rule as a prospective benchmark for open-question screening, so I do not force an economics citation here. In economics and adjacent bibliometric work, co-mention, co-citation, and keyword co-occurrence are more often used descriptively to map fields than prospectively to rank next questions. The role of the benchmark here is narrower: it asks whether repeated joint mention alone is already enough, or whether directed claim structure adds screening value on top of that.

methods tend to attract still more work. If that cumulative-advantage process does most of the work, a popularity rule should already rank future questions well. It does not. In both historical families, the transparent graph score beats preferential attachment, and the learned reranker improves further.

The two families then diverge in form. Direct-to-path yields denser shortlists. In plain language, a larger share of the top 100 suggestions later becomes part of the published record. At $h = 5$, the learned reranker surfaces about 17.0 future links per 100 suggestions in direct-to-path, versus 10.7 in path-to-direct, which is equivalent to a 17.0% versus 10.7% shortlist hit rate. At $h = 10$, the comparison is 29.5 versus 21.2. At $h = 15$, it is 41.2 versus 26.0. Path-to-direct instead captures more of the later-realized stock and ranks it earlier on average. The learned reranker reaches Recall@100 of 11.9, 15.5, and 15.3 across $h = 5, 10, 15$, versus 8.8, 9.6, and 7.7 in direct-to-path (Figure 8 and Table 2).

Figure 8: The two historical families are informative in different ways



Notes. Each row is one historical family and each column is one shortlist metric. Path-to-direct asks whether a locally supported direct relation later appears. Direct-to-path asks whether an existing direct relation later acquires a supporting path. Within each family, the transparent score beats preferential attachment and the learned reranker improves further. Across families, the comparison is not one-dimensional: direct-to-path delivers more later realizations per 100 suggestions, while path-to-direct has higher Recall@100 and MRR.

The top-100 shortlist recovers 10–15% of the eventual realization stock from candidate pools that contain thousands of pairs. The graph helps, but it helps in different ways. In direct-to-path, the main gain is a

denser flow of later realizations through the shortlist: Precision@100 rises. In path-to-direct, the main gain is bringing a larger share of the later-realized stock nearer the top of the ranking: Recall@100 and MRR rise.

The two families capture different forms of research progress and should not be collapsed into one generic score. Direct-to-path is most useful when an editor, researcher, or funder wants a sharp shortlist of claims that look paper-ready but still underworked. Path-to-direct is most useful when the task is to see which accepted claims are still structurally thin—where mechanism work is still missing. Much empirical economics moves from a reduced-form relation to channels, heterogeneity, and scope conditions, so direct-to-path is the more immediately readable family for most economists. But path-to-direct still matters: it closes missing direct claims that nearby support had already made plausible.

5.2 Mechanism thickening versus direct closure

Which motion is more common? One is direct closure: a supporting path already exists at $t - 1$, the direct link does not, and that direct link later appears. The other is mechanism thickening: a direct link already exists at $t - 1$, but a supporting mediator path appears only later. The distinction separates two uses of research effort. Direct closure states a missing claim. Mechanism thickening explains, disciplines, or extends a claim the literature already treats as meaningful. Work on conservative search and declining disruption suggests that later work often develops accepted lines rather than replaces them with wholly new ones (Foster et al., 2015; Park et al., 2023). I compare the two motions on the same dated length-2 graph structure (Figure 4).

The aggregate comparison is clear (Figure 9). Direct-to-path transitions dominate path-to-direct transitions in every cutoff-period block and at every horizon. The cleanest way to state this is in transition rates, because the eligible sets differ. At $h = 10$, the direct-to-path transition share rises from about 0.9% in the 1990s to 4.3% in the 2000s and 12.1% in the 2010s—a roughly 12-fold increase over three decades—while the corresponding path-to-direct shares are about 0.3%, 0.7%, and 1.4%.¹³

The gap is large. Economics more often elaborates mechanisms around claims it already states than closes a missing direct relation implied by nearby support. That is consistent with a broader view of scientific change in which the literature more often thickens around existing structure than opens wholly new direct connections (Foster et al., 2015; Park et al., 2023). The graph more often surfaces structured elaborations of accepted claims than missing direct links between otherwise distant concepts. That helps explain why direct-to-path is the more natural reader-facing object in economics: it lines up with a literature that often asks not only whether a relation exists, but through what channel, under what conditions, and with what supporting mechanism.

¹³These numbers use a length-2 definition of “supporting path.” Appendix F reports the same rates under a length-3 definition (any three-hop chain $u \rightarrow w_1 \rightarrow w_2 \rightarrow v$). The direct-to-path share roughly doubles at every horizon under L3 (e.g., at $h = 10$: 1.9% \rightarrow 10.3% \rightarrow 30.2% across decades), while the path-to-direct share is diluted by the larger eligible set; the direct-to-path over path-to-direct ratio therefore strengthens, not weakens, when the path definition is broadened (Table 8). The 12-fold figure in the text is the conservative length-2 version.

The pattern varies across the literature in economically interpretable ways. Adjacent journals are relatively more path-closure heavy than the core, but direct-to-path still dominates in both tiers. At $h = 5$, the share of realized transitions taking the path-to-direct form is about 0.45 in adjacent journals versus 0.28 in the core. Finance remains more direct-to-path heavy throughout. Figure 10 reports the journal-tier split. The asymmetry is therefore not an artefact of which journals are included in the core.

The paired evidence carries two lessons. In both historical families, graph structure improves on popularity and the reranker improves further. The larger substantive point is that much of scientific development takes the form of mechanism-deepening around existing claims. The most useful next question is therefore often richer than a single missing edge.

5.3 Reranked results in both families

The learned reranker improves on the transparent score in both families. Keeping the ontology, sample, and timing discipline fixed, it changes only how graph features are weighted. The best configuration differs by family and horizon—the appendix reports the full tuning table—but the direction is consistent: learned weighting moves more eventual realizations toward the top of the shortlist. The improvement comes from combination effects that the transparent fixed-weight score misses, not from additional data.

The transparent score remains useful alongside the reranker because it explains *why* a candidate looks promising in graph terms: path support, missingness relative to nearby support, topology, and provenance. Table 2 summarizes the paired comparison.

Table 2: Paired benchmark summary for the two historical families

Family	Model	$h=5$	$h=10$	$h=15$
Path-to-direct	Preferential attachment	1.7 / 3.8 / 0.0021	4.3 / 4.5 / 0.0020	5.0 / 3.6 / 0.0019
	Structural score	5.7 / 8.5 / 0.0078	11.2 / 8.9 / 0.0078	13.0 / 8.4 / 0.0073
	Learned reranker	10.7 / 11.9 / 0.0123	21.2 / 15.5 / 0.0104	26.0 / 15.3 / 0.0114
Direct-to-path	Preferential attachment	1.7 / 1.0 / 0.0012	2.0 / 0.7 / 0.0010	2.4 / 0.7 / 0.0011
	Structural score	3.0 / 1.6 / 0.0019	3.8 / 1.3 / 0.0015	4.6 / 1.2 / 0.0015
	Learned reranker	17.0 / 8.8 / 0.0056	29.5 / 9.6 / 0.0089	41.2 / 7.7 / 0.0061

Notes. Each row reports one family-model combination averaged across the 1990–2015 cutoff grid. Columns are grouped by metric, with the three main horizons nested underneath. Future links per 100 suggestions is the shortlist hit rate, that is, Precision@100 expressed in percent. Recall@100 is the share of all later realizations in that family captured in the top 100. Mean reciprocal rank rewards putting later realizations nearer the top. The table shows the same pattern as Figure 8: the graph beats preferential attachment in both families, the learned reranker improves further, and the two families remain substantively different rather than collapsing to one ranking problem.

The paired shortlist comparison already shows that graph structure adds screening value beyond popularity in both historical families. The reranker shows that the same candidate universe can be ordered better still. Additional paired extensions, including the current frontier and broader reading budgets, are therefore follow-on checks rather than the main claim (Appendix B and Appendix B.3). The paired budget comparison adds one practical lesson: path-to-direct is sharper when the reader wants a very short

shortlist, while direct-to-path becomes more useful once the reader is willing to inspect a broader set of candidates (Appendix B.3).

What the reranker loads on. A grouped Shapley decomposition (Appendix H) separates two things the reranker does. It loads *positively* on directed causal degree, recency, and evidence quality: pairs whose endpoints are central in the causal subgraph, have been the subject of recent research, and draw on well-evidenced concepts. It loads *negatively* on broad support-graph popularity: once directed causal degree and recency are controlled for, being a well-connected concept in the wider support graph reduces the predicted probability of realization. In plain terms, the model identifies concept pairs whose recent and causally central activity outpaces their long-run popularity. That is the opposite of cumulative advantage, and the decomposition is stable across logistic and tree-based model families (Figure 26).

6 Discussion and Extensions

These results speak to a structural feature of the economics literature, not merely to the performance of a ranking algorithm. At each historical cutoff, the graph stores the directional accumulation of what the literature has already asserted and what it has conspicuously left unsaid. When that structure predicts later publication at rates far above the popularity baseline, it means past claims carry systematic information about where the next ones will go. The asymmetry between the two families is equally informative. The literature more often extends accepted claims through new mechanisms than it closes locally implied direct relations—a pattern that is stable across eras and horizons. A screening tool calibrated on such a literature will surface a larger share of mechanism-thickening candidates: not because the algorithm favors them, but because that is the shape of research progress in economics.

What the distinction means. The two historical objects capture different parts of research progress. Direct closure is the case where nearby support already points toward a relation that later becomes explicit. Mechanism thickening is the case where a known relation later acquires a clearer channel. Economists care about both, but they help in different ways. Direct-closure questions are useful when a researcher, editor, or funder wants a sharper shortlist of claims that look paper-ready but still underworked. Mechanism-thickening questions are useful when the task is to see where an accepted result is still structurally thin. The historical record says the second move is more common. A useful screening tool should therefore not only point toward underworked direct relations. It should also help identify where an accepted claim still lacks a convincing mechanism.

Why to trust the pattern. One reason to take that result seriously is that it does not rest on one fragile ranking rule. The main paired comparison survives the move from a transparent score to a learned reranker, and the reranker carries forward to later vintages that were not used for model selection (Figure 11). The appendices show the same discipline more broadly: the extraction layer is auditable, the

directed claim graph is not just undirected co-mention in disguise, and the natural extensions confirm rather than reverse the main pattern.¹⁴

Where the graph helps. Two broader implications follow. First, the graph does not add the most value where the literature is maximally thin. The appendix regime splits show that its comparative advantage is larger where some usable local structure already exists. So this is not a machine for discovering the most remote idea in the graph. It is a screening device for the large middle of the literature, where there is enough nearby evidence to make a question legible but still enough missing structure that attention can be reallocated productively. Second, if the literature usually moves by thickening mechanisms around existing claims, then one useful role for computational tools is not open-ended idea generation. It is to help economists see where an accepted relation is still structurally underexplained.

Benchmark event and surfaced question. Keeping the benchmark and the reader-facing question separate follows from that logic. A dated graph event is useful because it can be tested prospectively. A surfaced question is useful because it can be read, argued over, and discarded if it is uninteresting. The historical benchmark gives discipline. The surfaced object gives the tool practical value. Supplementary human and LLM usefulness checks remain secondary for that reason: they help assess whether the surfaced questions are readable, but they do not replace the historical evidence.¹⁵

Budget interpretation. The budget results sharpen that interpretation. Scarce attention is not only an individual researcher's problem. It is also an editorial, funding, and field-level problem. On that margin, the two historical families do not play the same role. Path-to-direct is more useful for triage: it concentrates more realized direct closures near the top of a short list. Direct-to-path is more useful for exploration: once the reading budget widens, it recovers a much broader share of the future mechanism literature. That is the relevant object when the problem is to see where an accepted relation is still underexplained rather than merely underclaimed.

These budget results also connect the paper to a broader literature on how science allocates attention (Jones, 2009; Bloom et al., 2020; Azoulay et al., 2011; Foster et al., 2015; Bhattacharya and Packalen, 2020). That literature emphasizes the rising burden of reaching the frontier and the institutional tilt toward safer, more familiar directions. The present paper does not solve that problem at the level of incentives. But it does speak to scientific method in a narrower way: under severe screening constraints, a useful method is one that moves more plausible questions into the stage where economists, editors, and funders can exercise judgment. The graph is not a replacement for judgment. It is a disciplined way to widen the agenda that reaches judgment.

¹⁴The held-out temporal test is reported in Appendix H.9. The author-positioning null result is discussed in Appendix H. Credibility and stability audits are collected in Appendix J.

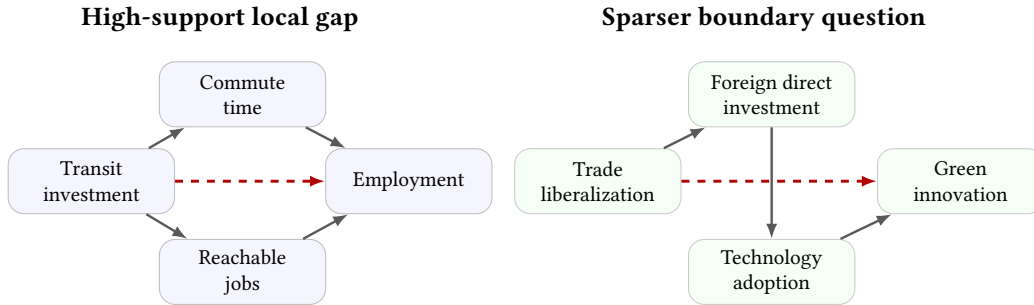
¹⁵Appendix K reports a small blinded human exercise and a larger appendix LLM screen. Both are useful as screening checks, not as substitutes for the historical benchmark.

Bundle uptake. Later papers could in principle realize a bundle of predicted edges rather than one. Work in science-of-science has emphasized that novel contributions often recombine nearby ingredients rather than arriving as isolated moves (Uzzi et al., 2013; Foster et al., 2015; Fortunato et al., 2018). The bundle-uptake analysis in Appendix L.1 suggests that the dominant downstream object in this setting is still a single paper-shaped question: only about five percent of realizing papers take up more than one historically predicted edge, and mixed-family bundles are very rare. But when multi-edge uptake does occur, it is strongly local in graph space. That result helps interpret the budget evidence. The wider-budget advantage of *direct-to-path* seems to come mainly from broader coverage across many separate papers, not from a large number of later papers each realizing multiple predicted edges across both families.

Extensions. Two extensions are most worth developing. The first is to test whether the mechanism-thickening asymmetry is specific to economics or a broader feature of cumulative science. Applying the same walk-forward benchmark to sociology, political science, and psychology would reveal whether directed claim graphs in other social sciences show the same imbalance between mechanism elaboration and direct closure. The second is dynamic concept activation. The current benchmark holds the active concept set fixed at each cutoff, but some of the highest-impact work introduces entirely new nodes. A model that predicts when a dormant concept pair becomes active—rather than when a missing edge closes—would complement the missing-link design and speak more directly to radical novelty.

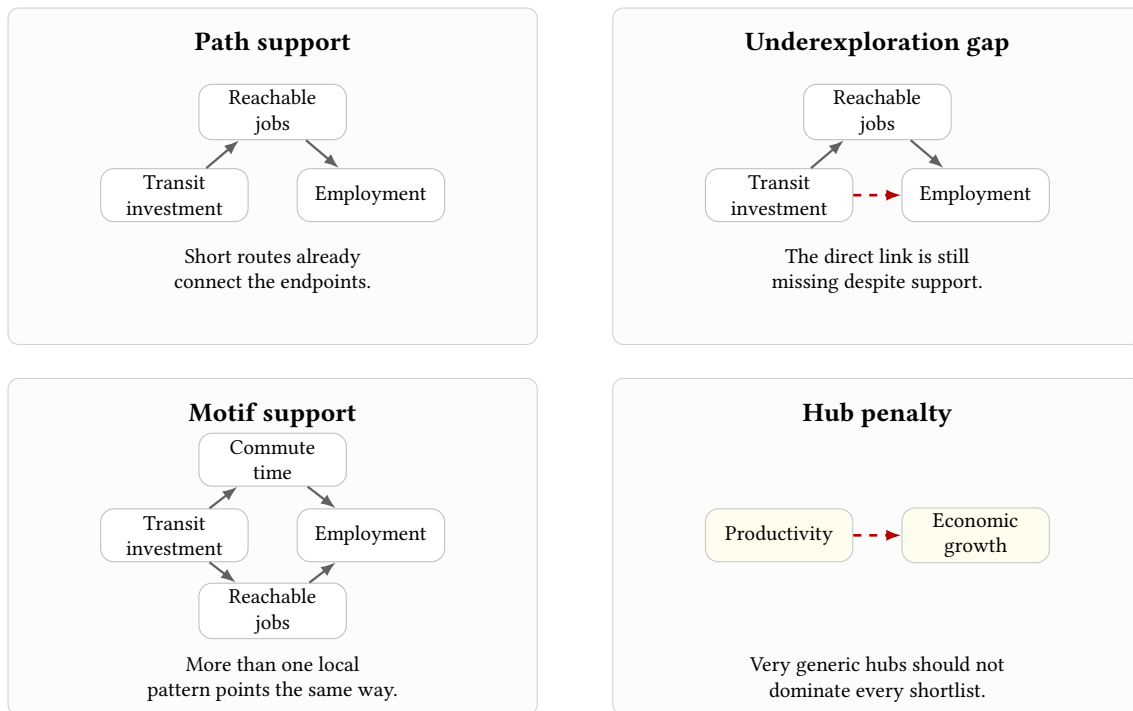
This paper provides a prospectively testable historical design, a readable surfaced object, and an honest comparison with the cumulative-advantage null. That is enough to establish that the nearby structure of the literature carries screening information that popularity alone misses. If ideas are harder to find and downstream research tasks are getting cheaper, the value of better upstream screening rises. Economics should not decide what to work on next only through popularity and familiarity. The graph offers a structured way to widen the set of questions that reaches the judgment stage—not to replace judgment, but to give it more to work with.

Figure 5: Gap and boundary questions are different local graph patterns



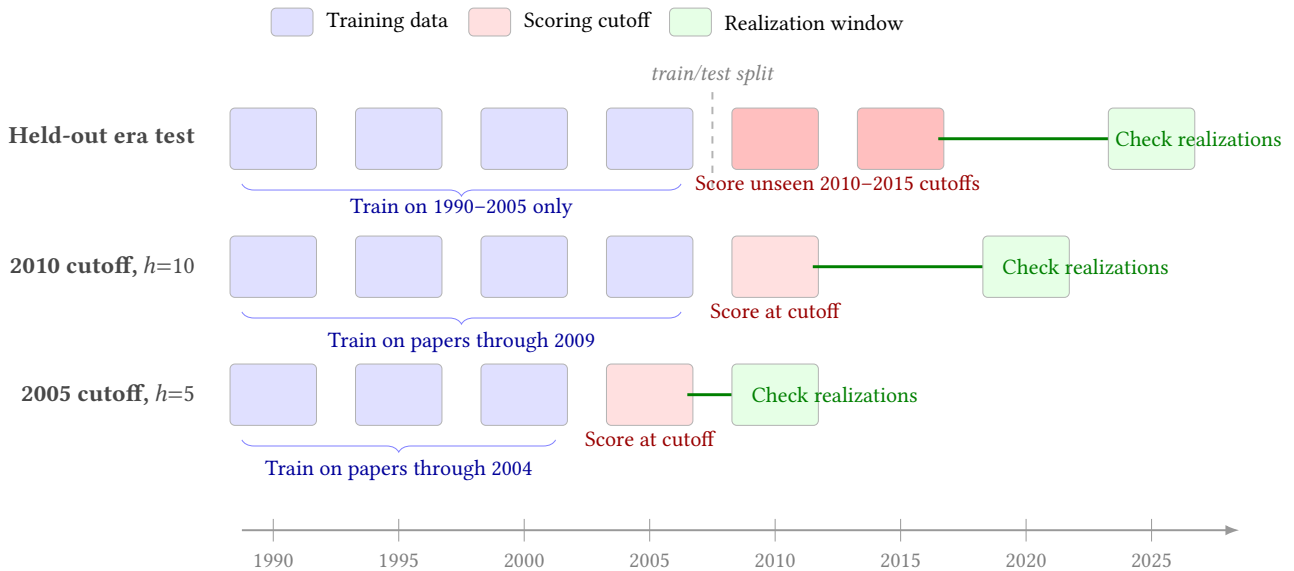
Notes. This figure exists to separate two screening cases that would otherwise look similar in prose. The left panel shows a gap-like candidate: nearby support is already dense, but the direct relation remains missing. The right panel shows a more boundary-like candidate: the two end concepts are connected only by a thinner mechanism chain. The node labels are hand-curated economics-facing examples used only to make the local graph pattern readable. The diagrams are conceptual rather than exhaustive.

Figure 6: The transparent score rewards supported but still undercompleted pairs



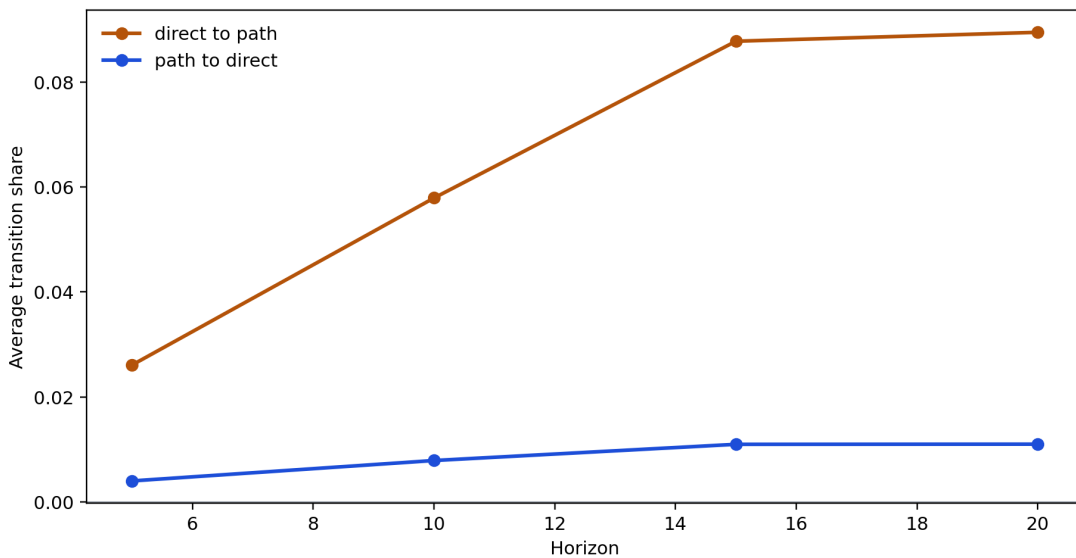
Notes. This figure makes the transparent score legible without the formula. The first three panels continue the same transit-investment worked example used in the earlier method figures. Path support asks whether short routes already connect the endpoints. The underexploration gap asks whether the direct relation is still missing despite that support. Motif support asks whether more than one nearby pattern points toward the same endpoint pair. The final panel intentionally switches to generic hubs to show the opposite force: the score should not rank a candidate highly only because both concepts are broad and heavily connected.

