TURNING CITATION NETWORKS INSIDE OUT: STUDYING SCIENCE USING CONTENT-BASED KNOWLEDGE GRAPHS FROM LLM-DERIVED TAXONOMIES

Seorin Kim[1,*] 0009-0001-5086-2704

Vincent Holst [1] 0009-0002-4117-4966

Vincent Ginis [1,2] 0000-0003-0063-9608

[1]Data Analytics Lab, Vrije Universiteit Brussel, 1050, Brussel, Belgium

[2]School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts, 02138, USA

September 21, 2026

ABSTRACT

Scientific fields are often mapped using citations and metadata, even though knowledge is transmitted primarily through content. We introduce an "inside-out" approach that reconstructs field structure directly from text by representing each paper as a small set of interpretable knowledge components. Using a large language model to induce domain-specific taxonomies and label papers, each publication is encoded as a triplet of measure, data type, and research question type. These triplets define a knowledge graph whose edges are weighted by the number of shared papers. Applied to 617 studies on intergenerational wealth mobility, the graph reveals a stable methodological backbone centered on regression-based mobility measures, alongside substantial temporal variation in the recombination of components. We further utilize normalized betweenness-to-connectivity ratios to identify components and pairings that act as structural bridges disproportionate to their prevalence. This content-derived, taxonomy-driven mapping complements citation-based approaches by exposing the evolving architecture of methods, data, and questions that define a field.

Keywords Knowledge Graph · LLM-generated Taxonomy · Citation Networks · Co-occurrence Analysis

* Corresponding author: seorin.kim@vub.be

1 Introduction

The structure and evolution of science have been studied through several complementary methodological lenses, each leveraging a different proxy for research content and supporting a different class of analytical questions. Citation networks, building on the seminal work of Garfield (1955) and Price (1965), use reference relations between articles to study intellectual lineage, the diffusion of ideas across communities, influential works, research fronts, and disciplinary boundaries (Radicchi et al., 2012). Keyword co-occurrence networks (Radhakrishnan et al., 2017) link publications through shared author-supplied terms and are most commonly used to identify thematic clusters, track the rise and decline of topics, and support systematic reviews (Grames et al., 2019; You et al., 2021; Yuan et al., 2022). Word-embedding approaches quantify conceptual proximity in high-dimensional semantic spaces and have been used to measure novelty, detect conceptual shifts, and characterize the position of a paper within a broader semantic landscape (Aceves & Evans, 2024; Shibayama et al., 2021).

Each of these approaches, therefore, captures a different aspect of scientific structure, but each also entails a different abstraction from the substantive content of papers. Citation networks encode explicit reference relations, yet citations may serve rhetorical, social, or strategic functions (Brooks, 1986; Cozzens, 1989; MacRoberts & MacRoberts, 1989; Nigel Gilbert, 1977), and may sometimes be incidental rather than substantively engaged with the cited work (Greenberg, 2009; Horbach et al., 2021; Moravcsik & Murugesan, 1975; Valenzuela et al., 2015). Keyword co-occurrence networks are closer to content, but author-supplied keywords are sparse, inconsistently assigned, and heterogeneous across journals and authors (Sampagnaro, 2023). Word-embedding approaches use textual content more directly, but their high-dimensional representations are difficult to interpret (Aceves & Evans, 2024) and sensitive to training data, model choice, and parameterization (Yin et al., 2023).

Within these traditions, recent work has sought to move beyond metadata toward more direct representations of content. Concept-extraction approaches such as SciKGraph (Tosi & dos Reis, 2022; Tosi & dos Reis, 2021) build co-occurrence knowledge graphs from full-text concept mentions identified via BabelNet, treating concepts as the fundamental units of analysis and clustering them into overlapping topics. This line of work demonstrates the analytical value of content-level graphs, but it depends on external knowledge-base coverage, hand-tuned pruning, and post-hoc cluster agglomeration, which limit its applicability to highly technical or interdisciplinary domains. In parallel, large language models (LLMs) have shown promise for extracting structured information from scientific texts (Lu & Wang, 2025; Reason et al., 2024; Seltmann et al., 2025) and for automating categorization tasks (Lan et al., 2025; Wu et al., 2025), opening the possibility of building content-level graphs whose category systems are induced from the corpus itself rather than imported from external resources.

Building on this literature, we propose a structured, content-based knowledge graph that shifts the unit of analysis from documents, citations, or author-supplied keywords to epistemic components extracted from publication content. Specifically, each paper is represented through three domain-specific dimensions: the measure it uses, the data type it relies on, and the type of research question it addresses. Thus, the resulting graph is neither a citation network nor a

conventional keyword co-occurrence network. Rather, it is a tripartite content-based representation in which nodes correspond to interpretable epistemic categories and edges capture their co-occurrence within papers.

We test this framework on a curated corpus of 617 publications on intergenerational wealth mobility, a domain in which methodological choices (e.g., regression-based measures, transition matrices), data infrastructure (e.g., administrative registries, panel surveys), and substantive questions (e.g., intergenerational transmission, demographic differences) are sufficiently well-defined to support a tripartite content representation. We consider abstracts rather than full texts in this study to avoid biasing toward only full-access texts. Given the case study's initial focus on theoretical and empirical studies of intergenerational wealth mobility measures, we naturally prioritize research reports over review papers or commentary. As most research reports include abstracts, our sole consideration of them is justified.

Using GPT-o3-mini, we induce three taxonomies (i.e., 8 measure categories, 13 data types, and 9 research question types), and assign each paper a triplet on these dimensions. The resulting graph reveals a stable methodological backbone centered on regression-based measures from the 2006-2010 period onward, while showing that structural variability arises primarily in how components are combined rather than in which components dominate. The capacitance diagnostics further surface bridging combinations that are disproportionately central relative to their prevalence, including transient bridges that may represent forgotten methodological linkages.

Our framework differs from existing content-based mappings in three ways. First, it imposes a tripartite structure on the content space, encoding each paper as a triplet across three disjoint dimensions (a measure, a data type, and a research question type), rather than treating all content tokens as interchangeable nodes. Second, the components themselves are drawn from domain-specific taxonomies induced by an LLM, which provides a level of conceptual consistency and methodological grounding that heterogeneous author keywords typically lack. Third, we introduce normalized betweenness-to-connectivity ratios, which we call *capacitance*, that identify components and pairings acting as structural bridges disproportionate to their prevalence, surfacing recombination that frequency-based summaries miss. In this sense, the framework turns the usual citation network perspective "inside-out": instead of mapping science through the relations between documents, we map it through the substantive components out of which documents are built.

1.1 The Framework

We consider a literature review of intergenerational wealth mobility research as a case study to test our content-based co-occurrence knowledge graph approach for exploring a scientific field. For this literature review, we aimed to investigate how different measures are used to study intergenerational wealth mobility over time. We therefore sought papers that either empirically study wealth mobility using data or theoretically discuss its measurement.

As shown in Fig. 1, an initial pool of 16,819 publications was retrieved from OpenAlex (Priem et al., 2022) using keywords related to intergenerational wealth mobility and income and earnings mobility, since these strands of literature often rely on similar statistical models. Subsequently, this pool is further filtered using OpenAlex metadata and LLM

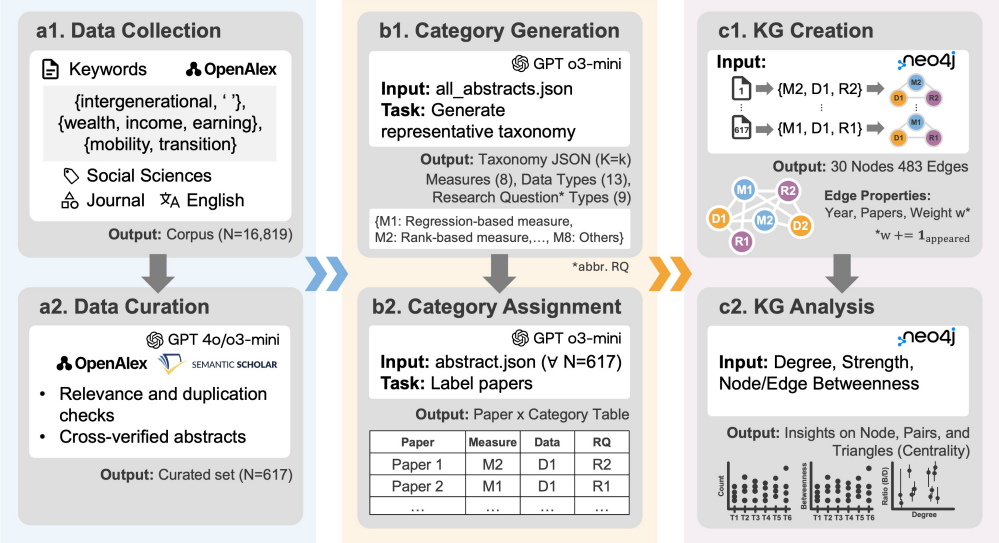


Figure 1: **Framework overview. a.** From an initial corpus of 16,819 publications, LLM-assisted curation identifies 617 relevant studies on intergenerational mobility. **b.** Using LLM-driven classification, each paper is assigned to three taxonomies – measures, data types, and research question types – derived from the LLM-generated category sets. **c.** These taxonomies form the basis of a knowledge graph where edges represent co-occurrence between categories, weighted by the number of shared papers, and the knowledge graph is analyzed with centrality measures.

categorizations. Using the metadata, we ensured that all publications are English-language journal articles in the field of Social Sciences. Using GPT-4o (`gpt-4o-2024-08-06`), we further ensured that the results were research papers by filtering out None entities in the LLM query to distinguish empirical and/or theoretical papers. Further, with GPT-o3-mini, study relevance is assessed using four sequential Yes/No questions, with one or two follow-ups for the first two (see the Supplementary Material, S1). This gave us a concise set of 742 abstracts that also fit within the GPT-o3-mini context window of 200,000 tokens, which we will use for label generation.

After the first major filtration, 742 abstracts are used altogether for GPT-o3-mini to obtain a list of categories for three dimensions (*Measure, Data Types,* and *Research Question Types*). While inspecting the LLM outputs, we noticed some duplicated work. Therefore, more manual inspection and cross-verification against metadata from Semantic Scholar and Crossref were conducted to correct erroneous metadata in OpenAlex, primarily to address duplicated abstracts. As a result, 617 English-language journal papers with valid abstracts were retained to assign the three-dimensional categories. During this inspection, one data type, *Rich List Data*, fell out because none of the 617 papers had this label. The LLM prompts for label generation and assignment are in S2.

With these LLM-empowered taxonomies, each paper is represented as a triplet of `{measure, data type, research question type}`. In the graph, only these paper properties are represented as nodes, and other paper information is stored cumulatively as an edge property, if necessary. In doing so, we assume that each resulting triplet corresponds to a

semantically coherent representation of a paper's content. This way, the structural relationships among measures, data types, and research question types are expressed through this content-based co-occurrence knowledge graph.