Figure 7: Walk-forward evaluation keeps scoring vintage-correct



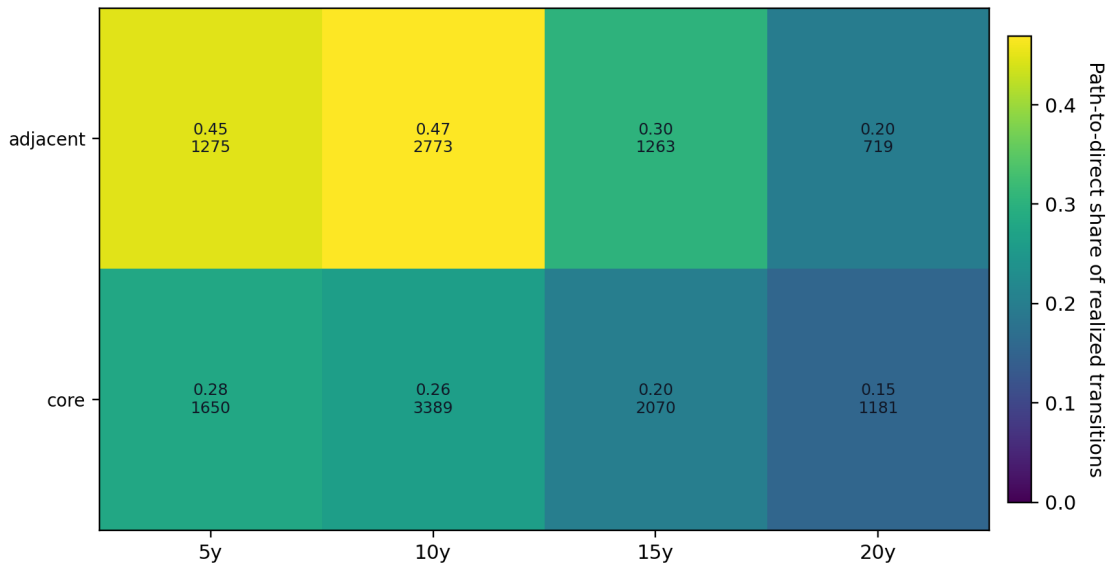
Notes. This figure exists to show the paper’s temporal discipline visually. At each cutoff, the graph uses only papers published before that date (blue). Candidates are scored at the cutoff (red), and realizations are checked over the horizon window (green). The bottom row shows the held-out era test from Appendix H.9: train on 1990–2005 and evaluate on the unseen 2010–2015 era.

Figure 9: Mechanism thickening is more common than direct closure



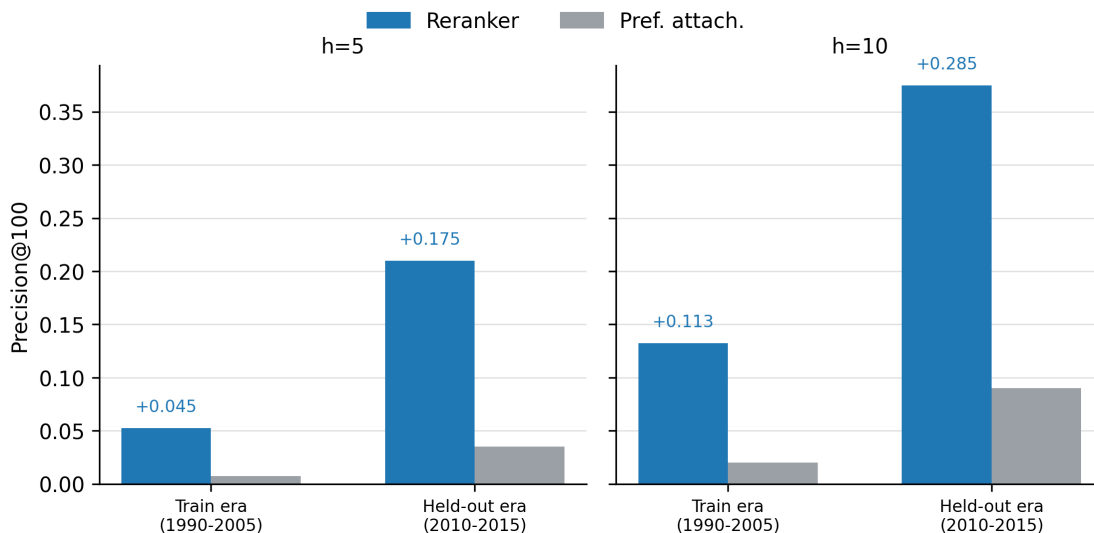
Notes. “Path to direct”: a local path exists first and the direct link appears later. “Direct to path”: the direct link exists first and a mediating path appears later. The figure exists to compare these two motions directly. Mechanism-deepening around existing claims dominates at all horizons.

Figure 10: Adjacent journals close more path-implied links, but direct-to-path still dominates



Notes. This figure asks whether the transition mix differs by journal tier. It reports the share of realized path-related transitions taking the path-to-direct form by journal tier and horizon. Adjacent journals are consistently more path-closure heavy than the core, but direct-to-path remains the larger transition type in both tiers.

Figure 11: The reranker generalizes forward in time



Notes. For each main horizon, the reranker is selected using only 1990–2005 cutoff cells and then evaluated on the fully held-out 2010–2015 era. Labels report the reranker’s absolute $P@100$ gap relative to preferential attachment. The held-out gap is larger than the earlier training-era gap at both $h = 5$ and $h = 10$. Full results in Appendix Table 19.

Appendix

What Should Economics Ask Next?

Table 3: Positioning relative to closest comparable work

	This paper	Impact4Cast	Sourati et al.	Tong et al.
Domain	Economics	All sciences	Biomedicine	Psychology
Edge type	Directed causal	Co-occurrence	Co-occurrence	Causal
Corpus	242,595 papers	2.4M papers	Varies	43,312 papers
Main null	Preferential attachment	ML baselines	Content-only	TransE
Temporal eval	Prospective walk-forward	Holdout	Holdout	None
Human eval	Appendix checks	No	No	Yes

Notes. Impact4Cast: Gu and Krenn (2025). Sourati et al.: Sourati et al. (2023). Tong et al.: Tong et al. (2024). The table compares the closest systems along the dimensions most relevant for this paper: domain, edge type, scale, main null, temporal design, and whether any human check is reported.

A Guide to the Appendix

The appendix has five jobs. First, it documents how the graph is built from title and abstract text. Second, it records the benchmark tables and reranker detail that support the main text without carrying its argument. Third, it collects paired and family-specific extensions that are useful once the main comparison is clear. Fourth, it reports audit and presentation checks on the graph and on the surfaced questions. Fifth, it adds a short set of supplementary analyses on how later papers realize predicted ideas and which kinds of teams move toward them. The main text remains narrow: paired shortlist results first, then the path-versus-mechanism result, then the confirmatory reranker comparison.

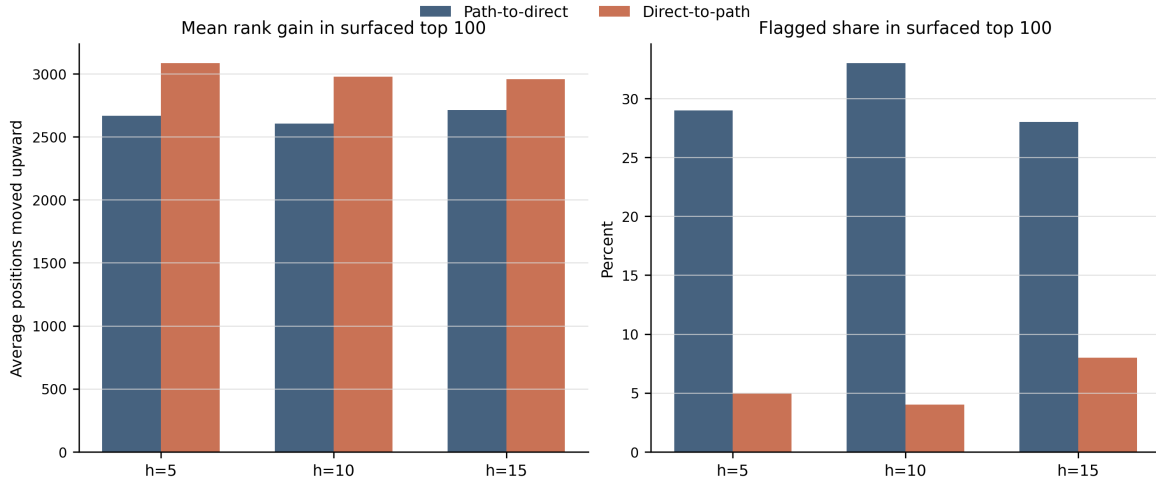
B Paired family extensions

This section extends the paired comparison beyond the strict historical shortlist. The two figures below keep path-to-direct and direct-to-path side by side on two additional objects that are already available for both families: the current frontier and the broader-shortlist heterogeneity splits. Family-specific extensions that still exist only for direct closure remain in the next section.

B.1 The two families surface different current frontiers

The historical benchmark and the current frontier answer different questions. The historical benchmark asks whether later papers validate a proposed relation or mechanism. The current frontier asks what each family would surface now after the reranker has reordered the large candidate pool. The paired current-frontier summary therefore compares the two families on the same simple diagnostics: how far the reranker moves surfaced items upward in the pool, and how often the surfaced top 100 still contains artifacts or generic endpoints.

Figure 12: The current frontier differs across the two question families



Notes. Each bar summarizes the surfaced top 100 for one family and one horizon on the current frontier pool. The left panel reports the mean upward movement in rank after reranking relative to the raw transparent ordering. Larger values mean the reranker is finding good candidates much deeper in the pool. The right panel reports the share of surfaced top-100 items that trigger the paper’s artifact or generic-endpoint flags. The comparison is descriptive rather than historical. Its purpose is to show that the two families are not cosmetic variants of the same current shortlist. Path-to-direct keeps a larger flagged share in the surfaced top 100, while direct-to-path looks cleaner on that margin and still receives large upward rank gains from the reranker.

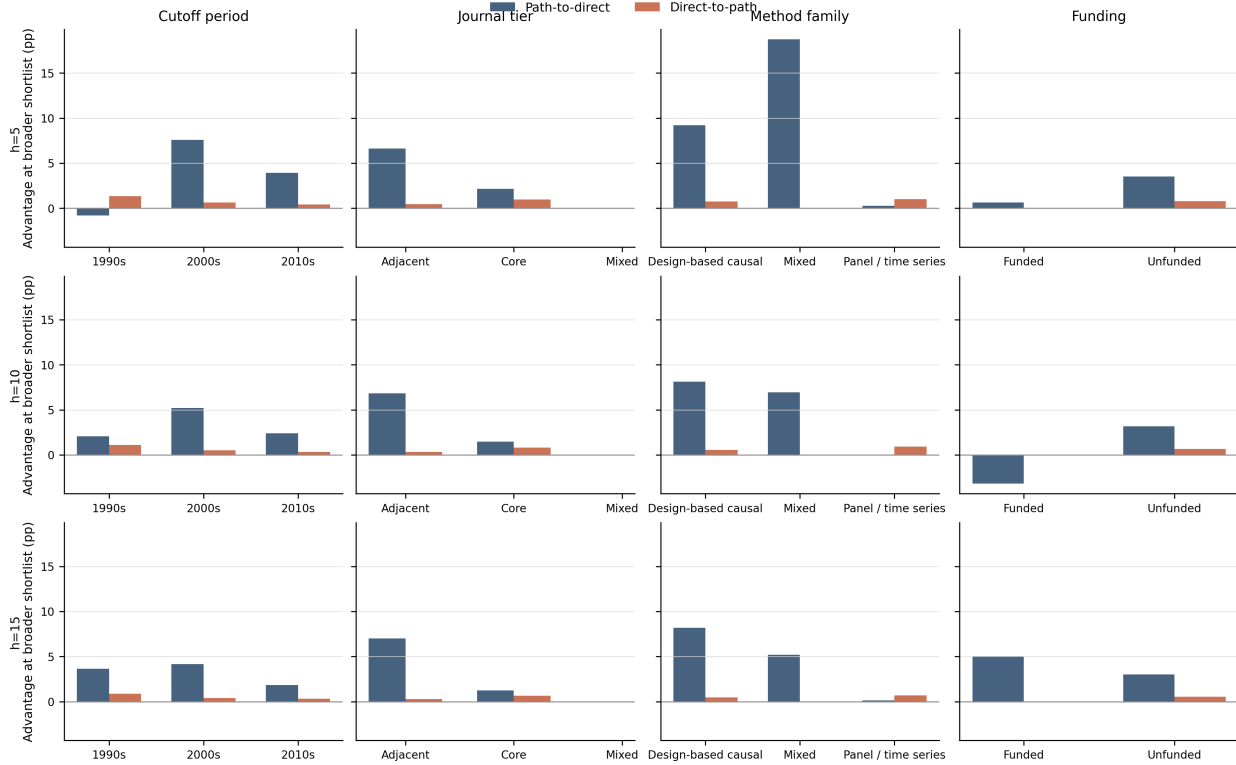
The two families differ in a way that matters for interpretation. The path-to-direct frontier is more tightly tied to strict historical validation, but it also surfaces more generic or noisy endpoints. The direct-to-path frontier looks cleaner on the current-reader margin and still benefits strongly from reranking. That is one reason to keep both families in view rather than treating one as a mere robustness check.

B.2 Heterogeneity is also family-specific

The main text reports the paired strict-shortlist comparison. The figure below asks a different question: where does each family gain more once the comparison is expanded beyond the strict top 100? The vertical axis reports the graph score’s average recall advantage over preferential attachment on broader shortlist-share cutoffs. Positive values favor the graph score.

This paired heterogeneity display helps discipline the interpretation. If the two families had looked identical under every split, the paper could have treated one as a convenient historical proxy for the other. They do not. The families appear to reward different kinds of local graph structure. That is why the paper compares them directly before giving either one rhetorical priority.

Figure 13: Where the graph helps differs across the two families



Notes. Rows correspond to horizons $h = 5, 10, 15$. Columns split the benchmark by cutoff period, journal tier, method family, and coarse funding status. Each bar reports the graph score’s average recall advantage over preferential attachment on broader shortlist-share cutoffs, expressed in percentage points. Positive bars favor the graph score. These are descriptive subgroup summaries, not regression coefficients. The figure shows that path-to-direct and direct-to-path do not vary across subgroups in the same way. Path-to-direct has larger gains in adjacent journals and in several design-based slices. Direct-to-path remains positive more quietly across many cells, with smaller but steadier advantages.

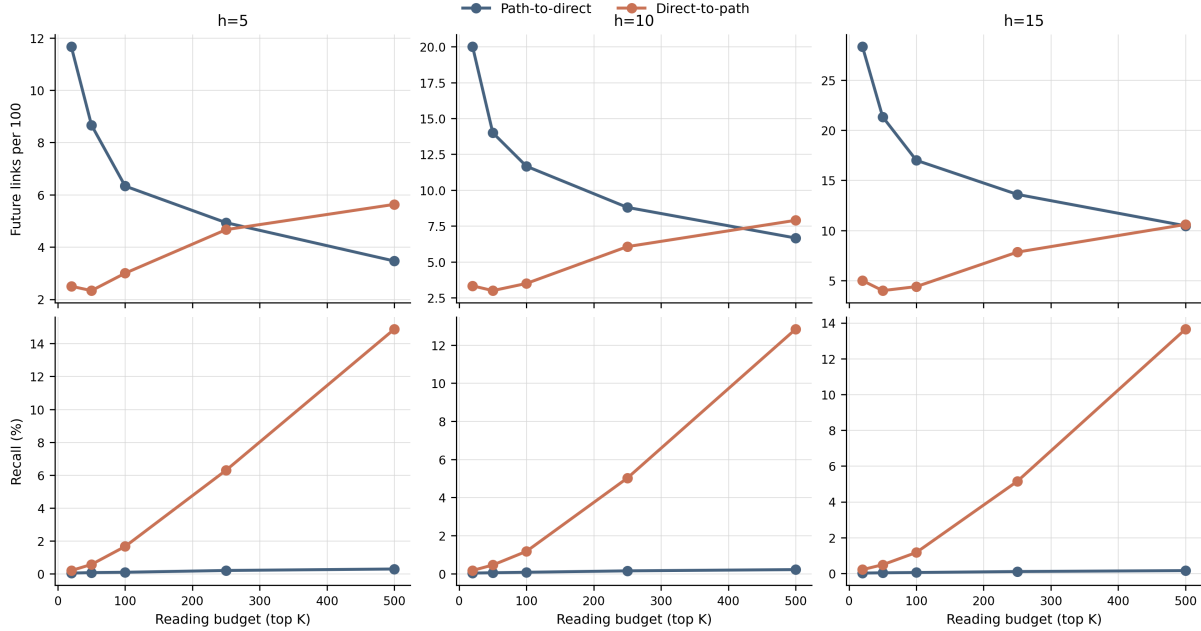
B.3 The two families behave differently as the reading list expands

The strict top-100 comparison in the main text is useful because it keeps the historical object common across the two families. But economists usually read a shortlist rather than a single question. This appendix therefore asks what happens as the reading budget widens from the first 20 questions toward 500.

Figure 14 fixes the candidate pool at the largest common value used in both families and varies only the number of surfaced questions K . The top row reports realized future links per 100 surfaced questions. The bottom row reports recall. This is the relevant trade-off. A family can surface a sharper first tranche, or it can recover a broader share of eventual future links once the reading list becomes longer.

The figure shows a stable crossover pattern. At tight budgets, path-to-direct delivers more realized links per 100 suggestions. At horizon 15, for example, the first 100 surfaced path-to-direct questions yield about 17 realized future links per 100 suggestions, compared with about 4.4 for direct-to-path. But direct-to-path catches up as the reading list broadens. By the top 500, the same comparison is about 10.6 versus 10.5, and

Figure 14: The two families solve different reading-budget problems



Notes. Each panel fixes the candidate pool at the common pool size of 5,000 and varies the reading budget K . The top row reports future realized links per 100 surfaced suggestions. The bottom row reports recall. Path-to-direct is sharper in the very first reading tranche. Direct-to-path recovers a much larger share of future links once the shortlist broadens.

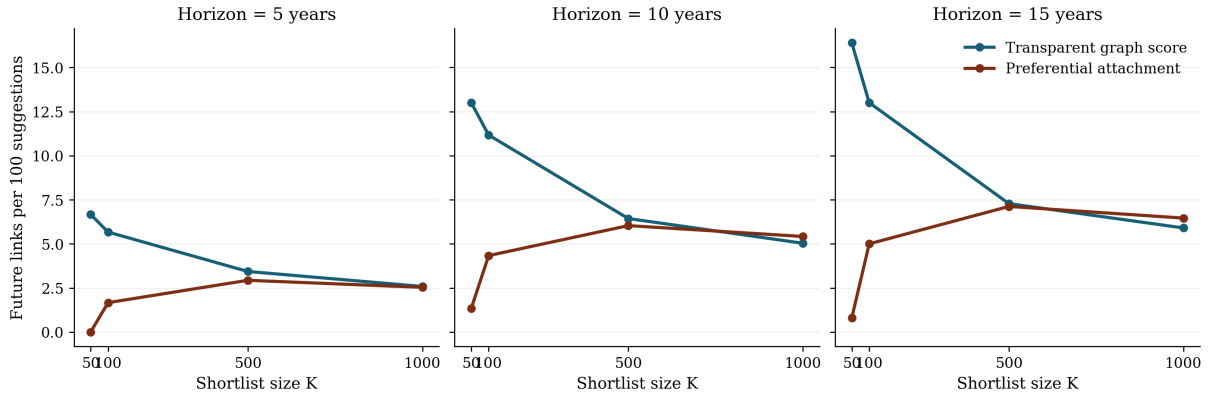
recall is much higher for direct-to-path at every horizon.

The interpretation is substantive. Path-to-direct behaves like a stricter screening rule: it concentrates successful direct closures near the top of the ranking. Direct-to-path behaves like a broader discovery rule: once a reader is willing to inspect a wider set of candidates, the mechanism-thickening family retrieves a much larger share of the eventual future literature. The two families therefore do not address the same reading-budget problem, even when they are evaluated on the same historical horizon structure.

C Direct-closure screening extensions

The main text keeps only the result objects that are already paired across path-to-direct and direct-to-path. This appendix collects the richer direct-closure extensions that have not yet been rebuilt for the direct-to-path family: the reading-budget frontier, the value-weighted comparison, the broader-shortlist heterogeneity displays, and the current surfaced examples drawn from the direct-closure shortlist.

Figure 15: As the reading list expands, the graph edge narrows



Notes. Each panel reports future links per 100 surfaced suggestions as the shortlist expands from $K = 50$ to $K = 1000$ on the main 1990–2015 direct-closure benchmark. The figure is meant to answer one question only: where does graph structure help relative to popularity when attention is scarce? The transparent score leads at $K = 50$ and $K = 100$ for all three main horizons. By $K = 500$, the gap is small. By $K = 1000$, the two rules are nearly tied at $h = 5$ and preferential attachment is slightly ahead at $h = 10$ and $h = 15$.

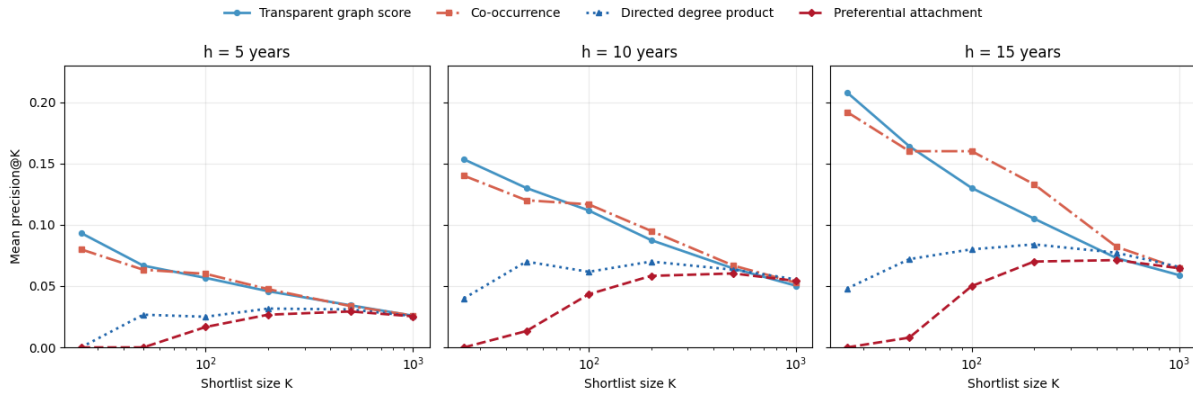
C.1 The gains are largest when the reading list is short

The first extension asks what happens when the reading budget expands beyond the first 100 questions. Economists rarely inspect one suggestion and stop; they inspect a shortlist. The attention-allocation frontier therefore asks how the comparison changes as that shortlist expands from $K = 50$ to $K = 1000$. I summarize that margin using “future links per 100 suggested questions.” Out of 100 suggestions, how many later appear? The object in this subsection is still the transparent direct-closure score, because its mechanics can be read directly question by question.

Figure 15 shows the answer. On the 1990–2015 benchmark, the transparent score already leads preferential attachment at $K = 50$ and $K = 100$ at all three main horizons. At $h = 10$, for example, the transparent score yields about 13.0 future links per 100 suggestions at $K = 50$ and 11.2 at $K = 100$, versus 1.3 and 4.3 for preferential attachment. The key frontier lesson is therefore not that popularity wins the tight shortlist and then fades. It is that the transparent score’s advantage is concentrated in tighter shortlists and narrows sharply once the reading list expands toward $K = 500$ or $K = 1000$.