2 Materials and Methods

C1 of Fig. 1 illustrates the methodological workflow of our case study. Here, we focus on the content of papers to isolate patterns in how measures, data types, and research question types co-occur. For each analysis, we construct a tripartite graph whose three disjoint node sets correspond to these categories. Thus, edges connect only across partitions and encode period-specific co-occurrence. To study temporal dynamics, we partition the corpus into size periods based on publication years: `pre-2000 (T1)`, `2001-2005 (T2)`, `2006-2010 (T3)`, `2011-2015 (T4)`, `2016-2020 (T5)`, and `2021-2025 (T6)`. Note that the first period spans 1976-2000, given the small number of publications before 2000. Accordingly, the data structure is a multi-period multigraph: within each period, the network is simple (i.e., at most one weighted edge between any pair of nodes), but across all periods, the same pair of nodes may be connected by multiple parallel edges.

2.1 Nodes

Our corpus comprises 617 English-language papers in the Social Sciences. For each paper, we extract three categorical attributes from the abstract: Measure ($\mathcal{M}$), Data type ($\mathcal{D}$), and Research Question type ($\mathcal{R}$), which constitute the three disjoint node sets of the tripartite graph. All the categories are listed in Table 1. In total, 30 nodes are included in the graph with 8 $\mathcal{M}$, 13 $\mathcal{D}$, and 9 $\mathcal{R}$ nodes. One data type category, $\mathcal{D}_8$ (Rich List Data), is excluded because it does not appear in any of the 617 papers and would therefore be an isolated node with zero degree in every period.

Table 1: **LLM-generated nodes and their codes.** Three categories (Measures, Research question types, and Data types) are generated by GPT o3-mini based on 742 abstracts.

	Measures		Research Question Types		Data Type
M1	Regression-based Measures	R1	Measurement and Methodological Advances	D1	Panel/Longitudinal Surveys
M2	Rank-based Measures	R2	Empirical Estimates and Determinants	D2	Administrative/Registry Data
M3	Transition Matrix / Probability Measures	R3	Policy, Institutional, and Geographic Impacts	D3	National Survey Data
M4	Absolute Mobility Measures	R4	Intergenerational Wealth Mobility and Inheritance	D4	Opportunity Atlas
M5	Multigenerational Measures	R5	Demographic Differences in Mobility (Race, Gender, etc.)	D5	Natural/Experimental Data
M6	Decomposition / Structural Approaches	R6	Mobility and Non-Income Outcomes (Health, Wellbeing, etc.)	D6	Linked Administrative Data
M7	Non-parametric Approaches	R7	Theoretical and Structural Models	D7	International Panel Data
M8	Others_Measure	R8	Perceptions of Mobility and Attitudes	D8	University/Institution Data
		R9	Others_RqType	D9	Pseudo-Panel/Household Budget Survey
				D10	Archival/Historical Data
				D11	Big Data
				D12	No dataset
				D13	Others_DataType

2.2 Edges

For each period p, we constructed a tripartite, undirected, weighted graph on the node sets $\mathcal{M}$, $\mathcal{D}$, and $\mathcal{R}$. We include exactly three edge types: `[:CO_MEASURE_DATATYPE]`,`[:CO_DATATYPE_RQTYPE]` `[:CO_RQTYPE_MEASURE]` linking $(\mathcal{M}, \mathcal{D})$, $(\mathcal{D}, \mathcal{R})$, and $(\mathcal{R}, \mathcal{M})$, respectively. Edges are undirected, and their weights capture period-specific document frequency: for eligible node pairs u, v,

$$w_{uv}^{(p)} = \sum_{d \in \mathcal{C}_p} \mathbf{1}\{u \in d \ \wedge \ v \in d\}, \tag{1}$$

where $\mathcal{C}_p$ is the corpus (set of papers) in period p and $\mathbf{1}\{\cdot\}$ is the indicator function. Thus, multi-edges are collapsed and $w_{uv}^p = w_{vu}^p$ by construction; self-loops are excluded.

Consider two papers in the same period: the first contains $(\mathcal{M}_1, \mathcal{D}_1, \mathcal{R}_1)$ and the second contains $(\mathcal{M}_1, \mathcal{D}_2, \mathcal{R}_1)$. Then the following edges and weights are created:

$$\begin{aligned} &(\mathcal{M}_1, \mathcal{D}_1) \ : \ w = 1, \ (\mathcal{D}_1, \mathcal{R}_1) \ : \ w = 1, \ (\mathcal{R}_1, \mathcal{M}_1) \ : \ w = 2, \\ &(\mathcal{M}_1, \mathcal{D}_2) \ : \ w = 1, \ (\mathcal{D}_2, \mathcal{R}_1) \ : \ w = 1. \end{aligned} \tag{2}$$

The resulting graph is presented in Fig. 2.

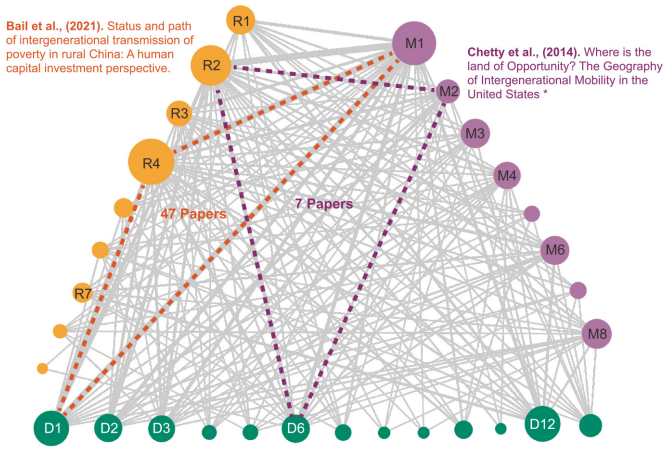


Figure 2: **Representing academic papers at the content-level summarizing all periods.** The knowledge graph is constructed by representing each of 617 papers as a triplet – i.e., each paper consists of a measure, a data type, and a research question type. For example, Bail et al. (2021) paper is represented as $\mathcal{M}_1$-$\mathcal{D}_1$-$\mathcal{R}_4$ and Chetty et al. (2014) paper as $\mathcal{M}_2$-$\mathcal{D}_6$-$\mathcal{R}_2$. Notice that although the used data structure is a multi-period multigraph, this visualization summarizes it across all six periods as a simple network.

2.3 Measures

For each period p, we analyze the tripartite, undirected, weighted graph $G_p = (V_p, E_p)$ on $\mathcal{M}$, $\mathcal{D}$, and $\mathcal{R}$, where V_p comprises these three disjoint node sets and E_p comprises the edges between these nodes. Let $w_{uv}^p \geq 0$ denote the period-specific edge weight between nodes u and v, counting the frequency of the publications, and thus, $w_{uv}^p = 0$ if no edge exists.

The *degree* of a node x in period p is given by

$$k_{x,p} = |\{v \in V_p : w_{xv}^{(p)} > 0\}|, \tag{3}$$

i.e., the number of distinct neighbors connected to x across the two opposite partitions (e.g., if x is a $\mathcal{M}$, x may be linked to different $\mathcal{D}$'s and $\mathcal{R}$'s). A high degree indicates that a node has a large number of neighboring nodes in a given period, denoting the *breadth* of partnering. When calculating the ratio of betweenness to degree, we normalize degree by dividing it by the network's maximum degree.

The *strength* sums the edge weights incident to node x, illustrating the *intensity* of the partnerships as follows

$$s_{x,p} = \sum_{v \in V_p} w_{xv}^{(p)} \ , \tag{4}$$

and it is also known as the weighted degree. In this paper, the weighted degree will be referred to as *strength*, and the unweighted degree as *degree*.

Strength increases when a node appears repeatedly with the same neighboring nodes (its partners), i.e., a high co-occurrence frequency, even if the set of distinct partners is modest. Thus, the degree captures the *breadth* of partnering, whereas the strength captures the *intensity* of those partnerships.

The *edge degree* for an edge, e with two end points u and v is given by

$$k_{e,p} = k_{u,p} + k_{v,p} - 2, \tag{5}$$

where $k_{u,p}$ is the degree of node u in period p and $k_{v,p}$ is node v thereof. This counts the number of partners of the two endpoints, discounting the connection between them. A high edge degree indicates that the edge is connected to many nodes.

The *betweenness centrality* of a node x in period p is given by

$$B_{x,p} \ = \sum_{\substack{s,t \in V_p \\ s \neq x \neq t}} \frac{\sigma_{st}(x)}{\sigma_{st}}, \tag{6}$$

where σ_{st} denotes the total number of shortest paths between nodes s and t, and $\sigma_{st}(x)$ is the number of those paths that pass through x. Nodes with higher $B_{x,p}$ values, therefore, play a more central role in bridging otherwise distant parts of the network.

To account for edge weights, shorter paths were computed using the inverse of the co-occurrence frequency between nodes u and v (i.e., $1/w_{uv}^{(p)}$), such that more frequently co-occurring properties are considered closer in distance. For better comparison, the node betweenness centrality scores are normalized by the total number of distinct unordered pairs, $\frac{2}{(n-1)(n-2)}$ in this paper. Both weighted and unweighted node betweenness centralities are used in our analysis.

The *edge betweenness centrality* of an edge $e = (u, v) \in E_p$ in period p is defined as

$$B_{e,p} = \sum_{\substack{s,t \in V_p \\ s \neq t}} \frac{\sigma_{st}(e)}{\sigma_{st}}, \tag{7}$$

where σ_{st} denotes the total number of shortest paths between node s and t, and $\sigma_{st}(e)$ is the number of those paths that pass through edge e. Edges with higher $B_{e,p}$ values, therefore, act as key bridges between otherwise weakly connected regions of the graph, contributing to the overall structural cohesion of the network. Like the node betweenness centrality, the same inverse weight as in the node betweenness centrality (i.e., $1/w_{uv}^{(p)}$) is used to calculate weighted edge betweenness scores, and the same normalization by $\frac{2}{(n-1)(n-2)}$ is applied in the analysis. Both weighted and unweighted edge betweenness centralities are used in our analysis.

In this paper, we additionally compute three diagnostic ratios: (1) unweighted node betweenness divided by degree, (2) weighted node betweenness divided by strength, and (3) weighted edge betweenness divided by pair count (Everett & Valente, 2016; Wang et al., 2017). With these ratios, we aim to assess the extent to which nodes and edges exert structural influence relative to their local environment. These ratios highlight the cases where a node or edge is disproportionately central given its few neighbors, limited repeated interactions (i.e., paper frequency), or low co-occurrence frequency, as well as the converse patterns. As these ratios quantify the shortest-path flow a node or edge can support relative to its local structural capacity, we refer to them as *capacitance* measures.