Figure 16 makes the direct-closure result more precise. Among the simple scores, there is no single winner across shortlist sizes. The transparent graph score is strongest at the first reading tranche, which is exactly where a researcher decides which 25, 50, or 100 questions deserve inspection first. Co-occurrence becomes slightly stronger at intermediate shortlist sizes, and by very broad lists the simple scores largely converge. The right reading is therefore narrow: the transparent score is an interpretable strict-shortlist screen, not a universal winner across every reading budget.

Figure 16: Across simple scores, the winner depends on shortlist size



Notes. Each panel reports mean precision@K across the main 1990–2015 horizon-valid cutoff grid, with log-scale K, for the direct-closure family. The figure’s point is not that one simple score wins everywhere. It is that different transparent scores are useful at different shortlist sizes. The transparent graph score is strongest at the very tight shortlist margin ($K = 25$ and $K = 50$). Co-occurrence is slightly strongest at intermediate shortlist sizes. By $K = 1000$, the simple-score family largely converges.

C.2 The pattern survives when later reuse is weighted

The next extension asks whether the graph signal survives once future links are weighted by downstream reuse rather than counted equally. Future appearance is not the only margin that matters. A later realized link can also be weighted by later 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 this way rather than treated as binary appearance alone.

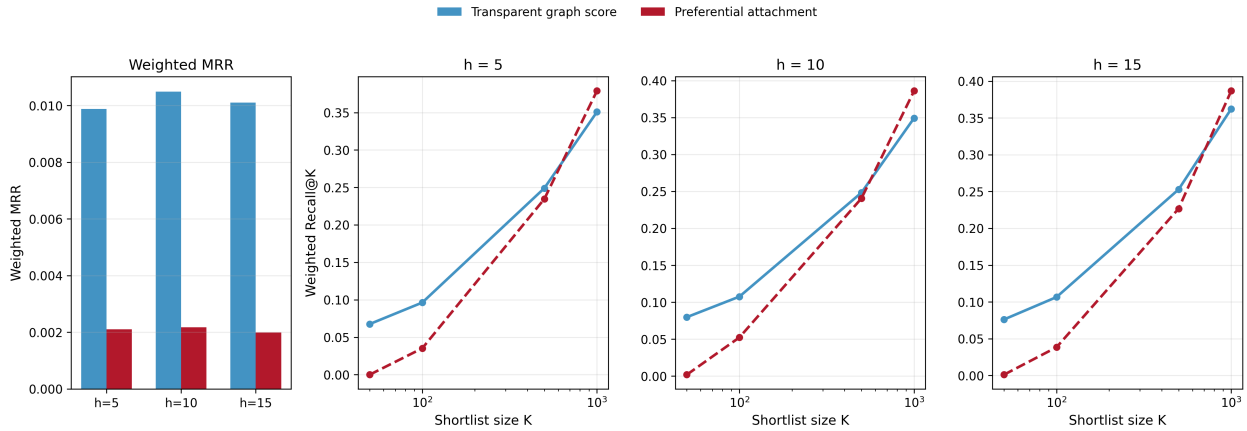
The weighted result is disciplined and favorable to the graph score. Once future links are weighted by later reuse, the transparent graph score beats preferential attachment on weighted MRR at all three main horizons. At $h = 10$, for example, weighted MRR is about 0.0105 for the transparent score versus 0.0022 for preferential attachment. The weighted recall frontier still bends back toward popularity as K becomes very broad, but the strict-shortlist weighted margin is clearly on the graph side (Figure 17).

Weighting by downstream reuse does not erase the graph signal. But it does not make popularity irrelevant either. At very broad reading lists, central concepts still reclaim some of the weighted frontier. So the value-weighted evidence reinforces the same budget logic as the binary frontier: local graph structure helps most when the reading problem is selective, while popularity becomes a stronger guide once the shortlist becomes very broad.

C.3 Where structure helps more

The next extension asks where the graph should help more if it is carrying real screening information rather than noise. The strict top-100 comparison hides meaningful variation across reading budgets. Once the same candidates are evaluated over broader fixed-K and pool-share shortlists, the transparent score

Figure 17: Value weighting still favors the graph at the strict shortlist



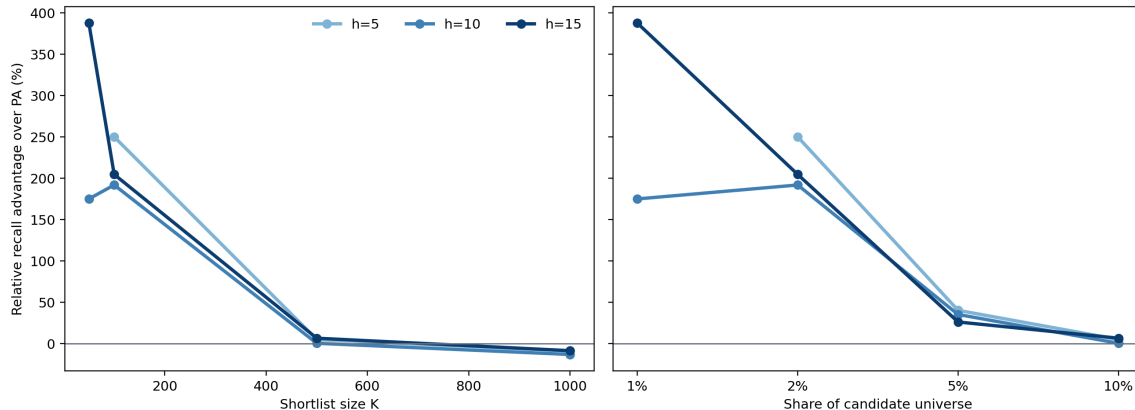
Notes. The left panel reports weighted MRR by horizon, where future realized links are weighted by later reuse. The right panels report weighted recall frontiers over shortlist size K . The figure is shown here as a direct-closure extension. Its message is straightforward: once later reuse is taken seriously, the strict-shortlist comparison still favors the transparent score. The graph edge narrows as the shortlist becomes very broad, but it does not disappear at the budget margin that matters most for screening.

looks materially stronger than the strict top-100 comparison suggests (Figure 18). The right interpretation is not that the graph fails outside a winner-take-all contest. It is that its comparative advantage is largest when the reading problem is still selective and then flattens as the shortlist becomes very broad.

The subgroup results sharpen this interpretation because they match a plausible ex ante story rather than a random set of subgroup differences (Figure 19). If graph structure is most useful where the literature already contains legible local support, then adjacent journals and design-based causal work should be more favorable terrain than the most popularity-dominated or methodologically diffuse slices. That is what the panel shows. Adjacent journals are clearly more favorable than the core on the broader-shortlist comparison, with the graph score's average recall advantage over preferential attachment around +0.066 to +0.070 in adjacent journals versus roughly +0.013 to +0.022 in the core. Design-based causal slices remain strongly positive at all main horizons, while panel- and time-series-heavy slices are close to zero by $h = 10$ and $h = 15$ rather than strongly negative.

Funding adds nuance rather than one clean sign pattern. In the coarse funded-versus-unfunded split, unfunded work is consistently positive at all three main horizons, while funded work is close to zero at $h = 5$, negative at $h = 10$, and positive again at $h = 15$. That is not a stable enough pattern to make funding central. It is still useful in the appendix because the interaction plot shows that the relevant contrast is not simply funded versus unfunded. The strongest positive cells are unfunded-adjacent slices, while funded-core slices are the most popularity-dominated.

Figure 18: Broader shortlists are more favorable to the graph than the top 100 alone



Notes. This figure keeps the comparison simple: at each cutoff year and horizon, it compares the direct-closure graph score with preferential attachment on the same candidate pool. The left panel uses fixed shortlist sizes K . The right panel uses broader shortlist-share cutoffs equal to 0.01, 0.05, 0.1, and 0.5 percent of the 5,000-candidate retrieval pool. The vertical axis reports the graph score’s percent recall advantage over preferential attachment, averaged across the benchmark cells. Positive values favor the graph score. The main point is that the graph score looks better once the shortlist expands beyond the strict top-100 slice, but that advantage still fades as the shortlist becomes very broad.

C.4 Path-to-direct surfaced examples

The benchmark is still recorded on direct links that later appear or do not appear. But the question a researcher actually reads is usually a path or mechanism question built around that direct link. The examples below show that object directly for the path-to-direct family.

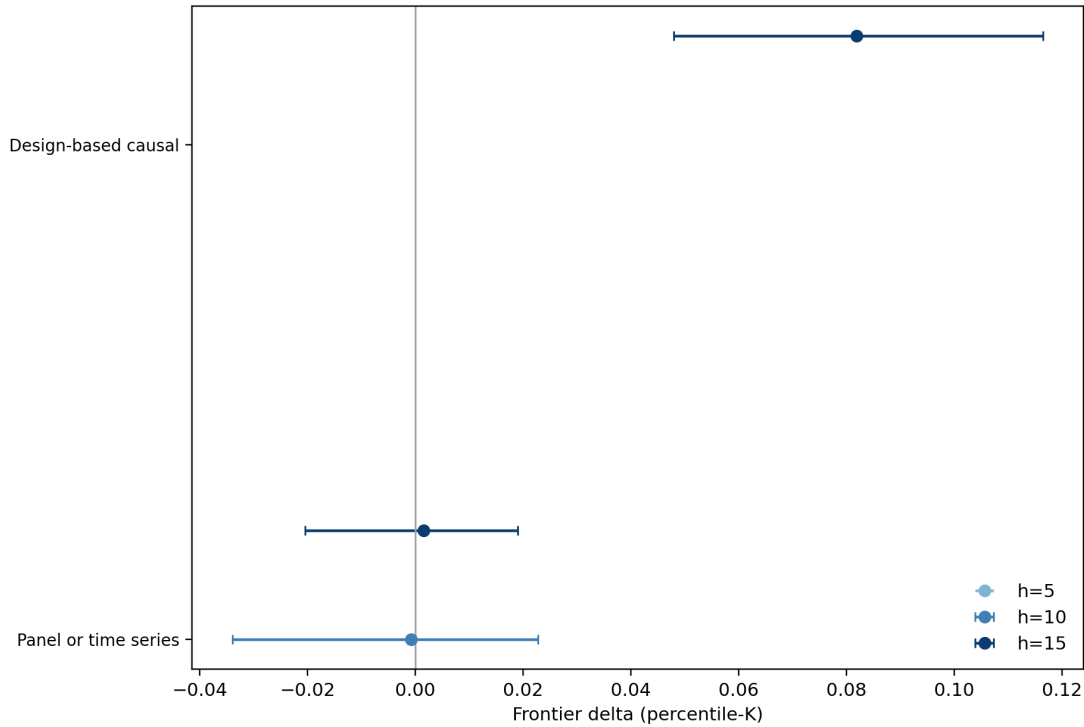
These examples make the same point in plain terms. The benchmark event is a missing direct link. But the most readable suggestions are rarely just “connect these two nodes.” They are usually mechanism questions around a relation the literature already partly supports. That is not an embarrassment. It means the system is most useful when it tightens, clarifies, or identifies a plausible relation rather than when it proposes a completely unstructured leap.

D Paper-local Extraction

This appendix documents how title-and-abstract text is turned into the paper-local graphs used in the benchmark. The core object is simple: one paper becomes one local graph of concepts and relations. Garg and Fetzer (2025) show that this is feasible for economics papers. The present paper uses the same starting point for a different downstream purpose: a reusable graph that can support missing-link construction, candidate ranking, and later concept matching across papers.

Three design choices matter most. First, the schema includes undirected relations, because contextual support is 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

Figure 19: Graph structure helps more in adjacent journals and design-based slices



Notes. Each point is an average over cutoff-year benchmark cells within one journal-tier or method slice. The horizontal axis reports the graph score’s average percent recall advantage over preferential attachment across the broader shortlist-share cutoffs used in Figure 18. This is not a regression coefficient. It is a descriptive summary of how much more recall the graph score delivers in each slice. Adjacent journals and design-based causal work are materially more favorable terrain for the graph score than core-journal and panel/time-series-heavy slices.

causally while using a weak design, or it can use a stronger design while describing the finding cautiously. Third, local scope qualifiers are stored separately from the node label whenever possible, so that later normalization can target concept identity rather than local sample wording.

D.1 Extraction prompts

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

System prompt.

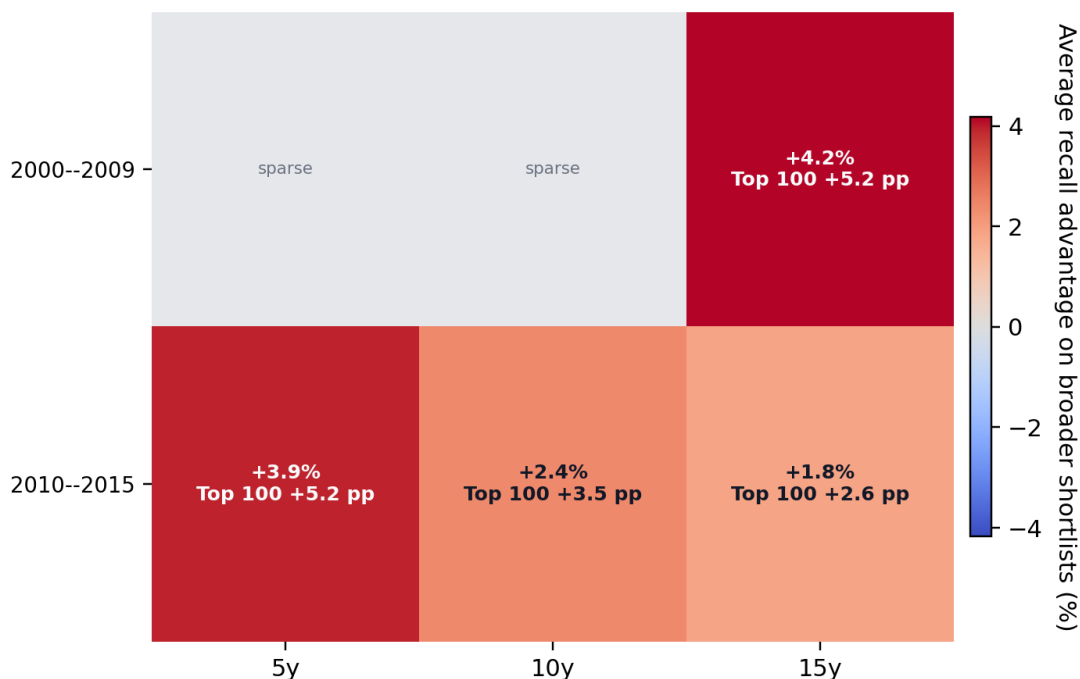
```

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.
    
```

Figure 20: The graph’s broader-shortlist advantage is strongest in the 2000s



Notes. Each cell averages over the benchmark cutoff-year cells that fall in that time period. This is a descriptive comparison, not a regression. Cell color shows the graph score’s average percent recall advantage over preferential attachment when the shortlist is allowed to expand to the broader shortlist-share cutoffs used in Figure 18. The number inside each cell reports the strict top-100 recall gain, in percentage points. Positive values favor the graph score. Gray cells are too sparse for interpretation under the paper’s display rule. Among the displayed cells, the broad-shortlist advantage is strongest in the 2000–2009 row, while the 2010–2015 row remains positive but smaller.

```

- Do not use outside knowledge.
- Do not infer relationships that are not supported by the title or abstract.

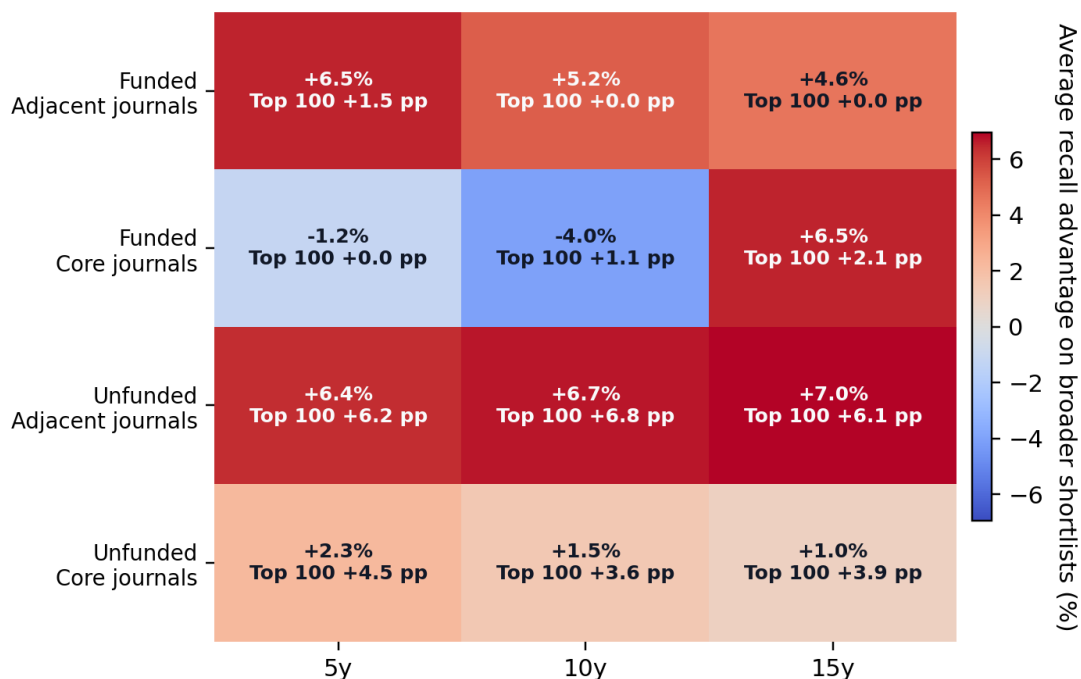
Purpose:
- The output will later be turned into a larger deterministic research graph.
- Downstream systems depend on consistent paper-local node reuse.
- If the abstract contains a chain like `A -> B`, `B -> C`, and `X -> B`, the shared concept `B` should be represented by the same paper-local node 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: []`

```

Figure 21: Funding and journal tier interactions



Notes. Each cell averages over the benchmark cutoff-year cells in one funding-by-journal-tier slice. Rows separate papers that do or do not mention grant funding in the metadata, and whether the paper comes from a core or adjacent journal. Cell color again shows the graph score’s average percent recall advantage over preferential attachment on the broader shortlist-share cutoffs, while the number inside the cell reports the strict top-100 recall gain in percentage points. Positive values favor the graph score. The main message is not that funding has one stable sign. It is that the clearest positive cells are the adjacent-journal slices, especially the unfunded adjacent slice, while the funded-core cells are the most popularity-dominated. Because the funded cells are thinner and composition-sensitive, this figure is suggestive rather than central.

```
- `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.

Table 4: Direct-closure examples from the main shortlist

Example role	Anchor or pair	Surfaced question	What it illustrates
Anchored policy mechanism	Digital economy → environmental regulation	Does green innovation mediate the digital economy → environmental regulation relation?	Clean policy-facing mechanism question from the top shortlist.
Trade and innovation bridge	Trade liberalization → R&D	Does productivity mediate the trade-liberalization → R&D relation?	Cross-domain bridge with a readable mediator.
Environmental upgrading	Environmental quality → green innovation	Does policy uncertainty help explain the environmental-quality → green-innovation relation?	Anchored progression question in an environmental-policy neighborhood.
Innovation and market design	R&D → carbon emission trading	Does green innovation connect R&D to carbon-emission trading?	Mechanism-deepening question that links innovation to a policy instrument.

Notes. Each row is a surfaced question chosen from the path-to-direct shortlist. The second column reports the narrow benchmark anchor or endpoint pair. The third column reports the richer question a reader would actually inspect after the local support graph is compressed into readable prose. The table is included here because the paper’s current main examples have not yet been rebuilt in a paired family format.

```

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}}

D.2 Extraction schema

The model returns a paper-local graph with nodes and edges. The underlying output is JSON. Tables 5 and 6 restate the fields in reader-facing labels.

Table 5: Schema for paper-local nodes

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 6: Schema for paper-local edges

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.

Field	Allowed values / type	Meaning	Why it exists downstream
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 <i>describes</i> 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.
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.

Field	Allowed values / type	Meaning	Why it exists downstream
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.

D.3 Extraction 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 are available at github.com/prashgarg/frontiergraph.

E Normalization and Ontology

Node normalization is central because candidate generation, path counts, and missingness all depend on node identity. Paper-level concept strings vary in wording, scope, and granularity, so node definition cannot be treated as a secondary detail.

E.1 Why open-world normalization

Garg and Fetzer (2025) use paper-local extracted 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 benchmark needs to preserve distinctions such as public debt versus public investment, or monetary policy versus energy consumption.

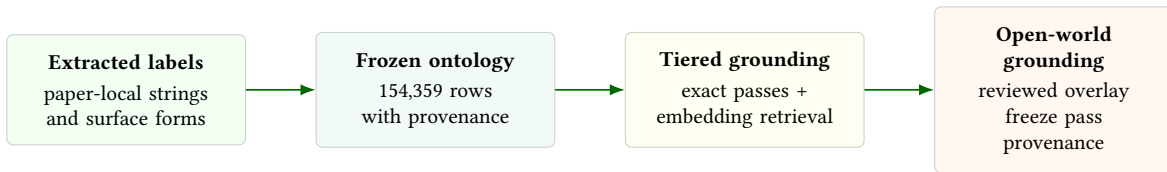
The difficulty is that the extraction layer and the ontology layer speak slightly different languages. Extraction produces paper-local phrases such as “gdp per capita”, “renewable energy consumption”, or “financial frictions”. The ontology contains more canonical labels. A useful grounding system therefore cannot be binary. If it only keeps exact or near-exact matches, it overselects the cleanest vocabulary and risks dropping real economics concepts from the graph. If it accepts every nearest neighbor, it creates false precision. The frozen ontology design used here is therefore *open-world*: it allows exact grounding, broader grounding, lower-confidence candidate bands, and unresolved labels that remain visible for later review rather than disappearing from the graph.

E.2 Implemented pipeline

The ontology build proceeds in five stages (Figure 22; Table 7 summarizes each stage).

Ontology assembly. The frozen ontology baseline starts from five structured source families rather than from the paper corpus alone. JEL provides a controlled economics taxonomy. Wikidata adds identifier-grounded concepts and aliases. OpenAlex topics and keywords add current research vocabulary. An economics-filtered Wikipedia crawl adds fine-grained named concepts, policies, instruments, and episodes

Figure 22: Node normalization and concept matching



Notes. This figure summarizes the frozen ontology grounding pipeline used in the benchmark. Paper-local labels are matched to a structured-source ontology by exact label and surface-form passes first, then by embedding retrieval. Lower-confidence labels are not simply dropped: broader grounding is allowed, reviewed outcomes remain explicit, and unresolved labels remain visible with provenance.

that appear in extraction labels but are absent from the structured sources. A small reviewed family layer is also carried forward. The resulting baseline contains 154,359 rows.

Label inventory and exact grounding. The extraction corpus contributes 1,389,907 unique normalized labels from 242,595 papers. Mapping begins with exact matches on the extracted label itself and then exact matches on stored surface forms, including stripped parenthetical variants. These deterministic passes solve the easy cases without using embeddings.