For a node x in period p, the *unweighted node capacitance* is defined as

$$R_{x,p} = \frac{B_{x,p}}{\tilde{k}_{x,p}}, \tag{8}$$

where $B_{x,p}$ denotes the unweighted betweenness of node x in period p, and $\tilde{k}_x$ is its degree in the same period, normalized by the corresponding network size. A high unweighted node capacitance indicates that a node bridges a disproportionately large share of shortest paths relative to its number of direct connections. In other words, its effectiveness as a bridge is high. Note that both numerator and denominator are normalized, ensuring that for a star-shaped graph, the central node always has a ratio of 1, regardless of the network size.

The *weighted node capacitance* is defined as

$$R_{x,p}^{(w)} = \frac{B_{x,p}^{(w)}}{\tilde{s}_{x,p}}, \tag{9}$$

where $B_{x,p}^{(w)}$ is the weighted betweenness of a node x in period p, and $\tilde{s}_{x,p}$ is its strength (i.e., weighted degree) in the same period, normalized by the maximum strength in the corresponding periodic network. As strengths are based on weights (i.e., paper frequency), the weighted betweenness is used here. A high-weighted node capacitance implies that a node bridges a large portion of the shortest paths relative to the total empirical weight of its connections or the total number of its connections. As for the unweighted node capacitance, the central node of a star-shaped graph has a strength-wise ratio of 1.

Lastly, for an edge $e = (u, v) \in E_p$ in period p, the *unweighted edge capacitance* is defined as

$$R_{e,p} = \frac{B_{e,p}}{\tilde{k}_{e,p}}, \tag{10}$$

where $B_{e,p}$ denotes the unweighted betweenness of the edge e between nodes u and v in period p, and $\tilde{k}_{e,p}$ is the edge degree, normalized by $2(n-2)$. Pairs with high unweighted edge capacitance exert disproportionately strong bridging influence relative to the number of neighbors this edge has.

2.4 Connectivity of the Network

Before turning to the results, we note that the knowledge graph contains no disconnected subgraphs in any of the six periods or in the aggregated network. All nodes are reachable through at least one path at every time slice. This guarantees that distance-based measures, most notably betweenness centrality and the ratios derived from it, are well-defined and comparable across periods with the normalization.

At T1, 16 nodes with 32 edges for 23 papers; at T2, 18 nodes with 41 edges for 23 papers; at T3, 23 nodes with 64 edges for 65 papers; at T4, 27 nodes with 103 edges for 141 papers at T5, 27 nodes with 120 edges for 183 papers; and at T6, 28 nodes with 123 edges for 182 papers are formed.

2.5 Consistency and Reliability of LLM Outputs

While LLMs offer an efficient means of categorizing a large number of papers, the categories they generate may diverge from the intended classification. We assess three aspects: (1) the reproducibility of LLM-generated taxonomies, (2) the stability of the LLM assignment, and (3) the reliability of the LLM assignment with respect to the human annotators.

First, to assess the reproducibility of LLM-generated taxonomies, 10 independent lists of categories were generated for each of the three taxonomy types (measures, data types, and research question types). One list served as the benchmark ($\mathcal{M}$) and the remaining 9 lists ($\mathcal{M}^1 - \mathcal{M}^9$) were each evaluated using an LLM judge that mapped generated categories to the benchmark in both directions, enabling computation of recall (the fraction of benchmark categories recovered in the generated categories) and precision (the fraction of generated categories that matched a benchmark category). To clarify, the recall checks how many out of, e.g., 8 $\mathcal{M}$ benchmark categories are recovered in the generated categories; the precision checks how many of, e.g., 5 $\mathcal{M}^1$ generated categories are matched to the $\mathcal{M}$ list. Recall that for the benchmark, we have 8 measures, 13 data types, and 9 research question types.

The three primary measure types (Regression-based, Rank-based, Transition Matrix), three primary data types (Panel/Longitudinal Surveys, Administrative/Registry Data, No dataset), and four research question types (Measurement and Methodological Advances, Policy, Institutional, and Geographic Impacts, Intergenerational Wealth Mobility and Inheritance, Demographic Differences in Mobility) were recovered in all 9 lists (recall 100%). Mean recall was 84% for RQ types, 69% for measures, and 53% for data types. Mean precision was high across all taxonomies (RQ types 100%, data types 92%, measures 88%), indicating that independently generated lists rarely produced spurious

Table 2: Within-group vs. between-group abstract similarity comparisons across labeled groups.

Dimension	Within-group	Between-group	Gap	p-value
Measures ($\mathcal{M}$)	0.915	0.909	0.006	< 0.001
Data Types ($\mathcal{D}$)	0.916	0.910	0.006	< 0.001
Research Question Types ($\mathcal{R}$)	0.912	0.910	0.002	0.003

categories without a benchmark counterpart. Mean F1 scores across 9 lists were 91% for RQ types, 77% for measures, and 66% for data types. The lower F1 scores for data types are driven more by recall than by precision, reflecting the larger, more granular benchmark space (13 categories). To validate the LLM judge, a field expert independently annotated benchmark category recovery counts (i.e., recall). Chi-square tests found no significant difference between the expert and LLM judgments for measures (χ^2=6.74, p=0.565) or data types (χ^2=3.73, p=0.988), and the difference for RQ types (χ^2=21.38, p=0.011) was attributable to the LLM assigning more specific labels where the manual annotator more often used 'Others'.

Second, to assess the LLM's stability in category assignment, we compared two sets of LLM outputs. In the first set, categorization was performed using only the paper abstracts, while the second additionally included the paper titles in the input prompt. The two outputs largely overlapped in the assignment of measure categories, but diverged more strongly for data types and, more notably, for research question types (Cohen's $\kappa_{\mathcal{M}} = 0.85, \kappa_{\mathcal{D}} = 0.82, \kappa_{\mathcal{R}} = 0.70$). While this pattern again illustrates that the inherent fuzziness of the task, especially for conceptual classifications, can affect the consistency of the LLM's outputs, a substantial level of agreement is observed in our case.