Embedding retrieval and confidence bands. Labels that remain unresolved after the exact passes are embedded and searched against the ontology with FAISS nearest-neighbour retrieval. The rank-1, rank-2, and rank-3 ontology candidates are stored for audit. The resulting grounding is tiered rather than binary: linked for scores at or above 0.85, soft for 0.75–0.85, candidate for 0.65–0.75, rescue for 0.50–0.65, and unresolved below 0.50. Across the 1,389,907 normalized extracted labels, 316,292 unique labels clear the primary 0.75 threshold before any reviewed open-world layer is applied. Lower-confidence bands remain visible rather than being silently dropped.

Reviewed open-world grounding. Lower-confidence labels are then audited with an overlay-first review pass. That layer keeps broader attachment to an existing concept, alias addition, proposed concept-family promotion, explicit rejection, and unresolved outcomes separate rather than forcing everything into one nearest-neighbour merge. Broader grounding is allowed when the ontology only contains a more general concept. That is intentional. For example, grounding “gdp per capita” to the broader concept “GDP” can be more informative than dropping the label entirely. Raw labels and raw edges are preserved throughout, so ontology weakness does not mechanically create spurious graph gaps.

Freeze pass and reviewed hierarchy. The conservative freeze turns that review work into a stable paper-facing ontology layer. It preserves the raw label, source, parent label, and root label fields, adds cleaned display labels where needed, and records reviewed effective-parent and effective-root overlays. The key design choice is conservative rather than exhaustive: the freeze accepts only a small number of

Table 7: Stages of the ontology pipeline

Stage	What it does	Why it is needed
Ontology assembly	Combines JEL, Wikidata, OpenAlex topics, OpenAlex keywords, and filtered Wikipedia into one deduplicated concept inventory.	Gives the graph a broad structured concept vocabulary rather than relying on a single taxonomy.
Exact grounding	Applies exact label, exact surface-form, and stripped-parenthesis matches.	Solves easy cases deterministically before any embedding step.
Embedding grounding	Uses nearest-neighbour embedding retrieval and stores rank-1 to rank-3 candidates.	Grounds the harder middle of the label distribution while keeping retrieval provenance.
Reviewed open-world grounding	Keeps broader attachments, alias additions, proposed family rows, explicit rejections, and unresolved outcomes separate.	Prevents the ontology tail from being treated as either fully trusted or silently dropped.
Freeze pass	Adds cleaned display labels, reviewed hierarchy overlays, and conservative duplicate review while preserving raw fields.	Stabilizes the paper-facing ontology baseline without rewriting source truth.

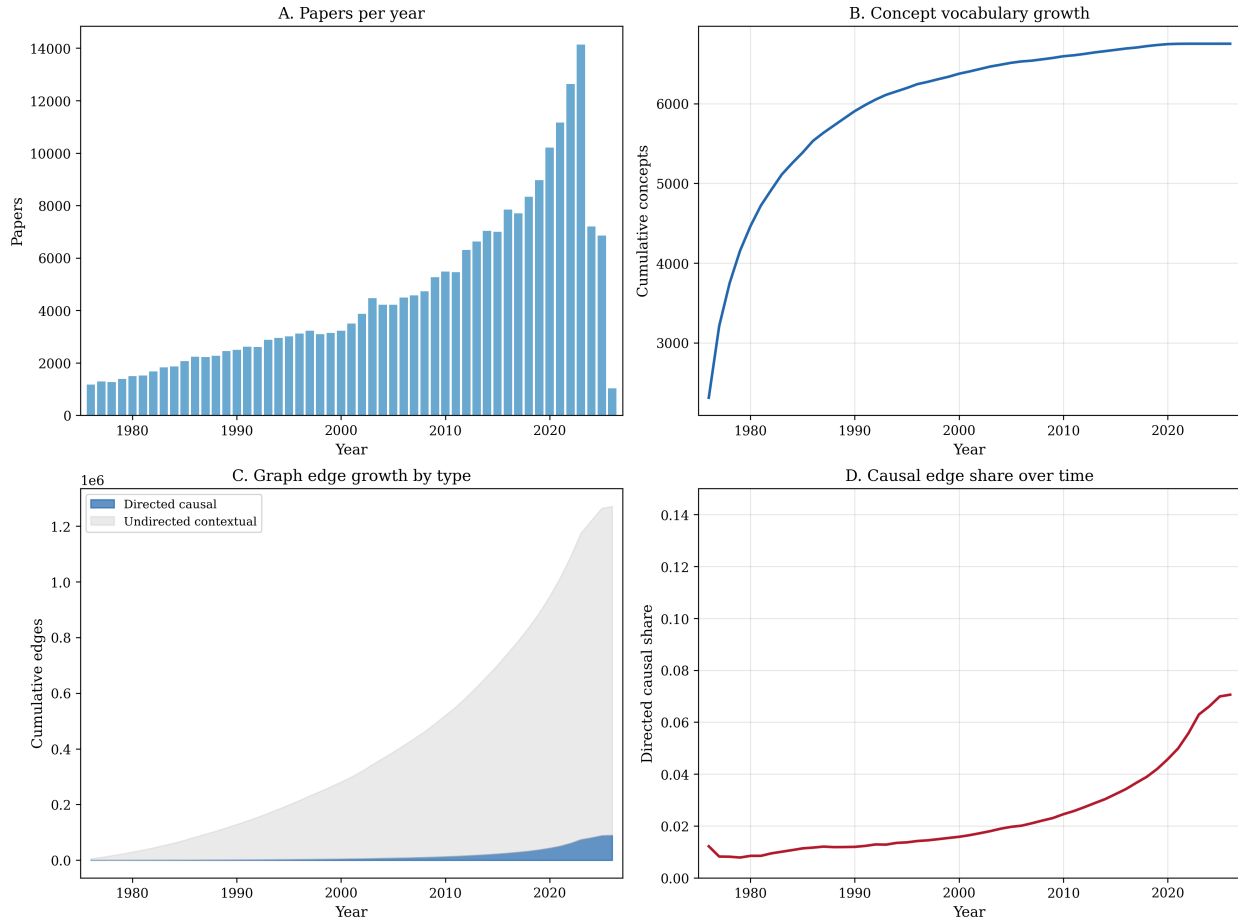
Notes. This is a process table, not an estimation table. Each row is one stage in the ontology pipeline that turns paper-local labels into the concept vocabulary used for candidate generation and benchmarking. The main distinction is between automatic grounding steps, which handle exact or embedding-based matches, and reviewed open-world steps, which keep lower-confidence cases visible instead of silently dropping them.

additional duplicate merges and leaves the remaining too-broad or unresolved cases explicit rather than silently patching them.

E.3 Preserving the raw graph and tiering the ontology

The key choice is not to let ontology confidence determine which literatures count as candidate-generating. If low-confidence labels were simply dropped, the benchmark would overselect the cleanest, most repetitive, and most canonical concept strings. It would also risk creating spurious novelty whenever a real but underrepresented concept failed to receive a clean ontology attachment. The solution used here is therefore to preserve the raw extraction graph while treating ontology grounding as a tiered interpretive layer. High-confidence grounded labels can be aggregated immediately. Lower-confidence labels can still receive broader grounding or reviewed unresolved status. The freeze then stabilizes the paper-facing ontology layer without pretending the remaining tail has been solved.

Figure 23: Corpus and graph growth over time



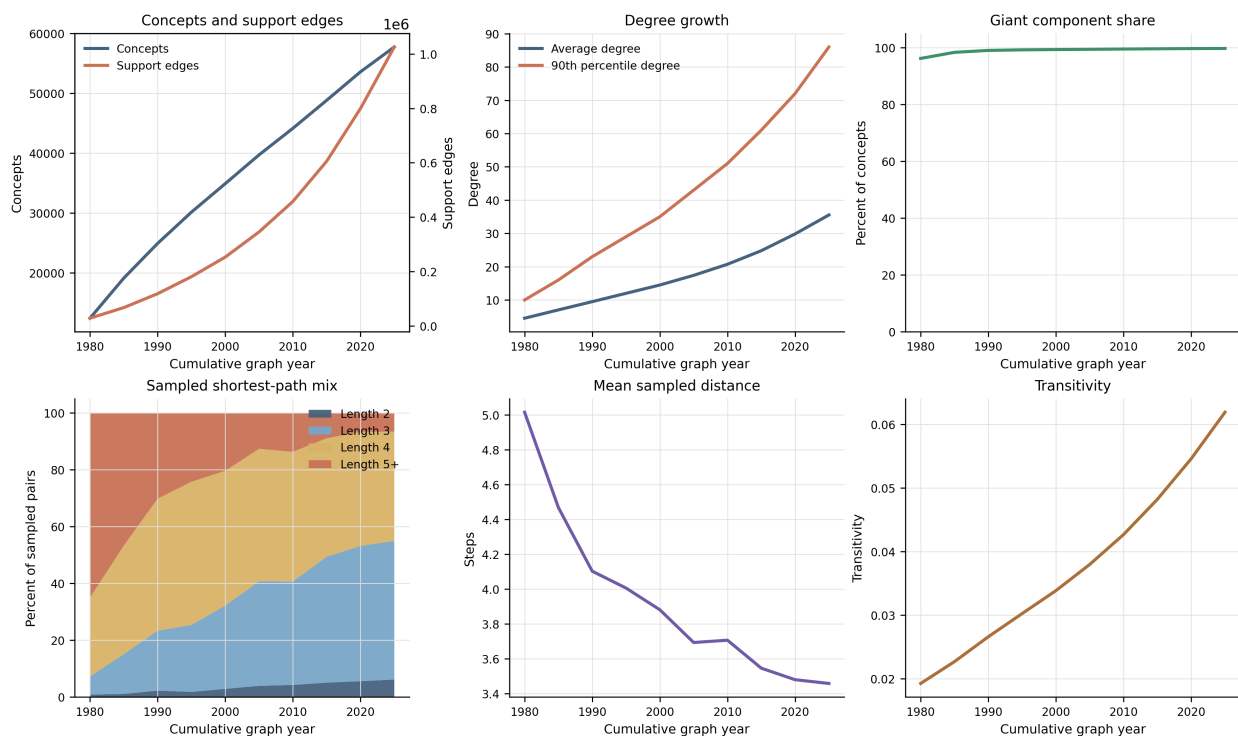
Notes. Panel A: papers per year in the selected journal corpus. Panel B: cumulative concept vocabulary growth. Panel C: cumulative edge growth by type (directed causal in blue, undirected contextual in gray). Panel D: causal edge share over time, reflecting the credibility revolution in economics. The share of directed causal edges rises from under 4% in the early corpus to over 10% by the 2020s.

F How the graph evolves over time

The benchmark uses a cumulative graph. Its behavior therefore depends not only on how candidate families are defined, but also on how the underlying graph changes as the literature grows. The figure below gives a simple descriptive account of that evolution using the support graph built from ordered and contextual links.

Four facts matter for the later path-length exercise. First, the corpus itself scales rapidly over time, which means the benchmark graph is being built on a much thicker published literature in the later decades than in the early ones. Second, the graph does not just add concepts; it becomes much denser. The support edge count rises from 28,055 in 1980 to 1,026,064 by 2025, while the average degree rises from 4.5 to 35.5. Third, the graph is already close to fully connected by the early benchmark era: the giant component covers about 96 percent of concepts in 1980 and nearly 100 percent by the end of the sample. Fourth, the graph

Figure 24: The live support graph becomes denser, shorter-path, and more connected



Notes. Each panel uses the cumulative support graph at five-year intervals. The top-left panel reports the number of unique concepts and support edges. The top-middle panel reports average degree and the 90th-percentile degree. The top-right panel reports the share of concepts in the giant connected component. The bottom-left panel reports the distribution of sampled shortest-path lengths inside that giant component. The bottom-middle panel reports the mean sampled path length. The bottom-right panel reports transitivity. Together the panels show that the graph becomes larger and much denser over time, while the typical distance between reachable concepts falls. That pattern helps explain why longer support paths may become empirically feasible even when they are harder to interpret.

also becomes shorter-path. In the sampled giant component, mean path length falls from about 5.0 to 3.5, and the mass of length-3 and length-4 paths becomes dominant.

These descriptives do not by themselves say that longer support paths should be used in the benchmark. They do say that a path-length comparison is worth doing. Once the graph is this dense, a longer path is no longer a rare accident. It is a common object. The real question is whether those longer paths still carry useful signal once screening quality and interpretability are judged side by side.

F.1 Path-definition robustness for the direct-to-path asymmetry

Section 5.2 reports the direct-to-path versus path-to-direct transition asymmetry using a length-2 definition of “supporting path” (one mediator, $u \rightarrow w \rightarrow v$). Because the support graph becomes denser and shorter-path over time, one natural robustness question is whether the asymmetry is an artefact of that path definition. Table 8 re-runs the same per-cutoff, per-horizon transition-rate computation under a length-3 definition (any three-hop chain, $u \rightarrow w_1 \rightarrow w_2 \rightarrow v$), pooled to the 1990s, 2000s, and

2010s cutoff-period blocks. Under L3, the direct-to-path rate roughly doubles at every horizon (e.g., at $h = 10$: 1.9% \rightarrow 10.3% \rightarrow 30.2% across the three decades, versus 1.1% \rightarrow 4.6% \rightarrow 12.9% under L2), while the path-to-direct rate falls because the eligible set grows faster than the realized-direct set. The asymmetry ratio therefore strengthens, not weakens: at $h = 10$ it goes from 8.2 (L2) to 35.9 (L3) in the 2010s. The mechanism-thickening finding is not an artefact of the length-2 graph.

Table 8: Mechanism thickening and direct closure under length-2 vs length-3 path definitions.

Period	Horizon	Direct \rightarrow path share		Path \rightarrow direct share		d2p/p2d ratio	
		L2	L3	L2	L3	L2	L3
1990s	$h = 5$	0.0020	0.0043	0.0023	0.0017	0.9	2.6
1990s	$h = 10$	0.0107	0.0191	0.0032	0.0025	3.4	7.6
1990s	$h = 15$	0.0251	0.0462	0.0036	0.0031	6.9	15.0
2000s	$h = 5$	0.0207	0.0416	0.0037	0.0025	5.6	16.9
2000s	$h = 10$	0.0464	0.1031	0.0074	0.0052	6.2	19.8
2000s	$h = 15$	0.0867	0.2056	0.0126	0.0094	6.9	21.9
2010s	$h = 5$	0.0621	0.1527	0.0073	0.0038	8.5	40.2
2010s	$h = 10$	0.1286	0.3022	0.0156	0.0084	8.2	35.9
2010s	$h = 15$	0.1597	0.3682	0.0181	0.0128	8.8	28.9

Notes. L2 uses the length-2 definition ($u \rightarrow w \rightarrow v$) of supporting paths used in the main text; L3 extends to length-3 chains ($u \rightarrow w_1 \rightarrow w_2 \rightarrow v$). The direct-to-path share is the rate at which pairs whose direct link already existed at $t - 1$ acquire a new supporting path within horizon h . The path-to-direct share is the rate at which pairs whose supporting path already existed at $t - 1$ later acquire a direct link. The ratio is direct-to-path over path-to-direct, with the 1990s, 2000s, and 2010s cutoff periods pooled across the six evaluation cutoffs. Under L3 the direct-to-path rate roughly doubles at every horizon, while the path-to-direct rate is diluted by the larger eligible set; the asymmetry is therefore not an artefact of the length-2 graph but strengthens as the path definition broadens.

G Benchmark Design and Significance

This appendix records the sample counts, benchmark tables, and significance summaries behind the main comparison. It does not introduce a second empirical object. Table 9 summarizes the two most important transitions: from selected journal papers to papers with extracted edges, and then from raw extracted edges to normalized evaluation links. The frozen ontology baseline is reported separately because the ontology inventory and the active benchmark graph are distinct objects.

The stricter identified-causal-claim layer is retained as a nested continuity benchmark. I do not use it as the main task, because it is much sparser than the broader causal-claim anchor used in the main text. But it is still useful as a conservative reference object, especially for readers who want continuity with the credibility-focused specification. Table 10 reports that continuity benchmark, and Table 11 reports the paired bootstrap comparison against preferential attachment on the same stricter layer.

Table 12 documents the reranker comparison inside the direct-closure family only. It is not a summary of the whole paper. The result is not one universal model. It is a short list of rerankers that all beat the transparent screening layer on the main metrics within that family. The paired family comparison stays

Table 9: 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
Frozen ontology baseline concepts	154,359
Directed causal rows	89,737
Undirected contextual rows	1,181,277
Total normalized links	1,271,014

Notes. This table reports the sample counts behind the benchmark. The rows move from the selected journal corpus to papers with extracted relations, then to the normalized graph used for candidate generation and evaluation. “Frozen ontology baseline concepts” refers to the paper-facing concept inventory after the reviewed freeze pass. “Total normalized links” is the active graph edge count used in the benchmark.

Table 10: Continuity benchmark on the identified-causal-claim layer

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

Notes. This table repeats the benchmark on a stricter continuity layer in which only identified-causal claims count as realized links. Recall@100 is the share of later-realized links captured in the top 100 candidates. MRR rewards putting realized links nearer the top of the ranking. The levels are small because this stricter layer is much sparser than the main benchmark.

in the main text. The family-specific model detail stays here.

Table 13 makes the timing issue concrete. At every main horizon, the early cells contain far fewer realized positives and the surfaced top-100 draws on much younger support. The endpoint recent-share measures are also systematically higher in the early era. The later benchmark therefore should not be read as the same environment with less noise. It is a different graph regime with thicker and older local structure.

H Reranker Design

This appendix documents the learned reranker used in the expanded benchmark comparison. The reranker operates on the same missing-link candidate set as the transparent graph score. Its purpose is simple: ask whether graph-derived features, when allowed to be reweighted from data rather than fixed by design, can beat the stronger transparent baselines that the fixed-weight score does not.

Table 11: Paired bootstrap continuity comparison: graph-based score minus preferential attachment

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

Notes. Each entry compares the transparent graph score with preferential attachment on the same stricter continuity benchmark as Table 10. Δ is graph minus preferential attachment, so negative values favor preferential attachment. The p -values come from paired bootstrap resampling across cutoff-year benchmark cells.

Table 12: Expanded reranker results for the main family

Horizon	Selected reranker	P@100	Recall@100	MRR
$h = 5$	Logistic GLM + composition family ($\alpha = 0.01$, pool = 5000)	0.107	0.1189	0.0123
$h = 10$	Logistic GLM + composition family ($\alpha = 0.05$, pool = 5000)	0.212	0.1551	0.0104
$h = 15$	Logistic GLM + composition family ($\alpha = 0.01$, pool = 5000)	0.260	0.1528	0.0114

Notes. Each row reports the reranker specification that performs best at that horizon on the 1990–2015 direct-closure benchmark. Precision@100 is the share of the top 100 candidates that later realize, Recall@100 is the share of all later-realized links captured in that top 100, and MRR rewards putting realized links nearer the top of the ranking. This table is intentionally narrower than the paired main-text comparison. It documents how the richer reranker behaves inside one family while the paper’s main comparison keeps both families visible.

H.1 Model inventory

Table 14 summarizes each model in the benchmark family.

H.2 Feature families

The reranker’s features are organized into five nested families. Each family adds to the previous, so the complexity gradient is itself interpretable. Table 15 lists the families and their contents.

Every feature is computed from the historical graph through year $t - 1$. The reranker sees no paper text, no future edges or degrees, and no author or institutional identity. That constraint is what keeps the exercise a graph-screening benchmark rather than a free-form prediction model.

The same-field indicator in the Structural family deserves a brief note because its earliest version of this paper used a cheap proxy—whether the two endpoint identifiers shared a first character—that was confounded with identifier source (JEL, OpenAlex, Wikidata) rather than substantive field membership. The

Table 13: Early and late benchmark cells are different benchmark regimes

Horizon	Era	Cutoffs	Mean eval positives	Winner R@100	Support age	Endpoint recent share
$h = 5$	Early (1990–1995)	2	12.5	0.071	11.1	0.413
$h = 5$	Late (2000–2015)	4	103.0	0.171	20.6	0.314
$h = 10$	Early (1990–1995)	2	32.5	0.109	11.7	0.430
$h = 10$	Late (2000–2015)	4	214.5	0.154	20.4	0.320
$h = 15$	Early (1990–1995)	2	55.5	0.174	11.5	0.435
$h = 15$	Late (2000–2015)	3	268.0	0.141	18.5	0.323

Notes. Early means the 1990 and 1995 cutoffs; late means 2000 through 2015 where the horizon is valid. Mean eval positives is the average number of realized positives in the cutoff-year cell. Winner R@100 uses the horizon-specific learned reranker winner from Table 12. Support age and endpoint recent share are computed on that winner’s surfaced top-100. The early cells are smaller and more recent-surge-like.

Table 14: Benchmark model inventory

Model	Input signals	Tunable?	Role in the paper
Preferential attachment	Source out-degree \times target in-degree	No	Cumulative-advantage null
Degree + recency	Endpoint support degree + recent support prominence	No	Stronger transparent baseline
Directed closure	Path support + mediator count + local closure density	No	Stronger transparent baseline
Transparent graph score	Path support + gap + motif support – hub penalty (fixed weights)	No	Interpretable screening layer
Learned reranker	Up to 34 graph-derived features across 5 nested families	Yes (L_2)	Strongest graph-based benchmark

Notes. Each row is one ranking rule evaluated on the same dated candidate pool. “Tunable” indicates whether the method has fitted parameters or hyperparameters. The paper’s empirical comparisons are therefore model-for-model on a common benchmark object, not separate samples or separate candidate universes.

current feature is a proper multi-label field overlap: each of the 154,490 ontology concepts is labeled into a 14-field taxonomy (macro, micro, labor, trade, finance, development, public, urban, environment, health, IO, methods, history, political economy) by an LLM classifier run on every concept’s label and description (details in Appendix H). The indicator fires when the two endpoints share at least one labeled field. On the 75,000-row evaluation panel, the new feature agrees with the old first-character heuristic only 47.6% of the time—12.9% of pairs were wrongly flagged as same-field by the identifier-coincidence proxy, and 39.7% of genuine cross-source same-field pairs were missed. Re-tuning the reranker on the corrected feature leaves the qualitative result intact: the composition family wins at every horizon, and the reranker’s Recall@100 advantage over the transparent score is preserved to within 0.01 (Appendix H).

H.3 Training design

The walk-forward panel is constructed as follows. At each cutoff year t , the training corpus contains all edges published through year $t - 1$. The candidate pool is the set of missing directed links at that cutoff.

Table 15: Feature families in the learned reranker

Family	Features	What it adds
Base	1	The transparent graph score itself.
Structural	14	Path support, motif bonus, gap bonus, hub penalty, mediator and motif counts, co-occurrence count and trend, same-field indicator (see below), endpoint degree products for both direct and support subgraphs.
Dynamic	21	Support age, recency of most recent supporting edge, recent-window degree and incident counts for each endpoint, recent-share fractions.
Composition	31	Mean stability, evidence-type diversity, venue diversity, source diversity, and mean field-weighted citation impact at each endpoint and at the pair level.
Boundary + gap	34	Whether the two endpoints sit in different field groups with no co-occurrence, whether the pair has path support despite a missing direct link, and the local closure density around the pair.

Notes. The families are nested and cumulative. “Features” is the total number available up to that family, not the number added only at that step. The table is included to show how the reranker moves from one transparent graph score to a broader but still interpretable feature set.

Features are enriched from the training corpus only. The binary label records whether the candidate edge first appears during $[t, t + h]$.

At evaluation time, the model is trained on all cutoff-year cells strictly before t and evaluated on the cell at t . This prevents any information from the evaluation cutoff from entering the training set.

Two model families are tested. The first is a class-balanced logistic regression that produces calibrated probabilities. The second is a pairwise ranking model (RankNet-style) that learns from positive-versus-negative feature differences, optimizing the order of candidates directly. Both use L_2 (ridge) regularization. The regularization strength α is drawn from $\{0.01, 0.05, 0.10, 0.20\}$. Features are standardized using training-set statistics before fitting.

H.4 Best models

On the current main benchmark (path-to-direct, corrected `field_same_group` feature), the best reranker at every horizon uses the interpretable logistic model with the composition feature family; the chosen L_2 strength is $\alpha = 0.01$ at $h = 5$, $\alpha = 0.05$ at $h = 10$, and $\alpha = 0.01$ at $h = 15$. The composition family adds evidence-composition features—mean stability, evidence-type and venue diversity, and field-weighted citation impact at each endpoint—on top of structural and dynamic signals. For the direct-to-path family, the retune (Appendix H.5) selects the structural family at $\alpha = 0.01$ for $h = 5$ and $h = 10$ and a pairwise-logistic structural specification at $\alpha = 0.01$ for $h = 15$. The two families therefore favor slightly different reranker bundles, but both prefer the lightest L_2 shrinkage in the tested grid.

H.5 Same-field indicator: source and robustness check

The same-field indicator is computed from a multi-label field classification over the full 154,490-concept ontology. Each concept is classified into a 14-field taxonomy (macro, micro, labor, trade, finance, development, public, urban, environment, health, IO, methods, history, political economy) by an LLM applied to its label and description; a concept may receive one to four field labels. The indicator for a candidate pair (u, v) fires when u 's and v 's field sets have non-empty intersection. Concepts with no recoverable field labels (roughly 42% of the ontology, including paper identifiers and very-specific entity concepts) contribute zero. The classifier output and helper code are versioned at `data/ontology_v2/concept_fields.parquet` and `src/field_overlap.py`.

Relative to the first-character identifier heuristic used in earlier drafts of this paper, the new feature flips 52.4% of rows on the 75,000-row evaluation panel: 12.9% of pairs the proxy flagged as same-field (e.g., two JEL codes that both begin with “j”) no longer fire, and 39.7% of genuine cross-source same-field pairs (e.g., `jel:F13:Tariff` paired with the OpenAlex keyword `trade-policy`) now fire. Re-tuning the full grid on the corrected feature (`outputs/paper/191_reranker_retune_field_overlap/`) returns the following best configurations and metrics by horizon:

Table 16: Best-config comparison: legacy first-character heuristic vs corrected field overlap

Horizon	Legacy (first-character)			Corrected (field overlap)		
	Family	α	R@100	Family	α	R@100
$h = 5$	boundary+gap	0.01	0.163	composition	0.01	0.162
$h = 10$	composition	0.20	0.201	composition	0.05	0.193
$h = 15$	boundary+gap	0.01	0.183	composition	0.01	0.179

Notes. All configurations use the glm-logit model on the same panel. Absolute Recall@100 falls by 0.001–0.008 per horizon on the corrected feature because the legacy heuristic was spuriously predictive through its correlation with identifier-source availability. The improvement of the learned reranker over the transparent score is preserved (Recall@100 deltas of +0.060, +0.107, +0.106 under the corrected feature at $h = 5, 10, 15$, versus +0.048, +0.111, +0.108 under the legacy feature). All three horizons now select the composition family.

The takeaway is that the reranker’s improvement over the transparent score is not driven by the spurious identifier proxy: the same composition-family configuration wins at every horizon under the corrected feature, with essentially unchanged deltas versus the fixed-weight score. The $h = 10$ regularization strength drops from $\alpha = 0.20$ to $\alpha = 0.05$, consistent with a broader true-positive rate for the corrected feature requiring less L_2 shrinkage. The paired benchmark table in the main text (Table 2) is itself built on the corrected feature and the retuned configs reported above; the direct-to-path side uses a parallel retune on its own panel, which selects the structural family at $\alpha = 0.01$ for $h = 5$ and $h = 10$ and a pairwise-logistic structural specification at $\alpha = 0.01$ for $h = 15$. The headline reranker-versus-transparent gap is therefore reported under the corrected same-field signal across both families.

Coverage check for the conservative zero. A natural concern with zero-coding unlabeled concepts is that 42% of the raw ontology carries no field labels, which could plausibly be dead weight for the reranker.

The held-out panel at $h = 10$ tells a different story. Only 9.7% of candidate pairs have at least one endpoint unlabeled, and just 0.5% (191 rows out of 35,000) have both endpoints unlabeled. Realization rates fall monotonically with coverage: the baseline rate is 2.85% for pairs where both endpoints carry a field label, 1.06% for pairs with exactly one labeled endpoint, and 0.00% across all 191 pairs where both endpoints are unlabeled. None of the reranker’s top-100 predictions at any cutoff come from the both-unlabeled region. Unlabeled concepts are therefore concentrated in precisely the part of the candidate space that does not realize, so the conservative zero-coding of `field_same_group` costs the reranker very little predictive information.

H.6 Single-feature importance

To understand what drives the reranker, I evaluate each of the 34 features as a standalone ranker in an auxiliary diagnostic run on the same candidate family. This is directly comparable to the single-feature ablation in Gu and Krenn (2025), who rank 141 features by predictive power in their cross-science knowledge graph. Table 17 reports the top 10 features by precision@100 in that diagnostic.

Table 17: Top 10 single features

Rank	Feature	$h = 5$			$h = 10$			
		Family	P@100	Hits	Feature	Family	P@100	Hits
1	Direct degree product	structural	0.100	10.0	Target direct in-degree	structural	0.197	19.7
2	Target recent incident count	dynamic	0.083	8.3	Direct degree product	structural	0.183	18.3
3	Target direct in-degree	structural	0.082	8.2	Target recent incident count	dynamic	0.173	17.3
4	Path support	structural	0.075	7.5	Path support	structural	0.143	14.3
5	Hub penalty	structural	0.075	7.5	Hub penalty	structural	0.142	14.2
6	Co-occurrence count	structural	0.057	5.7	Co-occurrence count	structural	0.113	11.3
7	Transparent score	base	0.057	5.7	Transparent score	base	0.112	11.2
8	Motif bonus	structural	0.048	4.8	Motif bonus	structural	0.093	9.3
9	Motif count	structural	0.042	4.2	Motif count	structural	0.088	8.8
10	Closure density	boundary	0.042	4.2	Closure density	boundary	0.088	8.8
		+ gap				+ gap		

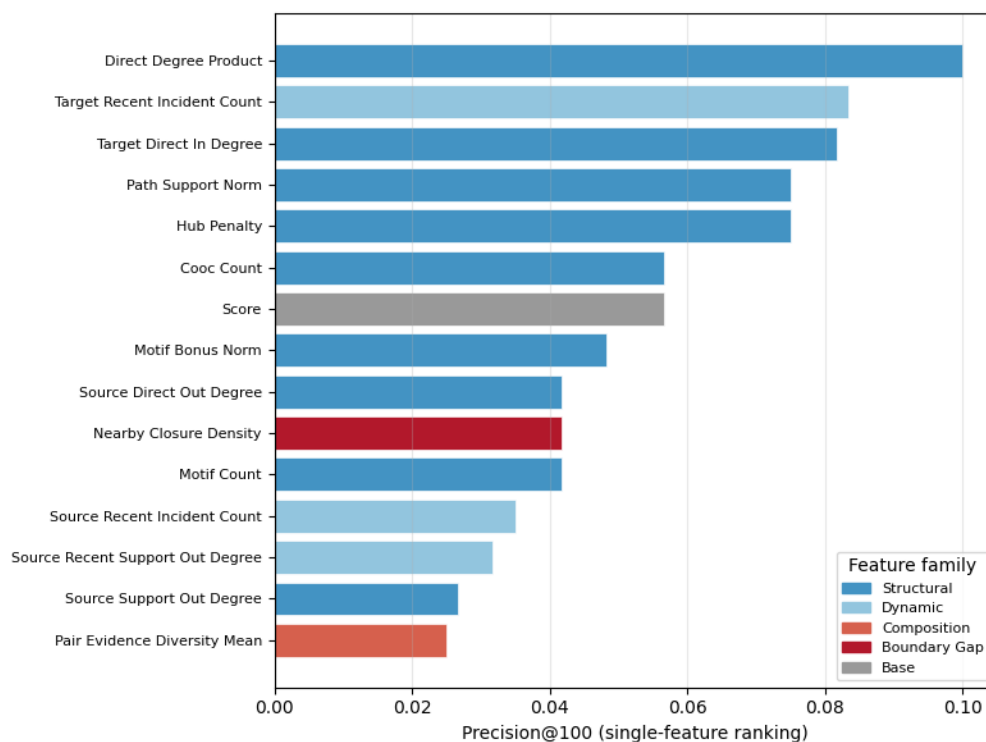
Notes. Each feature is used as a standalone ranker on the same candidate family. P@100 is precision at 100: the share of the top 100 candidates that later realize. “Hits” is the equivalent count of later-realized links per 100 surfaced candidates. The two panels compare the $h = 5$ and $h = 10$ diagnostics directly.

Three results stand out. First, directed-causal-degree features remain the strongest standalone signals. At $h = 5$, the direct degree product, the product of source causal out-degree and target causal in-degree, reaches 0.100 precision@100, compared with 0.057 for both raw co-occurrence count and the transparent score. At $h = 10$, target causal in-degree leads at 0.197, followed by the direct degree product at 0.183. These are popularity-style signals, but they are computed on the *directed causal subgraph*. They cannot be recovered from undirected co-occurrence alone.

Second, topology features still carry independent screening signal. Path support and the hub penalty rank fourth and fifth at both $h = 5$ and $h = 10$, ahead of raw co-occurrence count and ahead of the transparent score. So the reranker is not just reweighting centrality. Local path structure matters even as a standalone ranking rule.

Third, no single feature matches the full reranker on the same walk-forward benchmark. The value of the reranker is therefore still in the combination. But the single-feature ranking shows clearly that the strongest individual signals are not generic co-occurrence alone; they come from directed causal centrality plus local topology. Figure 25 shows the $h = 5$ ranking as a bar chart colored by feature family.

Figure 25: Single-feature importance by family ($h = 5$)



Notes. Each bar shows precision@100 when candidates are ranked by that single feature alone. Color indicates feature family: blue = structural, red = boundary/gap, salmon = composition, light blue = dynamic, gray = base. The top feature (direct degree product) requires the directed causal extraction and cannot be computed from co-occurrence.

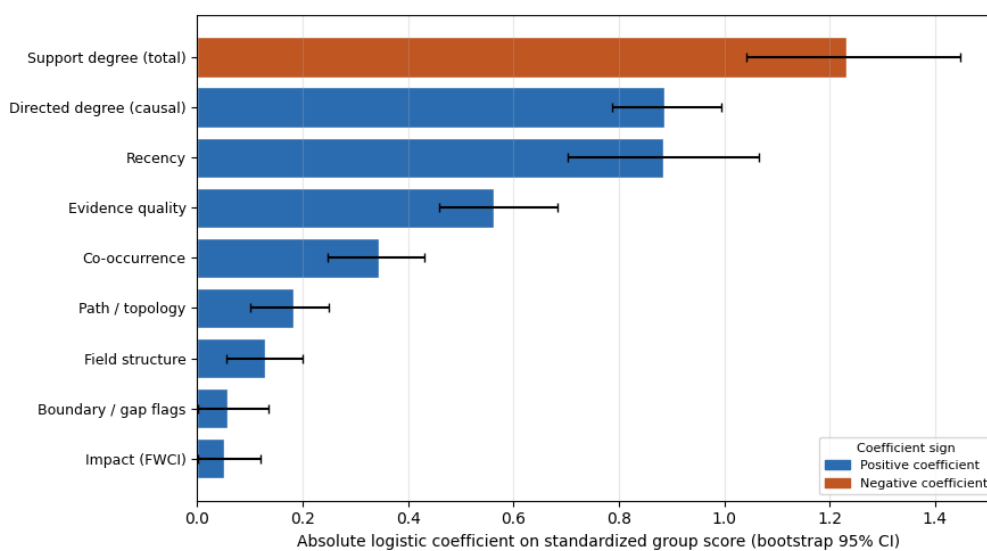
H.7 What the reranker uses

The single-feature ranking above shows which features carry standalone screening signal. But the raw-feature decomposition remains collinear. In this diagnostic, the top recent-support and support-degree measures have VIFs between 14 and 32 (Figure 29), and raw-feature importance rankings vary materially across model families: Spearman rank correlations are 0.28 for logistic versus gradient boosting, 0.39 for logistic versus random forest, and 0.53 for gradient boosting versus random forest. When features overlap

that much, Shapley or coefficient credit can move sharply across correlated inputs even when the model's total prediction is stable.

To resolve this, I group the 34 features into nine interpretable families, chosen *a priori* by what each feature measures rather than by statistical clustering: directed causal degree (endpoint degree on the causal subgraph), support degree (broader subgraph), recency (recent activity and incident counts), co-occurrence (paper co-mentions), path/topology (path support, motifs, hub penalty), evidence quality (stability, evidence diversity), impact (FWCI), boundary/gap flags, and field structure.¹⁶ I compute the first principal component within each group as a summary score, estimate a logistic reranker on those standardized group scores, and then inspect both group-level coefficient magnitudes and grouped SHAP plots (Figure 26).

Figure 26: Group-level importance after resolving multicollinearity



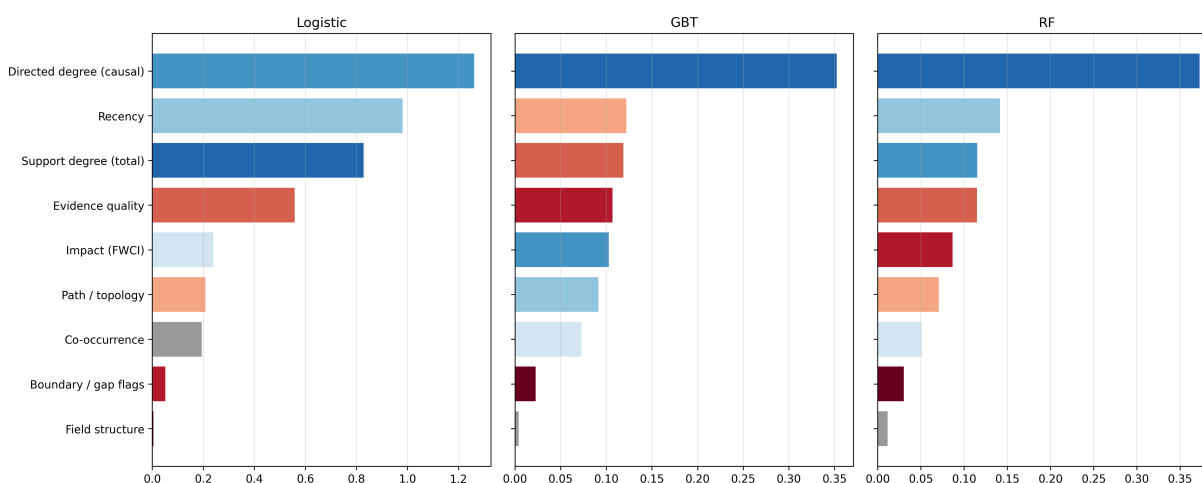
Notes. Each bar shows the absolute logistic coefficient on the standardized within-group PC1 score, with bootstrap 95% confidence intervals. Bar color shows the coefficient sign. Support degree is largest in magnitude but negative; directed causal degree, recency, and evidence quality are the strongest positive groups. Group-level VIF stays below 5.

Three results stand out. First, the largest coefficient by magnitude is support degree (1.23, bootstrap CI [1.04, 1.45]), but it enters with a *negative* sign. Once directed causal degree is controlled for, broad support-graph popularity hurts rather than helps. The model is learning to separate concepts that are prominent in causal work from concepts that are merely popular overall. Second, the strongest *positive* groups are directed causal degree (0.89, [0.79, 0.99]), recency (0.89, [0.70, 1.07]), and evidence quality (0.56, [0.46, 0.68]). So the reranker loads on a mix of causal centrality, recent activation, and the quality of the underlying evidence base. Third, co-occurrence remains useful (0.34, [0.25, 0.43]), but it is not the whole story. In economics, the central distinction is not simply between more versus less popular concepts. It is whether

¹⁶The grouping is defined before seeing the decomposition results, which prevents data-snooping. Within each group, the first principal component (PC1) explains 40–81 percent of the variance, confirming the groups are internally coherent. Using PC1 + PC2 does not materially change the ranking of the top families.

that centrality lives in the causal graph or only in the broader support graph.

Figure 27: Model comparison on grouped feature families



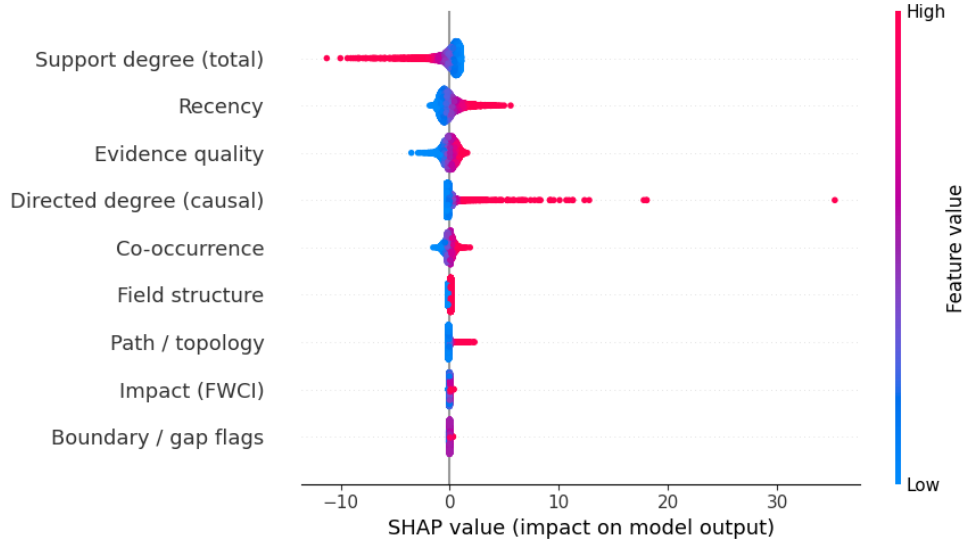
Notes. Tree-based models rank directed causal degree first. The grouped logistic model instead puts the largest weight on negative support degree, followed by recency and directed causal degree. Rank correlations across grouped models are moderate rather than perfect, but the main families are more legible than in the raw-feature decomposition.

The grouped decomposition improves interpretability more than it delivers a single invariant ranking. The gradient-boosting and random-forest models both rank directed causal degree first, while the grouped logistic model places the largest absolute weight on negative support degree and then on recency and directed causal degree. The grouped rank correlations are 0.30, 0.70, and 0.78 across model pairs: materially better than the raw-feature decomposition for two of the three pairs, but not close to unanimity. The right conclusion is therefore modest. Grouping makes the substantive families readable and reduces collinearity enough for stable sign interpretation, but model family still matters.

Robustness of the grouped decomposition. Figures 28 and 30 confirm that the grouping resolves the multicollinearity problem and produces stable, interpretable results.

Raw feature-level decomposition. The grouped analysis resolves collinearity but aggregates within families. Figure 31 complements it by showing raw feature-level Shapley importance for a logistic regression trained on the full 34-feature panel. The top six features by mean |SHAP| are all degree or topology measures: target support in-degree (1.30), target recent support in-degree (1.10), source support out-degree (0.92), target direct in-degree (0.70), normalized path support (0.51), and source direct out-degree (0.45). Two are total support degree, one is recency, two are directed causal degree, and one is path support. The coefficient signs confirm the grouped-level interpretation: total support-degree features enter negatively (broad popularity, once recency and directed degree are controlled for, lowers the predicted probability), while recency, directed causal degree, and path support enter positively.

Figure 28: Grouped SHAP beeswarm



Notes. Each dot is one candidate pair. Horizontal position shows the group’s SHAP contribution to the prediction; color shows the group’s PC1 value (blue = low, red = high). High directed degree, recency, and evidence quality tend to push predictions up. High support degree pushes predictions *down*, confirming the negative coefficient on broad support-graph popularity.

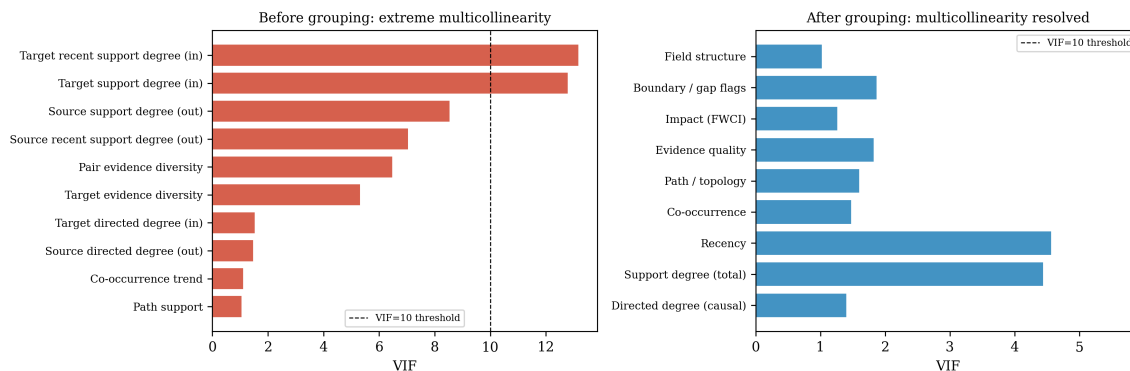
Prediction clusters. To see how the reranker applies this signal heterogeneously, I cluster the top-5,000 candidate pairs into five groups by their SHAP-value profiles and compare each cluster’s positive rate (Figure 32). The cluster driven by very high source recency (“hot source concept,” $n = 440$) has the highest positive rate at 13.6%, more than double the overall base rate of 6.2%. The cluster with low recency on both endpoints (“cold pairs,” $n = 2,944$) has the lowest at 4.7%. The remaining three clusters—high target recency (8.7%), high source recency (8.2%), and hot target concept (8.6%)—all beat the base rate. This confirms that the reranker’s screening value concentrates on concept pairs where at least one endpoint has recent research momentum, and the signal is not symmetric: the source concept’s recency profile matters more than the target’s.

Feature interactions. Pairwise interaction testing finds no significant synergy above main effects: the maximum interaction synergy is 0.004 (target recent support in-degree \times target recent incident count). This confirms that the logistic model is a reasonable approximation for this feature set and that the reranker’s screening value comes from combining individually informative features rather than from nonlinear interactions.

H.8 Failure mode profiles

To understand what the reranker gets wrong, I compare the feature profiles of its top-100 hits (realized links) and misses (unrealized links), averaged across cutoffs at $h = 5$ (Table 18).