Lastly, to verify the reliability of the LLM outputs for each paper, two annotators — one more experienced in wealth mobility studies and the other in economics — independently annotated 100 samples of the data, which were selected based on the number of citations. The agreement between the two annotators in terms of Cohen's κ is 0.39 on average (Cohen's $\kappa_{\mathcal{M}} = 0.38, \kappa_{\mathcal{D}} = 0.49, \kappa_{\mathcal{R}} = 0.31$). Annotator 1 was more aligned with the LLM's assignment for *measures*, whereas more similarity was found with annotator 2 for *data types* and *research question types*. We argue that this result does not undermine our experiment. Rather, it highlights the inherent difficulty of the labeling task and the challenge of maintaining consistency across a large corpus of papers, a setting in which LLMs can provide practical support.

2.6 Validation of Category Coherence

Despite the validations on the LLM outputs, the assumption of representativeness of the three labels needs to be validated. To do so, we tested whether papers assigned the same label are more semantically similar to each other than papers assigned different labels. We embedded all 617 abstracts with a scientific document embedding model, SPECTER2 (`allenai/specter2_base`) (Singh et al., 2022) and computed the full 617×617 matrix of pairwise cosine similarities. For each of the three dimensions, we then compared the mean within-group similarity (papers sharing the same label) against the mean between-group similarity (papers with different labels) and assessed significance using a permutation test with 1,000 random label reassignments.

For all three dimensions, papers sharing a label are significantly more semantically similar to each other than to papers with different labels (all $p \leq 0.003$) as shown in Table 2. The observed gaps for $\mathcal{M}$ and $\mathcal{D}$ exceed every one of the 1,000 permuted gaps. The absolute similarity values are uniformly high because the corpus is thematically narrow, as all 617 papers concern intergenerational wealth mobility. However, the pattern of higher within-similarity than between-similarity remains statistically robust relative to this baseline. Since the triplet $(\mathcal{M}, \mathcal{D}, \mathcal{R})$ is the conjunction of three independently validated coordinates, the per-dimension coherence supports the claim that the triplet as a whole is a semantically meaningful representation of a paper.

3 Results

Fig. 3 summarizes the co-occurrence patterns among Measures ($\mathcal{M}$), Data types ($\mathcal{D}$), and Research Question types ($\mathcal{R}$) across six consecutive periods. From T3 (2006-2010) onward, the blue, green, and yellow of the highlighted nodes, pairs, and triangles, representing the five most recurrent components of each type, show their dominance across the panels. This dominance indicates that a small subset of methodological and conceptual combinations (e.g., $\mathcal{M}_1$, $\mathcal{R}_4$, and their pairwise or triadic combinations) increasingly shape the literature. Notably, the popularity of $\mathcal{M}_1$ (*Regression-based Measures*), which begins to grow in T3, continues to increase, making it the only Measures element consistently present among the dominant pairs and triangles.

The black dashed line in each panel traces Kendall's τ_B, that is, Kendall's τ adjusted for ties, between consecutive periods. While rank similarity remains relatively high for nodes, it progressively decreases for pairs and triangles. This decreasing pattern is also stable across 1,000 bootstrap resamples. The 95% bootstrap confidence intervals (CIs) of the global mean for nodes are wider than those for pairs and triangles, although the difference between the pair and triangle CIs is less pronounced. Accordingly, the node popularity remains relatively stable over periods, whereas the popularity of their pairwise and triadic combinations varies more strongly.

It is vital to note that this decrease in temporal stability is not entirely due to randomness or network size. We compared the observed counts with those from 1,000 random models, in which the number of nodes per period is preserved while the three component labels are assigned at random. Part of the decrease in temporal stability (lower $\langle\tau\rangle$) in tandem with the increase in the component number (i.e., $\tau_{\text{node}} > \tau_{\text{pair}} > \tau_{\text{triangle}}$) can be attributed to network size, as the global mean appeared to decrease over the panels with the null models (See Fig. S4). However, the node, pair, and triangle counts have periodic Kendall's τ_B values at similar levels, consistent with their Clopper-Pearson 95% CIs, and the differences in their global means appear smaller than the observed ones. Thus, the consistently high and distinctive similarity in rankings of node counts across periods seems to reflect genuine structural continuity in the literature rather than an artifact of size or randomness. More details on the random null models are given in SM S5.

Together with the stability in node counts, the low τ_B values for pairs and triangles further support the fact that a small set of nodes form the dominant building blocks of higher-order structures. In other words, the dominant nodes, such as *Regression-based Measures* ($\mathcal{M}_1$), *Empirical Estimates and Determinants* ($\mathcal{R}_2$), and *Intergenerational Wealth Mobility*

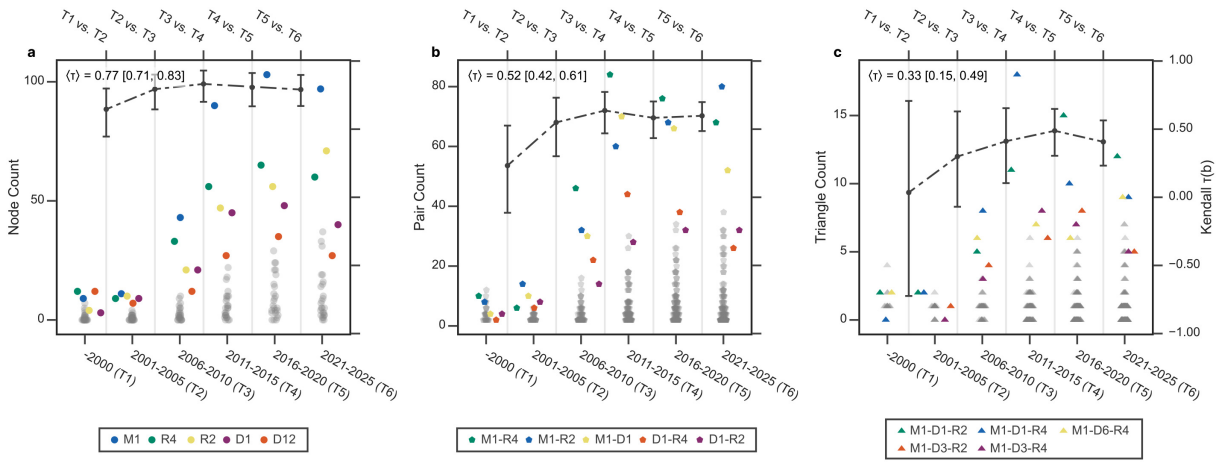


Figure 3: **Popular nodes, pairs, and triangles in the intergenerational wealth mobility literature across six periods.** Each panel summarizes the distribution of components per period. Highlighted markers denote the five most frequent components, presented in the order of frequency. The black dashed line indicates Kendall's τ_B between two consecutive periods, computed over the common set of components. A 95% bootstrap CI based on 1000 resamples is included to assess robustness given the small sample sizes per period. $\langle\tau\rangle$ denotes the mean observed coefficient across periods, with its 95% bootstrap CI reported alongside. **a. Node occurrences (strength/2).** Between T3-T5, highlighted nodes dominate, with an increase in $\mathcal{M}_1$ and $\mathcal{R}_4$ at T3. At T6, the three gray nodes between $\mathcal{D}_1$ and $\mathcal{D}_{12}$ correspond to $\mathcal{D}_6$, $\mathcal{D}_2$, and $\mathcal{M}_2$. Kendall's τ_B ranges between 0.64 and 0.83, indicating relatively high similarity between consecutive periods, and this similarity stays stable as reflected by the narrow bootstrap CIs (all above 0.5). **b. Pairwise co-occurrences.** While $\mathcal{M}_1$-$\mathcal{R}_2$ and $\mathcal{M}_1$-$\mathcal{D}_1$ peak at T4 and decline thereafter, $\mathcal{M}_1$-$\mathcal{R}_2$ continues to increase and become dominant at T5. Kendall's τ_B is lowest between the first two periods (0.23) but stabilizes thereafter, reaching a maximum of 0.64. Rank stability is low in the early comparison, and, the subsequent values remain above zero. **c. Three-way co-occurrences.** $\mathcal{M}_1$-$\mathcal{D}_1$-$\mathcal{R}_4$ peaks at T4 before falling below 10 observations by T6, allowing $\mathcal{M}_1$-$\mathcal{D}_1$-$\mathcal{R}_2$ to emerge as the dominant triangle from T5 onward. The first two periods exhibit minimal rank similarity ($\tau_B = 0.04$). Similarity increases over time, albeit insignificantly, reaching a maximum of $\tau_B = 0.5$, before slightly declining between the last comparison. Overall, stability remains low, particularly in the first three period-to-period transitions.

and Inheritance ($\mathcal{R}_4$), combine with different tripartite components across periods, producing less stable rankings at the pair and triangle level. Lastly, the particularly low coefficients between T1 and T2 for pairs and triangles likely reflect the limited volume of pre-2000 literature, as also evidenced by the wide 95% CIs from 1,000 bootstrap samples.

The normalized weighted node and edge betweenness centralities in Fig. 4 show which nodes and edges serve as bridges in the network. In general, the betweenness scores for edges are remarkably lower than those for nodes, with