Figure 29: Variance inflation factors before and after feature grouping



Notes. Left: the top raw-feature VIFs range from roughly 14 to 32 for recent-support and support-degree measures, indicating material multicollinearity. Right: after grouping into nine interpretable families, all grouped VIFs fall below 5. The grouping reduces the instability enough for coherent family-level interpretation.

Table 18: Failure modes in the reranker’s top 100 ($h = 5$)

	Hits (top-100, realized)	Misses (top-100, not realized)	Missed realized (rank > 500, realized)
Count	18	82	241
Mean co-occurrence	209	103	26
Mean direct degree product	3,413	3,288	290
Mean pair FWCI	5.35	5.55	5.71

Notes. Hits and misses are from the reranker’s top-100 predictions. “Missed realized” are links that the reranker ranks below 500 but that do appear within the horizon. Hits have much higher co-occurrence, confirming that the reranker succeeds on well-studied pairs. Missed realized links are in sparse neighborhoods with low degree products. This is the kind of sparse-literature setting where Sourati et al. (2023) argue human-aware models can add the most because content-only structure is thin.

The pattern is clear. The reranker’s successful predictions are concentrated among pairs that already have dense co-occurrence and high directed degree products. The links it misses, those that do realize but rank poorly, sit in sparse neighborhoods where popularity signals are weak and the graph provides little local structure to exploit. Those missed realizations have slightly *higher* endpoint FWCI, suggesting they are structurally surprising connections between reputable concepts rather than noise. This is the natural boundary of any graph-based screen: connections that arise from genuinely new combinations, serendipity, or shifts in methodology are hard to predict from existing structure alone.

H.9 Temporal generalization

A natural concern is that the walk-forward reranker may perform well only because all cutoff years were used during model development. To test temporal generalization, I select reranker configurations using only 1990–2005 cutoff cells and then evaluate them on the fully held-out era of 2010–2015, which the model

never saw during development or selection. During held-out evaluation, the graph itself is still built using only pre-cutoff evidence, but reranker training is frozen to the pre-2010 schedule.

Table 19: Held-out temporal generalization

Era	Model	Precision@100		Abs. reranker – PA	
		$h = 5$	$h = 10$	$h = 5$	$h = 10$
Train era (1990–2005)	Reranker	0.053	0.133	+0.045	+0.113
Train era (1990–2005)	Pref. attach.	0.008	0.020	n/a	n/a
Held-out era (2010–2015)	Reranker	0.210	0.375	+0.175	+0.285
Held-out era (2010–2015)	Pref. attach.	0.035	0.090	n/a	n/a

Notes. The table reports temporal holdout performance in levels rather than percentages. Precision@100 is the share of the top 100 candidates that later realize. The final two columns report the reranker’s absolute Precision@100 gap relative to preferential attachment, so positive values favor the reranker. The train-era winner is selected using only 1990–2005 cells and then evaluated again on the fully unseen 2010–2015 era.

The result is reassuring, but the right comparison is in levels rather than percentages. Because train-era preferential-attachment precision is very close to zero, percentage lifts are mechanically unstable. In absolute $P@100$ terms, the reranker still generalizes cleanly forward. At $h = 5$, its gap over preferential attachment rises from +0.045 in the 1990–2005 cells to +0.175 in the held-out 2010–2015 era. At $h = 10$, the gap rises from +0.113 to +0.285. The likely explanation is that the literature graph is thicker in later years, giving the reranker more structure to exploit. But the central conclusion does not depend on that interpretation: the learned graph features do not collapse out of sample when applied to a later era.

H.10 Regime and horizon checks

These appendix figures answer two interpretation questions that are useful but not central to the main comparison. Figure 33 asks where the transparent graph score gains the most over preferential attachment once the benchmark links are split by local density and by endpoint FWCI. Figure 34 asks how the transparent benchmark changes when the horizon is shortened or extended beyond the paper’s main 5, 10, and 15-year design.

I Support-Path Length as a Design Axis

The benchmark builds its candidate pool by traversing the directed support graph from each concept outward: for a pair (u, v) to enter the pool, there must be at least one directed path from u to v through one or two intermediate nodes in the historical graph at the relevant cutoff. The design parameter is `max_path_len`: the longest support path allowed. The production benchmark uses `max_path_len = 3`, which admits length-2 paths ($u \rightarrow w \rightarrow v$, one mediator) and length-3 paths ($u \rightarrow w_1 \rightarrow w_2 \rightarrow v$, two mediators). This section asks whether that choice is consequential, and extends the comparison to length-4 paths (three mediators) as well.

I.1 Comparison design

The comparison holds everything constant except `max_path_len`. Both runs use the same corpus, the same base and best-configuration hyperparameters, the same six evaluation cutoffs (1990, 1995, 2000, 2005, 2010, 2015), the same three horizons (5, 10, 15 years), the same pool ceiling of 5,000 candidates, and the same walk-forward training discipline. The reranker is tuned separately for each path length on the same grid. The comparison is therefore path-length-only: no other design difference is introduced between the two runs.

I.2 Headline results

Table 20 reports the headline metrics for both path lengths at the tuned-reranker stage. Table 21 reports the same metrics for the transparent score only.

Table 20: Learned reranker: path-length sensitivity

	$h = 5$			$h = 10$			$h = 15$		
	R@100	P@100	MRR	R@100	P@100	MRR	R@100	P@100	MRR
<code>max_path_len = 2</code>	0.163	0.103	0.0163	0.201	0.233	0.0120	0.183	0.340	0.0124
<code>max_path_len = 3</code>	0.169	0.108	0.0149	0.195	0.230	0.0118	0.179	0.338	0.0133
$\Delta(3 - 2)$	+0.007	+0.005	-0.0014	-0.007	-0.003	-0.0001	-0.004	-0.003	+0.0010

Notes. Each entry reports the best-adopted reranker specification at that horizon. R@100 is Recall@100: the share of all later-realized candidate edges in the evaluation window that appear in the top-100 shortlist. P@100 is Precision@100: the share of the top-100 shortlist that later realizes. MRR is mean reciprocal rank. Metrics are averaged over the four evaluation cutoffs valid at each horizon (1995, 2000, 2005, 2010). Bold entries mark the better path length at each metric-horizon cell where the margin exceeds 0.5 pp.

Table 21: Transparent score: path-length sensitivity

	$h = 5$			$h = 10$			$h = 15$		
	R@100	P@100	MRR	R@100	P@100	MRR	R@100	P@100	MRR
<code>max_path_len = 2</code>	0.115	0.060	0.0101	0.091	0.108	0.0094	0.075	0.153	0.0074
<code>max_path_len = 3</code>	0.127	0.063	0.0109	0.105	0.118	0.0091	0.085	0.168	0.0072
$\Delta(3 - 2)$	+0.012	+0.003	+0.0007	+0.015	+0.010	-0.0003	+0.011	+0.015	-0.0002

Notes. Same setup as Table 20, using the fixed-weight transparent score. The transparent score is not tuned; R@100 and P@100 differences here reflect the raw change in candidate composition. Bold entries mark the better path length where the margin exceeds 0.5 pp.

I.3 Reading the comparison

Four patterns stand out.

The transparent score favors path length 3 consistently. At every horizon, allowing paths up to length 3 improves the fixed-weight transparent score R@100 by 1.1–1.5 percentage points. The P@100 gain is 0.3–1.5 pp. These are not large improvements in absolute terms, but the direction is consistent and the mechanism is straightforward: length-3 paths expand the candidate set to include pairs with richer multi-hop graph support, and the transparent score’s path-support component directly rewards that breadth.

The learned reranker splits by horizon. At $h = 5$, path length 3 also improves the best-reranker R@100, by 0.6 pp (0.163 \rightarrow 0.169). At $h = 10$ and $h = 15$, path length 2 retains a small edge of 0.4–0.6 pp. All reranker differences are under 1 percentage point. The split reflects a timing effect: the reranker’s best $h = 5$ specification draws on boundary and gap indicators, which are sharp on a denser candidate universe; at longer horizons, the evidence-composition family dominates and benefits less from the expanded candidate pool.

P@100 is close across both path lengths. Precision@100 (the hit rate within the top-100 shortlist) differs by at most 0.5 pp for the reranker and at most 1.5 pp for the transparent score. The top-100 lists are compositionally different—roughly 7–8 candidate pairs swap into and out of each cutoff’s top-100 between path lengths—but the swapped pairs realize at similar rates in aggregate.

MRR is mixed but close. Reranker MRR favors path length 2 slightly at $h = 5$ (by 0.14 pp) but length 3 at $h = 15$ (by 0.10 pp). Transparent MRR slightly favors length 3 at $h = 5$ and $h = 10$. The differences are small and inconsistent in direction; MRR does not provide a decisive signal either way.

I.4 Which candidate areas shift

The two path lengths produce similar top-100 prediction distributions across JEL subject areas. The highest-hit-rate categories are the same under both: Development/Growth paired with Trade/Finance (hit rate 0.36 at $h = 10$), Development/Growth paired with OpenAlex keywords (0.29), and Math Methods paired with Trade/Finance (0.21). Low-hit categories also overlap: Health/Education concepts paired with OpenAlex keywords produce roughly 80 predictions per run but realize at near-zero rates.

The main compositional difference is that path length 3 introduces a Development/Growth \times Microeconomics cluster not present in the top prediction areas under path length 2: nine predictions at $h = 10$, three of which realize (33% hit rate), including Firm Size \rightarrow R&D (cutoff 2000) and Investment \rightarrow Economic Growth (cutoff 2010). Path length 2 surfaces two marginal Development pairs at the same cutoffs that path length 3 does not—Human Capital \rightarrow R&D and Foreign Direct Investment \rightarrow R&D—because those pairs slip below rank 100 when the denser candidate pool shifts relative scores at the margin.

I.5 Endpoint-anchored illustrations

To make the path-length comparison concrete for readers who want to see specific concept pairs rather than aggregate metrics, Figure 35 shows four curated endpoint pairs—one per JEL cluster (macro, trade, growth, health)—with their top supporting mediators under each path length. The pairs are fixed; only the mediator enumeration changes. The length-3 column surfaces genuinely new routes (e.g., Productivity \rightarrow Human Capital and Growth \rightarrow Human Capital as mediators of Inflation \rightarrow Economic Growth), but the top-ranked length-2 mediators remain the strongest supports in every case. The picture is consistent with the aggregate result in Section I.5: path length shifts the tail of the mediator distribution rather than the head.

Endpoint-anchored sweep tables

Figure 35 shows four pairs in detail. The endpoint-anchored sweep tables below (Tables 22–29) extend the comparison to full per-source top-20 target lists for eight source concepts spanning the main clusters of the taxonomy: Monetary Policy, Inflation, Unemployment, Productivity, Economic Growth, Human Capital, Emissions Trading, and Health Care Cost. For each source, the table shows the targets the reranker surfaces under length-2 and length-3, tagged “=” when the target appears in both top-20 lists and “+” when it only appears at that length. Across the eight sources, six of eight show 20/20 target overlap between the two path lengths; the remaining two (Human Capital and Health Care Cost) differ by a single target each. The re-orderings are small—target ranks shift by a few positions in either direction—confirming the aggregate result that path length reshuffles the ranked shortlist without systematically introducing or removing top targets.

I.6 Path length 4

Extending to path length 4 introduces a third mediator: $u \rightarrow w_1 \rightarrow w_2 \rightarrow w_3 \rightarrow v$. The enumeration cost is an order of magnitude larger than length 3, and the candidate pool per cutoff grows from roughly 2–5 million ordered pairs to 15–20 million at the neighbor cap we use. Running the full six-cutoff benchmark at length 4 required two targeted pipeline changes — streaming the feature panel to disk one cutoff at a time, and pruning length-4 paths whose support contribution falls below 10^{-1} of a unit edge weight before they reach the top-pool aggregation. Both are described in Appendix H. The quality-neutrality of the pruning threshold was verified separately on length-3 panels by confirming that the top 5,000 candidate support values are bitwise identical at thresholds up to 10^{-1} .

With those changes in place, the length-4 benchmark runs end to end on the same hardware as length 2 and length 3. Table 30 reports the tuned-reranker and transparent-score numbers for length 4 alongside length 3, restricted to the four cutoffs (1995, 2000, 2005, 2010) that both runs share.

Length 4 is strictly worse than length 3 on the learned reranker at every horizon and essentially flat on the transparent score. Reranker $R@100$ loses between 0.7 and 4.9 percentage points depending on the horizon,

Table 22: Top-20 targets for source **Monetary policy** (Macro / monetary) at cutoff 2010, horizon 10. “=” = appears under both path lengths; “+” = only at this length. Overlap: 20/20.

Length 2		Length 3	
5	= [paper 14704282]	5	= [paper 14704282]
9	= Inflationary Expectations	9	= Inflationary Expectations
12	= Inflation Targeting	13	= Inflation Targeting
15	= Real Exchange Rate	16	= Real Exchange Rate
23	= Exchange Rate	20	= Exchange Rate
28	= Exchange Rate Regime	26	= Money Supply
29	= Money Supply	27	= Exchange Rate Regime
36	= Fiscal Policy	35	= Fiscal Policy
42	= Central Bank Independence	40	= Central Bank Independence
58	= Welfare	56	= Welfare
79	= Stock Returns	82	= Stock Returns
154	= Interest Rates	150	= Interest Rates
163	= [paper 29583998]	158	= [paper 29583998]
204	= Price Level	212	= Output
215	= Output	216	= Price Level
236	= Phillips Curve	247	= [paper 498022]
246	= Monetary Target	255	= Phillips Curve
263	= [paper 498022]	263	= Monetary Target
297	= Nominal Interest Rates	285	= Nominal Interest Rates
308	= Central Banking	311	= Central Banking

with the largest loss at $h = 10$ where the denser candidate pool hurts the most. The transparent score is roughly unchanged: two small positives (+0.3 pp at $h = 5$ and $h = 15$) and one small negative (−0.9 pp at $h = 10$), none of them substantively meaningful. This holds despite the fact that length-4 paths do introduce new candidate pairs (a top-300-concept subgraph diagnostic at cutoff 2010 confirmed a roughly +29% increase in unique (u, v) pairs at length 4 versus length 3). The new pairs exist; they are just not predictive of realized co-occurrence.

Two mechanisms are consistent with the degradation. First, a three-mediator path appears to be a weaker prior on causal relevance than a two-mediator path: as the support chain lengthens, its correlation with the eventual realization of a causal edge decays faster than the raw weight of the path would suggest, even after the hub-discount and cycle filters applied during enumeration. Second, the reranker is tuned to pick out good candidates from a fixed-size top-5000 shortlist, and at length 4 the shortlist is noisier because the additional 29% of new candidates all compete for the same 5,000 slots. The learned features are less discriminative on this denser pool, which is why the reranker loss (−4.9 pp at $h = 10$) is larger than the transparent-score loss (−0.9 pp at the same horizon). Taken together, the length-4 result confirms that the axis is saturated at length 3: path length is not a margin on which the benchmark can be improved by going deeper.

Table 23: Top-20 targets for source **Inflation** (Macro / prices) at cutoff 2010, horizon 10. “=” = appears under both path lengths; “+” = only at this length. Overlap: 20/20.

Length 2		Length 3	
34	= Inflation Targeting	32	= Inflation Targeting
35	= Inflationary Expectations	33	= Inflationary Expectations
41	= Economic Growth	46	= Economic Growth
92	= [paper 14704282]	94	= [paper 14704282]
125	= Real Exchange Rate	132	= Real Exchange Rate
135	= Monetary Policy	142	= Monetary Policy
177	= Exchange Rate	167	= Exchange Rate
202	= Central Bank Independence	186	= Central Bank Independence
245	= Interest Rates	234	= Interest Rates
251	= Real Interest Rates	244	= Real Interest Rates
300	= [paper 29583998]	251	= Human Capital
304	= Money Supply	284	= Money Supply
341	= Human Capital	287	= Financial Development
348	= Exchange Rate Regime	304	= [paper 29583998]
367	= Nominal Interest Rates	323	= Exchange Rate Regime
498	= Financial Development	351	= Nominal Interest Rates
504	= Monetary Target	496	= Fiscal Policy
511	= Welfare	511	= Monetary Target
531	= Fiscal Policy	557	= Welfare
609	= Output Growth	631	= Output Growth

I.7 Production choice

The main benchmark uses `max_path_len = 3`. The evidence supports this from both directions of the axis. Moving from length 2 to length 3, the transparent score—the fixed-weight baseline that does not tune to historical outcomes—improves at every horizon by 1.1–1.5 pp on $R@100$, and at $h = 5$ the learned reranker also favors length 3; the $h = 10$ and $h = 15$ reranker difference slightly favors length 2 but the margin is under 0.7 pp. Moving from length 3 to length 4 (Table 30), the learned reranker loses at every horizon (by 0.7–4.9 pp on $R@100$, with the largest loss at $h = 10$) and the transparent score is essentially flat (max ± 0.9 pp). Length 3 is therefore not a tractability fallback but the upper bound of what the path-length axis can deliver. The richer candidate pool at length 3—relative to length 2—also better represents the diversity of questions the literature eventually asks, adding Development/Growth \times Microeconomics pairs that length 2 does not surface.

J Credibility Audit Summaries

The main score does not yet fully weight evidence quality, but the benchmark object is not blind to it either. This appendix exists to answer a narrow concern: is the graph mostly noisy co-mention, or does the benchmark sit on a substantively credible directed-claim layer? The extraction layer already records stability, causal presentation, evidence type, and related claim metadata. Tables 31, 32, and 33 should therefore be read as a quality audit of the empirical object rather than as a replacement ranking model.

Table 24: Top-20 targets for source **Unemployment** (Labor / macro) at cutoff 2010, horizon 10. “=” = appears under both path lengths; “+” = only at this length. Overlap: 20/20.

Length 2		Length 3	
97	= Unemployment Duration	113	= Unemployment Duration
203	= [paper 20581952]	202	= [paper 20581952]
220	= Unemployment Rate	229	= Unemployment Rate
265	= [paper 14704282]	265	= [paper 14704282]
321	= Inflation	317	= Inflation
338	= Welfare	406	= Welfare
419	= [paper 810954]	413	= Labor Demand
443	= Labor Demand	426	= [paper 810954]
486	= Wage Setting	434	= Wage Setting
526	= [paper 378588]	529	= Job Search
536	= Job Search	544	= [paper 378588]
587	= Trade Liberalization	611	= [paper 378430]
601	= [paper 378430]	665	= Wage Differentials
645	= [paper 38868622]	707	= [paper 38868622]
696	= Wage Differentials	735	= Trade Liberalization
824	= [paper 77212042]	793	= Wage Bargaining
830	= Wage Bargaining	822	= [paper 77212042]
858	= Inflation Targeting	876	= Inflation Targeting
939	= [paper 526807]	887	= Central Bank Independence
947	= Central Bank Independence	909	= [paper 526807]

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.

K Supplementary Usefulness Validation

These checks ask a different question from the historical benchmark. The historical benchmark asks whether a graph-grounded question later becomes part of the literature. The usefulness checks ask whether the rendered question looks useful and intelligible to a current reader. They speak to presentation quality and semantic coherence, not to historical forecasting. They remain in the appendix because they assess the surfaced question, not because they replace the benchmark.

The human exercise uses a small blinded pack built from the path-to-direct frontier. It compares 12 graph-selected items with 12 preferential-attachment-selected items drawn from the same candidate universe, with repeated sources and targets capped within arm so the pack does not collapse into one crowded endpoint neighborhood. Each item is shown as a compact graph-grounded question object plus a short construction note, and raters score readability, interpretability, usefulness, and artifact risk. Readability is intentionally narrow: it measures whether the labels are clean and easy to parse, not whether the underlying idea is substantively better. On this 24-item pack, the overall mean score is identical across arms at 3.22. Graph-selected items score somewhat higher on interpretability (3.17 versus 2.92), slightly higher

Table 25: Top-20 targets for source **Productivity** (Growth / firms) at cutoff 2010, horizon 10. “=” = appears under both path lengths; “+” = only at this length. Overlap: 20/20.

Length 2		Length 3	
6	= R&D	7	= R&D
54	= Economic Growth	64	= Economic Growth
136	= Firm Size	148	= Firm Size
209	= Labor Productivity	227	= Labor Productivity
238	= Spillovers	288	= New Economy
291	= New Economy	315	= Spillovers
325	= Unemployment	328	= Unemployment
376	= Wage Differentials	368	= Wage Differentials
452	= Trade Liberalization	637	= Environmental Regulation
469	= Human Capital	642	= Human Capital
518	= Foreign Direct Investment	655	= Foreign Direct Investment
623	= Welfare	658	= Trade Liberalization
647	= Environmental Regulation	682	= International Transfer of Technology
704	= International Transfer of Technology	686	= Welfare
754	= [paper 20747072]	736	= [paper 20747072]
810	= Terms of Trade	820	= Terms of Trade
970	= Labor Demand	934	= Labor Demand
1057	= Agglomeration	1132	= Agglomeration
1063	= [paper 23273]	1314	= [paper 23273]
1241	= [paper 1554529]	1351	= [paper 1554529]

on usefulness (3.00 versus 2.92), and slightly lower on artifact risk (1.58 versus 1.67), while preferential-attachment items score higher on readability (3.83 versus 3.50). The right reading is cautious external validation, not a broad human-rated advantage.

The appendix LLM usefulness sweep applies a related but more compressed question object on a much larger grid. The model sees a simplified graph-grounded question object plus a short construction note. That makes the exercise useful as a coarse screening check for readability and artifact risk, but not as a substitute for the historical benchmark or for the richer current-frontier objects used in the human exercise. The model is explicitly instructed not to judge novelty at the cutoff, likely future success, or topic importance. In practice, this comparison is informative mainly at the level of broad screening differences across arms rather than as a fine ranking of individual questions. For that reason, it is reported only as supplementary appendix evidence.

L Additional Analysis

L.1 Bundle uptake

The main paper treats one predicted edge as the unit of analysis. This section asks whether the downstream unit is often larger. When a later paper realizes one historically predictable edge, does it usually realize that edge alone, or does it take up a small bundle of nearby graph-supported ideas in the same paper?

Table 26: Top-20 targets for source **Economic growth** (Growth / development) at cutoff 2010, horizon 10. “=” = appears under both path lengths; “+” = only at this length. Overlap: 20/20.

Length 2		Length 3	
13	= R&D	12	= R&D
25	= Welfare	23	= Welfare
63	= Trade Liberalization	39	= Trade Liberalization
80	= Innovation	85	= Innovation
102	= Human Capital	108	= Human Capital
165	= Labor Productivity	175	= Labor Productivity
188	= [paper 14704282]	188	= [paper 14704282]
371	= Welfare Effects	281	= Energy Consumption
512	= Emissions	362	= Welfare Effects
583	= Endogenous Growth Model	383	= Emissions
638	= Energy Consumption	574	= Technological Change
787	= Human Capital Investment	581	= Endogenous Growth Model
813	= Technological Change	823	= Human Capital Investment
983	= Spillovers	1048	= [paper 1705653]
1080	= [paper 1705653]	1165	= Population Growth
1176	= Financial Development	1229	= Financial Development
1206	= Population Growth	1521	= Spillovers
1644	= International Transfer of Technology	1641	= International Transfer of Technology
1763	= Pollution	1764	= Pollution
1771	= [paper 2351125]	1973	= [paper 2351125]

We build a historical edge-to-paper uptake spine that links each realized in-pool prediction to the later paper or papers that realize it. For *path-to-direct*, the realizing paper is the paper that first writes the direct causal edge. For *direct-to-path*, the realizing paper is the paper that first supplies the path event. We then deduplicate to the first realizing paper for each prediction-horizon pair and aggregate to the paper level. For each later paper we count how many historically predicted edges it realizes, whether those realized edges come from one family or both, and whether the realized edges are close in the earlier graph. Distance is measured on the historical support graph available at the relevant cutoff.

Table 34 shows a clear pattern. Most later papers realize exactly one historically predicted edge. Across horizons, only about five percent of realizing papers realize more than one predicted edge in the same paper. Mixed-family bundles are rarer still: only about 0.1 to 0.2 percent of all realizing papers, or about 2 to 3 percent of the multi-edge subset. The downstream object in this setting therefore still looks like a single paper-shaped question, not a broad mixed package.

When multi-edge uptake does happen, however, it is local rather than diffuse. Among the multi-edge papers, about 77 to 79 percent share at least one endpoint across realized predicted edges, about 43 to 46 percent share a mediator, and the median minimum graph distance is zero at every horizon. Figure 37 shows both the dominance of single-edge uptake and the fact that the small set of multi-edge bundles is overwhelmingly local and mostly within-family.

The main implication is interpretive. The budget result in Appendix B.3 does not seem to be driven by large numbers of later papers absorbing broad mixed bundles of predicted ideas. Instead, the wider-budget

Table 27: Top-20 targets for source **Human capital** (Labor / growth) at cutoff 2010, horizon 10. “=” = appears under both path lengths; “+” = only at this length. Overlap: 16/20.

Length 2		Length 3	
39	= R&D	41	= R&D
573	= Innovation	608	= Trade Liberalization
652	= Trade Liberalization	737	= Earnings
727	= Earnings	967	= Innovation
1070	= Specific Human Capital	1052	= Specific Human Capital
1146	= Spillovers	1187	= Endogenous Growth Model
1232	= Endogenous Growth Model	1344	= Human Capital Investment
1368	= Human Capital Investment	1464	= Fertility
1491	= Fertility	1905	= Technological Change
1847	= Technological Change	2161	= Wage Differentials
2087	= Wage Differentials	2288	= Output Growth
2345	= Output Growth	3134	= Spillovers
3064	= Schooling	3150	= Schooling
3943	= [paper 378430]	3873	= [paper 378430]
4569	= Earnings Differentials	4630	= Earnings Differentials
4865	= Labor Mobility	4987	= Labor Mobility
4933	+ Returns to Education		

advantage of *direct-to-path* appears to come mainly from broader coverage across many separate later papers. When a later paper does realize more than one predicted edge, those edges tend to be tightly local in graph space.

L.2 Adopter profiles

This section asks which kinds of papers and teams independently move toward the same graph-supported questions. The exercise is descriptive rather than causal. The claim is not that these adopters use Frontier Graph. The question is whether different question families tend to be taken up by different kinds of scientific actors.

We start from the first-realizing-paper version of the uptake spine and merge it with paper metadata, authorship records, and a targeted OpenAlex enrichment pass. Local corpus data provides paper year, subfield, venue, funding fields, team size, and local author history. Targeted OpenAlex API pulls supply focal-paper affiliation country and institution type, as well as author-level counts-by-year that we use to recover global career age and prior output up to the focal paper year. Coverage is high for the author-history variables (missing share about 0.2 percent) and good enough for the focal affiliation variables (missing share about 11 to 12 percent).

Table 35 shows the main pattern. Relative to *direct-to-path*, papers that realize *path-to-direct* predictions tend to have larger teams, higher funding incidence, and more cross-country collaboration. At $h = 10$, for

Table 28: Top-20 targets for source **Emissions trading** (Environment / public) at cutoff 2010, horizon 10. “=” = appears under both path lengths; “+” = only at this length. Overlap: 20/20.

Length 2	Length 3
64 = Kyoto Protocol	62 = Kyoto Protocol
178 = Emissions	204 = Emissions
261 = [paper 3295052]	271 = [paper 3295052]
290 = Tradable Permits	280 = [paper 70004130]
298 = [paper 70004130]	291 = Tradable Permits
368 = Climate Policy	361 = Climate Policy
453 = Abatement Cost	440 = Abatement Cost
545 = [paper 11376541]	528 = [paper 11376541]
698 = [paper 15532928]	687 = [paper 15532928]
756 = [paper 2348980]	781 = [paper 2348980]
806 = Greenhouse Gas	821 = [paper 3210893]
836 = [paper 3210893]	889 = Greenhouse Gas
886 = R&D	941 = R&D
1095 = [paper 7180775]	1153 = [paper 7180775]
1179 = Energy Efficiency	1188 = Energy Efficiency
1322 = Environmental Regulation	1346 = Environmental Regulation
1354 = Uncertainty	1384 = Uncertainty
1396 = [paper 8200857]	1413 = [paper 8200857]
1900 = Compliance	1880 = [paper 32971000]
1913 = [paper 32971000]	1884 = Compliance

example, *path-to-direct* adopters average 2.86 authors versus 2.39 for *direct-to-path*; 18.5 percent versus 5.4 percent have any recorded funder; and 30.0 percent versus 20.4 percent are cross-country teams. Table 36 shows that these differences are statistically precise in simple large-sample comparisons. Incumbent-share and bridge-share differences are not. *Path-to-direct* adopters are also somewhat younger in global career age.

The team-composition split in Figure 38 sharpens that difference. *Path-to-direct* uptake is much less likely to be solo-authored and more likely to come from teams of two to six authors. The outlet mix differs too, but the difference sits within economics rather than between economics and non-economics. *Path-to-direct* uptake tilts more toward energy, environment, development, and applied-health outlets such as *Sustainability*, *Energy Policy*, *Energy Economics*, *World Development*, and the *Journal of Development Economics*. *Direct-to-path* uptake is relatively more present in canonical economics, finance, and health-policy outlets such as the *American Economic Review*, *Journal of Banking & Finance*, *Journal of Economic Dynamics and Control*, *The Journal of Finance*, *Health Affairs*, and *Health Economics*. The top-five general-interest economics journals remain a small share in both families, so the difference is not driven by a handful of flagship outlets. The cleaner venue-tier split confirms that both families are overwhelmingly in economics-facing journals, with only a small adjacent share. The sharper separation is in team structure and resource intensity, with a secondary tilt in venue content and topic composition. The paper-type split points in the same direction: *path-to-direct* is relatively more concentrated in energy and environment, while *direct-to-path* is relatively more concentrated in finance, health-policy, and theory-or-methods papers.

Table 29: Top-20 targets for source **Health care cost** (Health / public) at cutoff 2010, horizon 10. “=” = appears under both path lengths; “+” = only at this length. Overlap: 19/20.

Length 2	Length 3
18 = Uninsured	17 = Uninsured
73 = Physician	70 = Physician
75 = Managed Care	71 = Managed Care
121 = HMO	111 = HMO
194 = [paper 217635]	183 = [paper 217635]
196 = [paper 6337653]	194 = Medicare
207 = Medicare	198 = [paper 6337653]
230 = Medicaid	225 = Medicaid
309 = Physician Supply	293 = Physician Supply
401 = [paper 29354893]	391 = Incentives
408 = Incentives	396 = [paper 29354893]
549 = Medicaid Managed Care	548 = Medicaid Managed Care
559 = [paper 24476986]	554 = [paper 24476986]
674 = Prospective Payment System	652 = Prospective Payment System
676 = Pay For Performance	657 = Pay For Performance
829 = Health Care Reform	801 = Health Care Reform
961 = [paper 7304558]	1009 = [paper 7304558]
1044 = Health Spending	1236 = Reimbursement
1288 = Reimbursement	1241 = Health Spending
1290 + [paper 73721558]	1295 + Universal Health Care

These descriptives suggest a plausible division of labor across the two question families. *Path-to-direct* ideas appear to be taken up by somewhat more coordinated and better-resourced teams, with more cross-border collaboration and slightly lower mean global career age. *Direct-to-path* uptake is broader and more diffuse: it appears in more papers overall, but more often in smaller or solo teams and less often in funded projects. If that pattern survives fuller work, it would sharpen the interpretation from the budget analysis. The two families would differ not only in what they surface, but also in which scientific actors tend to move toward them on their own.

M Reproducibility Map

This appendix maps each main-text and appendix object to the script and output directory that produced it. The intent is to make every numeric claim and figure in the paper traceable to a single source artifact.

Table 37: Main-text objects and their source artifacts

Paper object	Source artifact (under outputs/paper/)
Figure 8, Table 2 (dual-family benchmark)	193_dual_family_main_pairing_field_overlap/; built by script- s/build_dual_family_main_pairing_field_overlap.py
Main-text hit rates (§5.1)	193_dual_family_main_pairing_field_overlap/dual_family_main_ove

Paper object	Source artifact (under outputs/paper/)
Figure 9 (direct-to-path dominance)	path_evolution_comparison.png; length-3 robustness in appendix_graph_evolution.tex
Figure 10 (journal-tier split)	path_transition_mix_by_source.png
Figure 26 (grouped SHAP)	140_grouped_shap_refresh/grouped_bootstrap.csv
Figure 11 (held-out era)	temporal_generalization.png; full table in Appendix H.9
Corpus waterfall (Table 9)	data/graph_v2/ build logs

Table 38: Appendix objects and their source artifacts

Paper object	Source artifact (under outputs/paper/)
Table 12 (path-to-direct reranker)	193_dual_family_main_pairing_field_overlap/dual_family_main_overs (path-to-direct rows)
Table 13 (early vs late regime)	outputs/paper/early_late_regime/ summary CSV
Table 16 (legacy vs corrected feature)	191_reranker_retune_field_overlap/tuning_best_configs.csv (corrected); run 70 (legacy)
Coverage-check statistics (§H.5)	193_classifier_coverage_diagnostic/; built by scripts/build_classifier_coverage_diagnostic.py
Table 17 (single-feature top 10)	139_single_feature_importance_refresh/
Figure 25	139_single_feature_importance_refresh/feature_importance_h5.png
Figure 28, Figure 30, Figure 27	140_grouped_shap_refresh/
Figure 29	140_grouped_shap_refresh/vif_comparison.png
Figure 31, top-6 raw SHAP values (§H)	141_shap_robustness_refresh/bootstrap_importance.csv
Figure 32 (cluster positive rates)	52_hypothesis_discovery/cluster_profiles.png
Table 18 (failure-mode profiles)	outputs/paper/reranker_failure_modes/ summary CSV
Table 19 (held-out era)	outputs/paper/temporal_generalization/
Figure 33, Figure 34	regime_split_delta.png, auxiliary_horizon_comparison.png
Path-length L2 vs L3 (§F, Appendix I)	path_evolution_L2_vs_L3.tex; underlying CSVs in outputs/paper/path_evolution/
Extended abstract / extra results	extra_results.tex and extended_abstract_research_allocation.tex

Scripts that produce paper-specific artifacts are under scripts/ (build_dual_family_main_pairing_field_overlap.py, build_classifier_coverage_diagnostic.py, and the sibling tuning runners). The public repository (github.com/prashgarg/frontiergraph) contains the full extraction, normalization, and benchmark pipeline. The src/field_overlap.py helper computes the corrected field_same_group indicator used in both

Table 30: Length-4 versus length-3 (tuned, shared four cutoffs)

horizon	metric	len = 3	len = 4	Δ (4-3)
$h = 5$	Reranker R@100	0.169	0.150	-0.020
$h = 5$	Transparent R@100	0.127	0.130	+0.003
$h = 10$	Reranker R@100	0.195	0.146	-0.049
$h = 10$	Transparent R@100	0.105	0.096	-0.009
$h = 15$	Reranker R@100	0.179	0.172	-0.007
$h = 15$	Transparent R@100	0.085	0.088	+0.003

Notes. Averages are computed over the four evaluation cutoffs available in both runs (1995, 2000, 2005, 2010), using the widened per-cutoff outputs from run 184 (len_3) and run 185 (len_4). Length-4 was run at a per-mediator neighbor cap of eight with `min_path_contribution = 0.1` and per-cutoff panel streaming. Reranker configurations were tuned separately for length 4 on the same grid used for lengths 2 and 3; the selected configurations were the composition family across horizons (pairwise-logit at $h = 5$ and $h = 10$, glm-logit at $h = 15$). The main takeaway is the $h = 10$ reranker gap, which is the dominant failure mode of length 4: the larger candidate pool dilutes the top-5000 shortlist with weakly supported three-mediator pairs that the reranker cannot distinguish from the good candidates.

Table 31: 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%

Notes. Each row summarizes one edge layer in the extracted graph. “Rows” is the number of extracted relations, “Papers” is the number of papers containing at least one such relation, “Mean stability” is the extraction-model stability score averaged across rows, and “Explicit-causal share” is the share of rows whose language is coded as explicitly causal. The point is descriptive: the directed layer is smaller, but it is also the more stable and more explicitly causal slice of the graph.

families’ final benchmark runs.

References

- Ajay Agrawal, John McHale, and Alexander Oettl. Artificial intelligence and scientific discovery: A model of prioritized search. *Research Policy*, 53(5):104989, 2024. doi: 10.1016/j.respol.2024.104989.
- Pierre Azoulay, Joshua S. Graff Zivin, and Gustavo Manso. Incentives and creativity: Evidence from the academic life sciences. *RAND Journal of Economics*, 42(3):527–554, 2011. doi: 10.1111/j.1756-2171.2011.00140.x.
- 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>.

Table 32: 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%

Notes. This table restricts attention to directed-causal rows and then splits them by the main evidence method recorded in the extraction metadata. “Rows” is the number of extracted directed-causal relations in that method family. “Mean stability” and “Explicit-causal share” are row-level averages within the method family. The table shows that design-heavy empirical methods remain well represented inside the directed layer.

Table 33: Stability bands 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%

Notes. Each row shows how extraction rows are distributed across the paper’s three stability bands. The percentages sum to 100 within edge kind. The comparison is included to show that low-stability rows are relatively uncommon overall and especially uncommon in the directed-causal layer.

Jay Bhattacharya and Mikko Packalen. Stagnation and scientific incentives. NBER Working Paper 26752, National Bureau of Economic Research, 2020.

David M. Blei and John D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 113–120, 2006. doi: 10.1145/1143844.1143859.

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>.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. Quantifying memorization across neural language models. In *International Conference on Learning Representations (ICLR)*, 2023. URL <https://arxiv.org/abs/2202.07646>.

Chaomei Chen. CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, 57(3):359–377, 2006. doi: 10.1002/asi.20317.

Mathieu d’Aquin. On the role of knowledge graphs in AI-based scientific discovery. *Journal of Web Semantics*, 84:100854, 2025. doi: 10.1016/j.websem.2024.100854.

horizon	n_papers	mean_predicted_edge_count	share_multi_edge	share_mixed_family	share_shared_endpoint	share_shared_mediator	median_min_graph_distance
5	1779	1.054525	0.051152	0.001124	0.039348	0.023609	0.000000
10	2277	1.055336	0.052701	0.001757	0.040843	0.023276	0.000000
15	2638	1.055345	0.053071	0.001516	0.041698	0.022745	0.000000

Table 34: Bundle uptake summary by horizon.

horizon	candidate_family_mode	n_papers	mean_team_size	share_any_funder	share_cross_country	share_incumbent_local	share_bridge_author_local	mean_career_age_local	mean_career_age_global	mean_prior_works_global	share_mixed_family
5	direct_to_path	1638	2.230000	0.034000	0.182000	0.539000	0.170000	6.163000	21.868000	55.001000	0.001000
5	path_to_direct	143	2.804000	0.140000	0.294000	0.530000	0.146000	5.758000	15.862000	55.057000	0.014000
10	direct_to_path	2038	2.388000	0.054000	0.204000	0.535000	0.167000	6.299000	22.092000	58.243000	0.002000
10	path_to_direct	243	2.856000	0.185000	0.300000	0.544000	0.136000	5.914000	18.301000	59.829000	0.016000
15	direct_to_path	2306	2.477000	0.077000	0.213000	0.533000	0.165000	6.300000	22.028000	61.400000	0.002000
15	path_to_direct	336	2.789000	0.182000	0.280000	0.539000	0.127000	5.766000	18.269000	59.275000	0.012000

Table 35: Adopter-profile overview by horizon and family.

Teresa C. Fort, Nathan Goldschlag, Jack Liang, Peter K. Schott, and Nikolas Zolas. Growth is getting harder to find, not ideas. Center for Economic Studies Working Paper CES-WP-25-21, U.S. Census Bureau, 2025. URL <https://www2.census.gov/ces/wp/2025/CES-WP-25-21.pdf>.

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, Dashun Wang, and Albert-Laszlo Barabasi. Science of science. *Science*, 359(6379):eaa0185, 2018. doi: 10.1126/science.aao0185. URL <https://doi.org/10.1126/science.aao0185>.

Jacob G. Foster, Andrey Rzhetsky, and James A. Evans. Tradition and innovation in scientists’ research strategies. *American Sociological Review*, 80(5):875–908, 2015. doi: 10.1177/0003122415601618.

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

Xuemei Gu and Mario Krenn. Forecasting high-impact research topics via machine learning on evolving knowledge graphs. *Machine Learning: Science and Technology*, 6(2):025041, 2025. doi: 10.1088/2632-2153/add6ef. URL <https://arxiv.org/abs/2402.08640>.

Aboozar Hadavand, Daniel S. Hamermesh, and Wesley W. Wilson. Publishing economics: How slow? why slow? is slow productive? how to fix slow? *Journal of Economic Literature*, 62(1):269–293, 2024. doi: 10.1257/jel.20221653. URL <https://www.aeaweb.org/articles?id=10.1257/jel.20221653>.

Daniel S. Hamermesh. Citations in economics: Measurement, uses, and impacts. *Journal of Economic Literature*, 56(1):115–156, 2018. doi: 10.1257/jel.20161326. URL <https://www.aeaweb.org/articles?id=10.1257/jel.20161326>.

Johannes Hoelzemann, Gustavo Manso, Abhishek Nagaraj, and Matteo Tranchero. The streetlight effect in data-driven exploration. NBER Working Paper 32401, National Bureau of Economic Research, 2024.

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>.

Horizon	Metric	Path-direct	Direct-path	Difference	CI low	CI high	p	N path	N direct
5	Team size	2.804000	2.230000	0.574000	0.137000	1.012000	p=0.010	143	1636
5	Any recorded funder	0.140000	0.034000	0.106000	0.048000	0.163000	p<0.001	143	1638
5	Cross-country team	0.294000	0.182000	0.112000	0.035000	0.189000	p=0.004	143	1636
5	Share incumbent authors	0.530000	0.539000	-0.008000	-0.072000	0.055000	p=0.794	143	1628
5	Share bridge authors	0.146000	0.170000	-0.025000	-0.080000	0.031000	p=0.389	121	1253
5	Mean global career age	15.862000	21.868000	-6.006000	-7.886000	-4.126000	p<0.001	143	1628
5	Mean global prior works	55.057000	55.001000	0.056000	-8.808000	8.920000	p=0.990	143	1628
10	Team size	2.856000	2.388000	0.468000	0.161000	0.775000	p=0.003	243	2035
10	Any recorded funder	0.185000	0.054000	0.131000	0.081000	0.181000	p<0.001	243	2038
10	Cross-country team	0.300000	0.204000	0.096000	0.036000	0.157000	p=0.002	243	2035
10	Share incumbent authors	0.544000	0.535000	0.008000	-0.041000	0.058000	p=0.738	243	2026
10	Share bridge authors	0.136000	0.167000	-0.031000	-0.074000	0.012000	p=0.155	206	1580
10	Mean global career age	18.301000	22.092000	-3.791000	-5.550000	-2.032000	p<0.001	243	2026
10	Mean global prior works	59.829000	58.243000	1.585000	-6.603000	9.774000	p=0.704	243	2026
15	Team size	2.789000	2.477000	0.312000	0.073000	0.551000	p=0.010	336	2302
15	Any recorded funder	0.182000	0.077000	0.104000	0.062000	0.147000	p<0.001	336	2306
15	Cross-country team	0.280000	0.213000	0.066000	0.016000	0.117000	p=0.010	336	2302
15	Share incumbent authors	0.539000	0.533000	0.006000	-0.037000	0.049000	p=0.799	335	2292
15	Share bridge authors	0.127000	0.165000	-0.038000	-0.075000	-0.001000	p=0.044	281	1801
15	Mean global career age	18.269000	22.028000	-3.759000	-5.309000	-2.208000	p<0.001	335	2292
15	Mean global prior works	59.275000	61.400000	-2.124000	-9.371000	5.123000	p=0.566	335	2292

Table 36: Descriptive difference tests for the adopter-profile comparison.

Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804, 2023. doi: 10.1016/j.patter.2023.100804.

Anton Korinek. Generative ai for economic research: Use cases and implications for economists. *Journal of Economic Literature*, 61(4):1281–1317, 2023. doi: 10.1257/jel.20231736.

Mario Krenn and Anton Zeilinger. Predicting research trends with semantic and neural networks with an application in quantum physics. *Proceedings of the National Academy of Sciences*, 117(4):1910–1916, 2020. doi: 10.1073/pnas.1914370116.

Donggyu Lee, Hyeok Yun, Meeyoung Cha, Sungwon Park, Sangyoon Park, and Jihee Kim. Econ-Causal: A context-aware causal reasoning benchmark for large language models in social science. arXiv:2510.07231, 2025. URL <https://arxiv.org/abs/2510.07231>.

David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 58(7):1019–1031, 2007. doi: 10.1002/asi.20591. URL <https://doi.org/10.1002/asi.20591>.

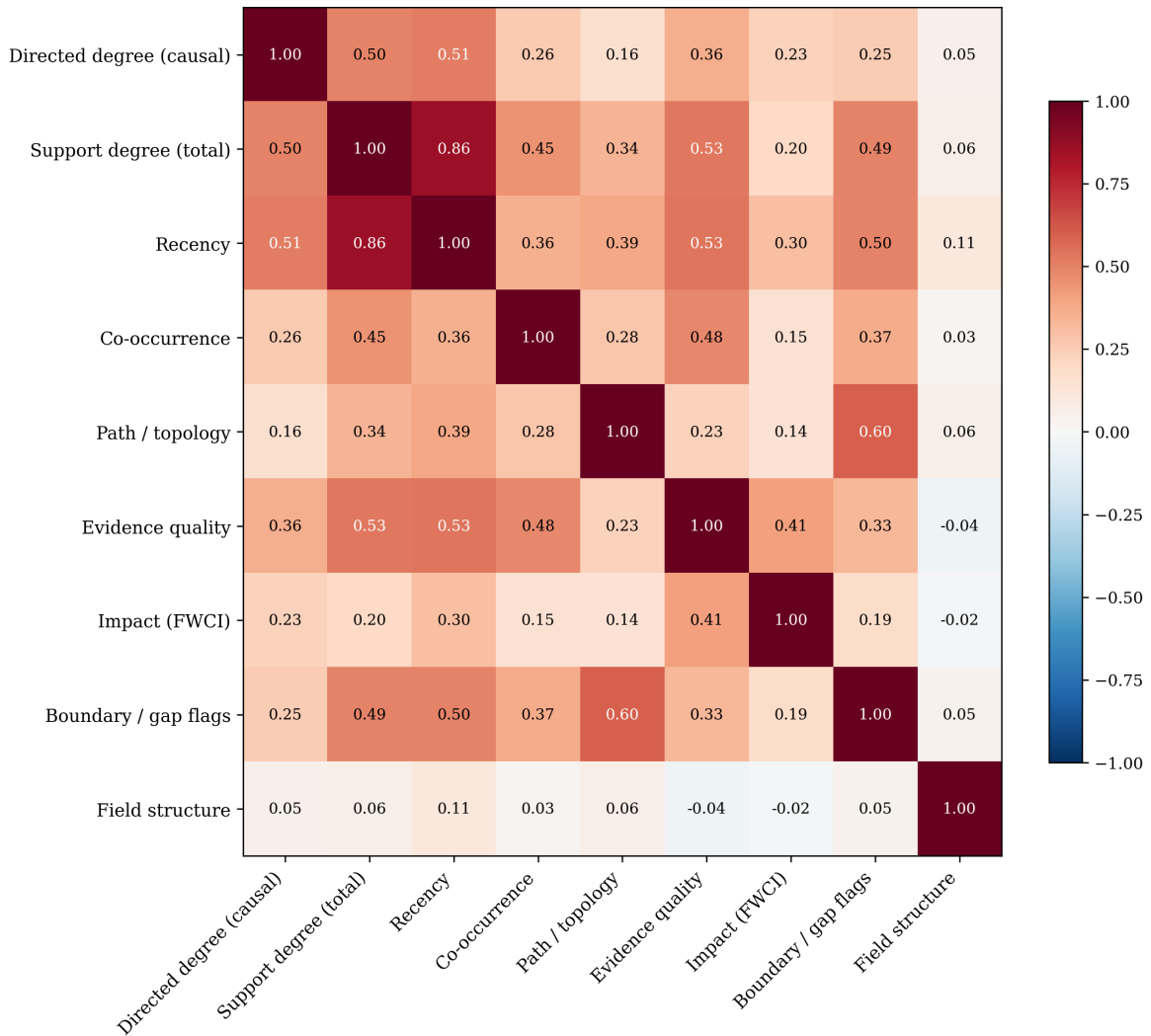
Tie-Yan Liu. *Learning to Rank for Information Retrieval*. Foundations and Trends in Information Retrieval. Springer, 2009. doi: 10.1561/1500000016. URL <https://doi.org/10.1561/1500000016>.

Victor Martinez, Fernando Berzal, and Juan-Carlos Cubero. A survey of link prediction in complex networks. *ACM Computing Surveys*, 49(4):1–33, 2016. doi: 10.1145/3012704. URL <https://doi.org/10.1145/3012704>.

- Michael Park, Erin Leahey, and Russell J. Funk. Papers and patents are becoming less disruptive over time. *Nature*, 613:138–144, 2023. doi: 10.1038/s41586-022-05543-x.
- Alexander M. Petersen, Felber Arroyave, and Fabio Pammolli. The disruption index is biased by citation inflation. *Quantitative Science Studies*, 5(4):936–953, 2024. doi: 10.1162/qss_a_00316.
- 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>.
- Andrey Rzhetsky, Jacob G. Foster, Ian T. Foster, and James A. Evans. Choosing experiments to accelerate collective discovery. *Proceedings of the National Academy of Sciences*, 112(47):14569–14574, 2015. doi: 10.1073/pnas.1509757112.
- 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>.
- Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. arXiv:2409.04109, 2024. URL <https://arxiv.org/abs/2409.04109>. NeurIPS 2024.
- Henry Small, Kevin W. Boyack, and Richard Klavans. Identifying emerging topics in science and technology. *Research Policy*, 43(8):1450–1467, 2014. doi: 10.1016/j.respol.2014.02.005.
- Jamshid Sourati, Fatemeh Faroughi, Esra Albayrak, and James A. Evans. Accelerating science with human-aware artificial intelligence. *Nature Human Behaviour*, 7:1682–1696, 2023. doi: 10.1038/s41562-023-01648-z.
- Don R. Swanson. Fish oil, Raynaud’s syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine*, 30(1):7–18, 1986.
- Song Tong, Kai Mao, Zhen Huang, Yukun Zhao, and Kaiping Peng. Automating psychological hypothesis generation with AI: When large language models meet causal graph. *Humanities and Social Sciences Communications*, 11(1):1606, 2024. doi: 10.1057/s41599-024-03407-5.
- 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>.

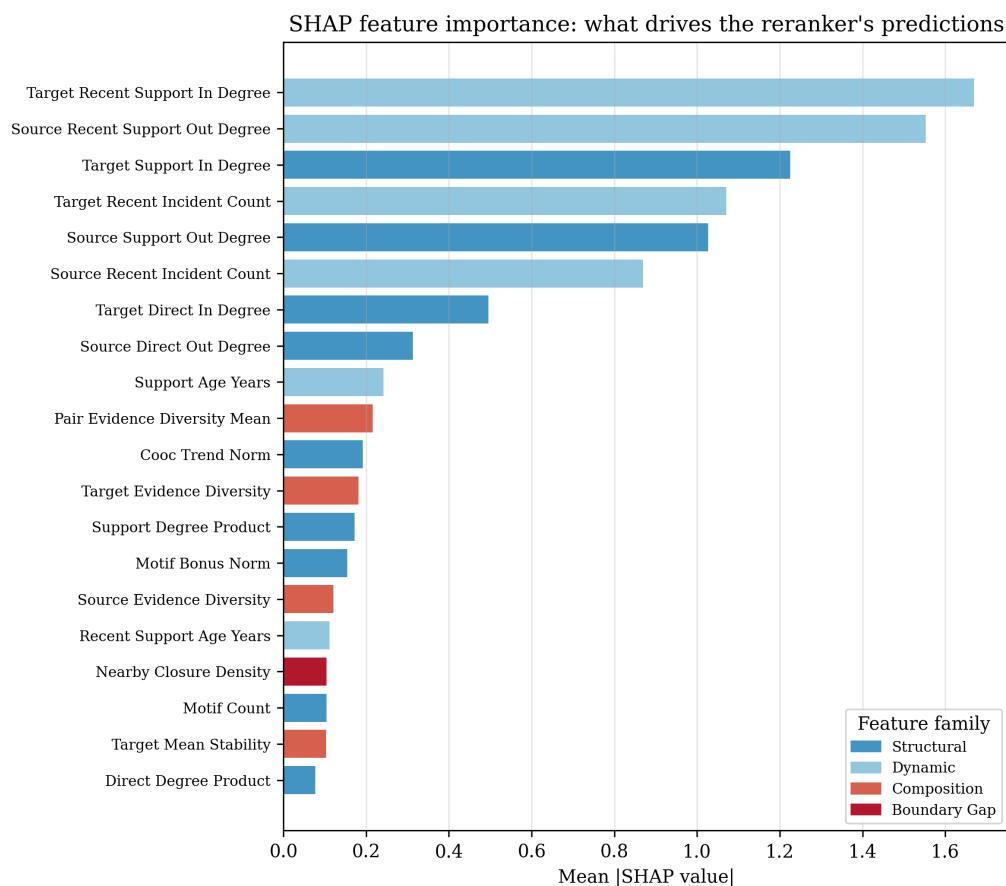
Yanbo Zhang, Sumeer A. Khan, Adnan Mahmud, Huck Yang, Alexander Lavin, Michael Levin, Jeremy Frey, Jared Dunnmon, James Evans, Alan Bundy, Saso Dzeroski, Jesper Tegner, and Hector Zenil. Exploring the role of large language models in the scientific method: from hypothesis to discovery. *npj Artificial Intelligence*, 1:14, 2025. doi: 10.1038/s44387-025-00019-5. URL <https://www.nature.com/articles/s44387-025-00019-5>.

Figure 30: Correlations across grouped feature families



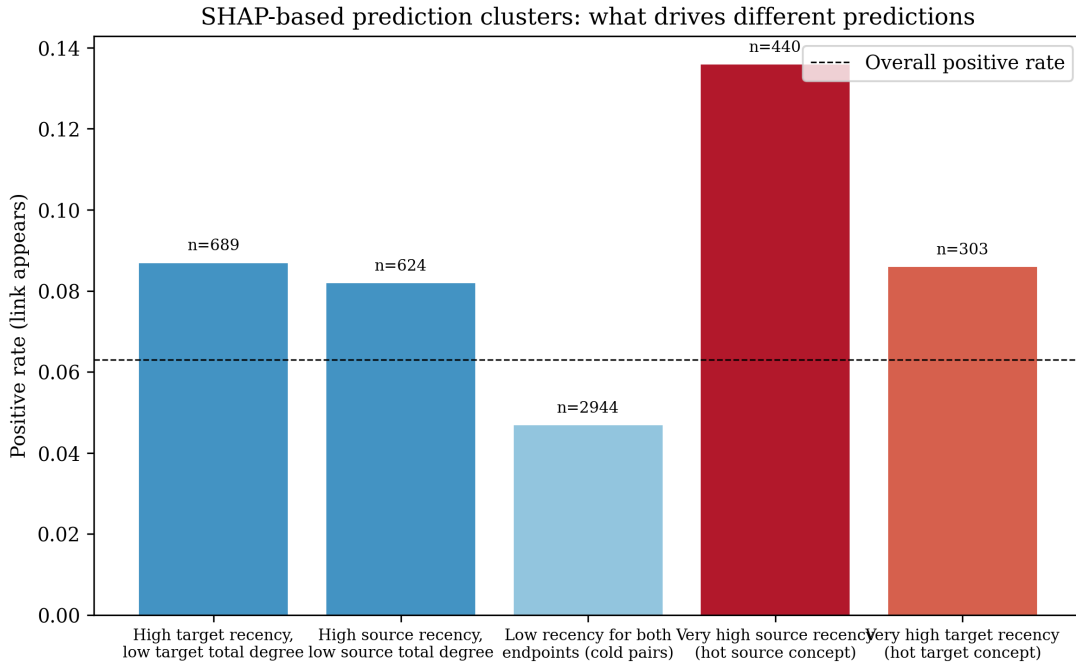
Notes. Pairwise correlations between the nine group-level PC1 scores. The strongest remaining correlation is between support degree and recency (about 0.86), so the grouped features are not orthogonal. But the grouped VIFs remain below 5, which is enough for substantially cleaner interpretation than at the raw-feature level.

Figure 31: Raw feature-level SHAP importance



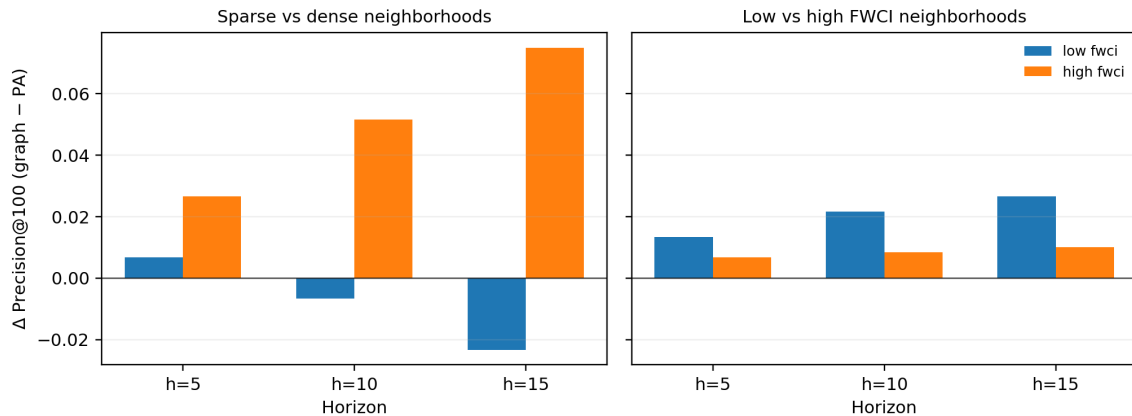
Notes. Each bar shows mean |SHAP| for the given feature across all candidate pairs in the training panel. Color indicates feature family. The top six features are all degree or recency measures. Composition, topology, and boundary features contribute individually small amounts but collectively provide the residual signal.

Figure 32: SHAP-based prediction clusters and positive rates



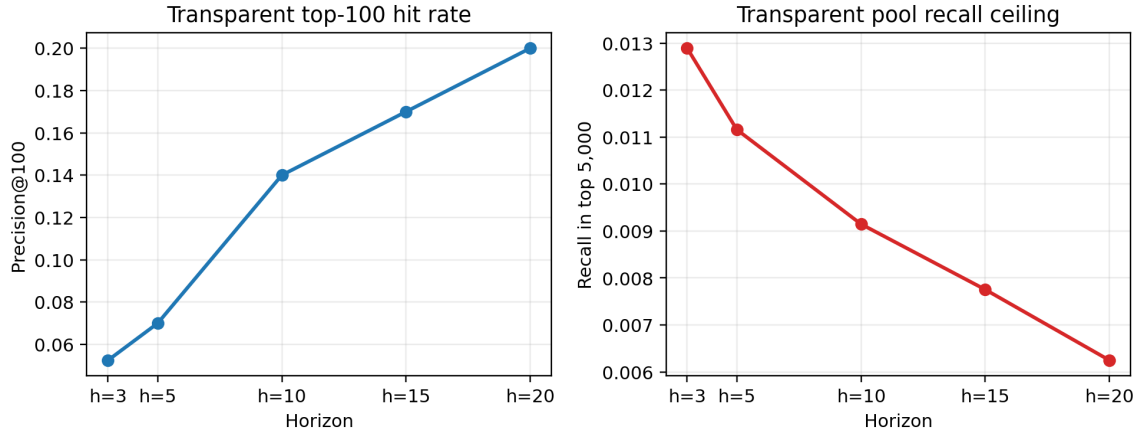
Notes. Each bar shows one k -means cluster defined on the SHAP-value vectors of the top-5,000 candidates, with the positive rate (share of edges that later realize) on the vertical axis. The dashed line is the overall positive rate. Cluster labels summarize the dominant SHAP driver. The “very high source recency” cluster is both the most predictive and the smallest, consistent with a “rising attention” signal that is rare but sharp.

Figure 33: The transparent score helps more in dense neighborhoods



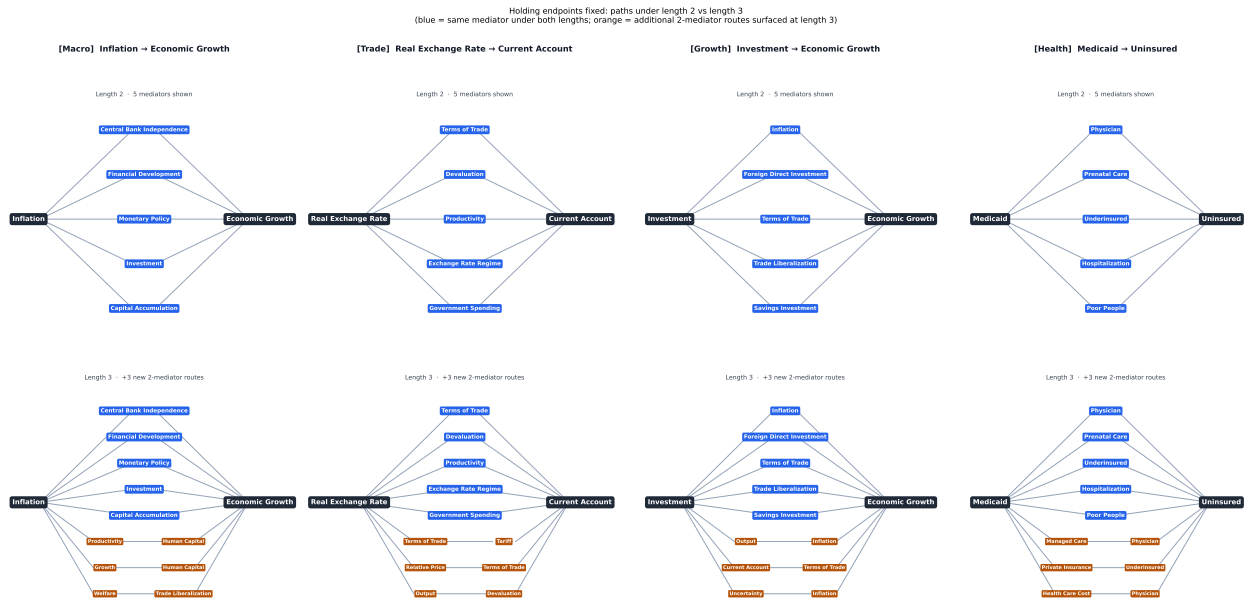
Notes. Each bar reports the transparent score’s mean Precision@100 advantage over preferential attachment on the main 1990–2015 benchmark panel, so positive values mean more later-realized links per 100 surfaced candidates. The left panel splits candidate pairs by whether they sit in denser or sparser local co-occurrence neighborhoods within the same benchmark cell. The right panel splits them by pair-level field-weighted citation impact (FWCI). The figure is included to show where the graph helps, not just whether it helps. Its edge is largest in denser local neighborhoods, and it is relatively larger in lower-FWCI slices than in the highest-FWCI ones.

Figure 34: Longer horizons raise hit rates, but the graph remains a screening layer



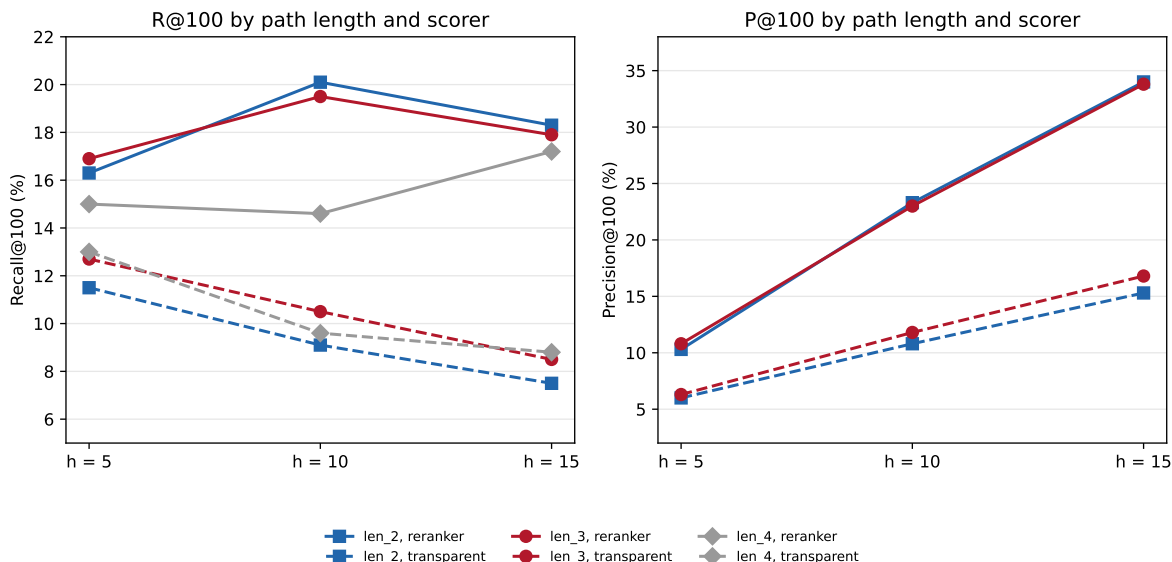
Notes. This figure uses the same fixed 5,000-candidate pool at each horizon. The left panel reports later-realized links per 100 surfaced candidates. The right panel reports Recall@100, the share of all later-realized links captured in the top 100 candidates, together with the much higher recall available if one inspected the full candidate pool. As the horizon lengthens, top-100 hit rates rise because more future links become eligible realizations. But Recall@100 remains small relative to the full-pool ceiling, which is why the paper treats the transparent score as a screening layer and studies reranking and reading budgets separately.

Figure 35: Curated endpoint pairs: length-2 vs length-3 supporting mediators



Notes. Each row fixes a concept pair from a different JEL cluster and lists its strongest supporting mediators under length-2 and length-3 enumeration. Blue nodes are length-2 mediators shared across both enumerations; orange nodes are routes that only appear under length 3 (i.e., require a two-mediator chain). Mediator rank is by log-hub-discounted support weight at the pair’s cutoff year. Paper-id mediators and bare Wikidata Q-codes are suppressed; only named economic concepts are shown.

Figure 36: Path-length sensitivity: R@100 and P@100 by horizon



Notes. Left panel: Recall@100 for each path length and scorer. Right panel: Precision@100 for lengths 2 and 3 (length 4 shown only in the left panel because the length-4 run reported R@100 only). Solid lines show the tuned reranker; dashed lines show the fixed-weight transparent score. All metrics are averaged over four evaluation cutoffs (1995, 2000, 2005, 2010). The transparent score consistently favors length 3; the reranker splits by horizon; length 4 falls below both on the reranker. All gaps between lengths 2 and 3 are under 1.5 pp.

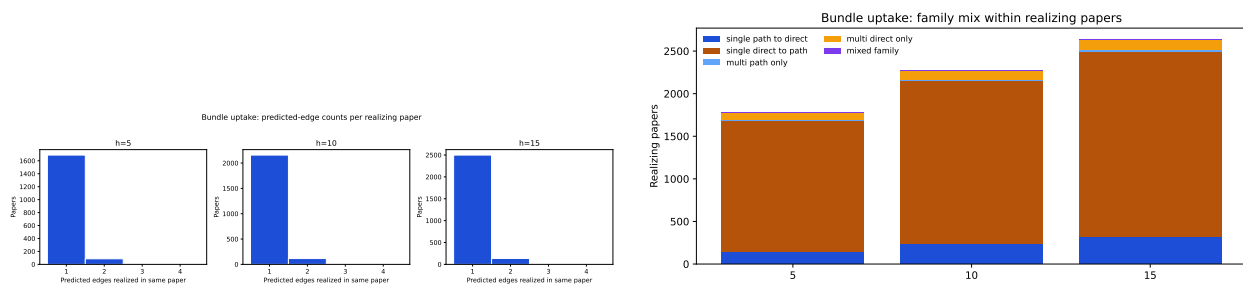


Figure 37: Bundle uptake descriptives. Left: the number of historically predicted edges realized in the same later paper is almost always one. Right: the small multi-edge subset is dominated by within-family bundles, especially *direct-to-path*.

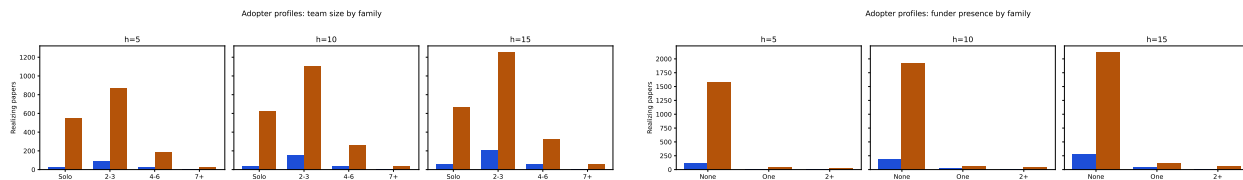


Figure 38: Adopter-profile descriptives. Left: *path-to-direct* uptake is less often solo-authored and more concentrated in two-to-six-author teams. Right: *path-to-direct* uptake is more likely to appear in papers with recorded funding.