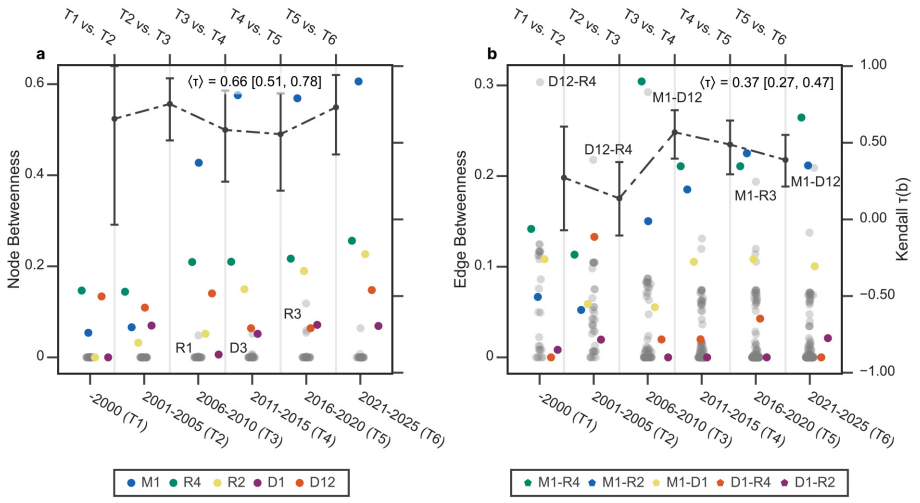


Figure 4: **Structurally central nodes and pairs in the intergenerational wealth mobility literature.** Each panel shows weighted betweenness centrality across periods. The same color scheme and Kendall's τ in the black dashed line as in Fig. 3 are used. Moreover, the same 95% bootstrap CI and the mean observed coefficient across periods, $\langle\tau\rangle$, are calculated. **a. Weighted Node Betweenness.** $\mathcal{M}_1$ emerges as the most prominent bridging node from T3 onwards, suggesting that most papers have employed *Regression-based Measures* ($\mathcal{M}_1$). Although highlighted, $\mathcal{D}_1$ exhibits lower betweenness than $\mathcal{R}_1$ (*Measurement and Methodological Advances*) at T3, and both $\mathcal{D}_1$ and $\mathcal{D}_{12}$ fall below $\mathcal{R}_3$ (*Policy, Institutional, and Geographic Impacts*) at T5. This indicates that high-frequency nodes do not necessarily occupy structurally diverse positions. Kendall's τ_B peaks at the comparison between T2 and T3 (0.75), despite the overlapping bootstrap CIs. All values remain above zero. **b. Weighted Edge betweenness.** The edge betweenness scores are generally lower than node betweenness, and the highlighted ones are no longer distinct from the gray ones. This suggests that pairwise connectivity is less diversified, and that frequently occurring pairs do not always correspond to structurally central edges. At T1 and T2, $\mathcal{D}_{12}$-$\mathcal{R}_4$ (*No dataset-Intergenerational Wealth Mobility and Inheritance*) in gray attains the highest betweenness. At T3, $\mathcal{M}_1$-$\mathcal{D}_{12}$, records the second highest after $\mathcal{M}_1$-$\mathcal{R}_4$; at T5, $\mathcal{M}_1$-$\mathcal{R}_3$ follows the highlighted, $\mathcal{M}_1$-$\mathcal{R}_4$; and at T6, it follows $\mathcal{M}_1$-$\mathcal{R}_2$. $\mathcal{M}_1$-$\mathcal{R}_2$ (*Regression-based Measures-Empirical Estimates and Determinants*) increases steadily and becomes the highest at T5, whereas other periods are dominated by $\mathcal{M}_1$-$\mathcal{R}_4$ (*Regression-based Measures-Intergenerational Wealth Mobility and Inheritance*). The rank similarity between T3 and T4 appears to be highest at the median.

the maximum score below 0.3. This suggests that the same edges may be shared by many triangles, and thus the components that distinguish different communities in the network are nodes rather than edges.

The generally lower edge-level betweenness scores than the node-level scores are also captured by the null models. The highest edge betweenness score lies around 0.047, and the highest node betweenness score lies around 0.08 – both scores being significantly lower than the observed values (see Fig. S5). It can be inferred that, structurally, edge betweenness scores are lower than node betweenness scores, while the significantly higher node and edge betweenness may be due to the characteristics of the literature.

The observed ranking of the betweenness between two consecutive periods seems to vary more than when considering the components' counts in Fig. 3. The highest similarity is observed between T2 and T3 ($\tau_B = 0.75$) for node betweenness, and between T3 and T4 ($\tau_B = 0.57$) for edge betweenness. The generally lower Kendall's τ_B alludes to more changes in pairs' dynamics across the six periods than in the nodes. Note that the random null models also show generally lower Kendall's τ_B values at the edge level, yet to a lesser extent.

Considering the node betweenness in Fig. 4 **a**, $\mathcal{M}_1$ (*Regression-based Measures*) appears to be most diversely connected, while $\mathcal{D}_1$ (*Panel/Longitudinal Surveys*), albeit one of the most frequent nodes, seems to be connected with a limited number of nodes of different partitions, given the node betweenness scores below 0.2 across all periods. Given the remarkably high count of $\mathcal{M}_1$ in T3 in Fig. 3, it can be thought that *Regression-based Measures* ($\mathcal{M}_1$) not only became more popular but also combined with different data types and research question types, serving as a bridge for shortest paths between nodes. Moreover, across periods, only 3 to 7 nodes have node betweenness scores above zero, which is contrary to the fact that more than 7 nodes showed a node count above zero in Fig. 3. These zero-scored nodes in Fig. 4 signify that they have appeared in only one triangle. In other words, they are not connected to other triangles with shared edges, which is further supported by the generally very low edge betweenness in Fig. 4 **b**.

Lastly, combining the insights from the previous two figures, we turn to the capacitance measures in Fig. 5. These capacitance measures highlight components that serve as disproportionately strong bridges relative to their overall connectivity, suggesting potentially underexplored combinations. Note that the figure shows median values, along with the range of ratios across six periods.

The unweighted node capacitance (Eq. 8) in Fig. 5 **a** demonstrates that the five most frequently appearing nodes also exhibit high capacitance, even when disregarding their large number of partnering nodes. This suggests that, within the literature of intergenerational wealth mobility, *Regression-based Measures*, research questions related to *Intergenerational Wealth Mobility and Inheritance*, and *Empirical Estimates and Determinants*, as well as data types such as *Panel/Longitudinal Surveys* and *No dataset*, occupy structurally central positions. When accounting for paper frequency, the weighted node capacitance (Eq. 9) in Fig. 5 **b**, on the other hand, reveals a striking contrast between these highlighted nodes and the rest, with the exception of $\mathcal{D}_3$, which displays a comparatively higher capacitance. Taken together, these results indicate that the most popular nodes act as key bridges between different triplets, in ways that cannot be explained solely by their degrees or strengths. This pattern points to a literature structured around a small set of dominant knowledge components (or paper properties) that recurrently connect otherwise distinct combinations of measures, data types, and research questions.

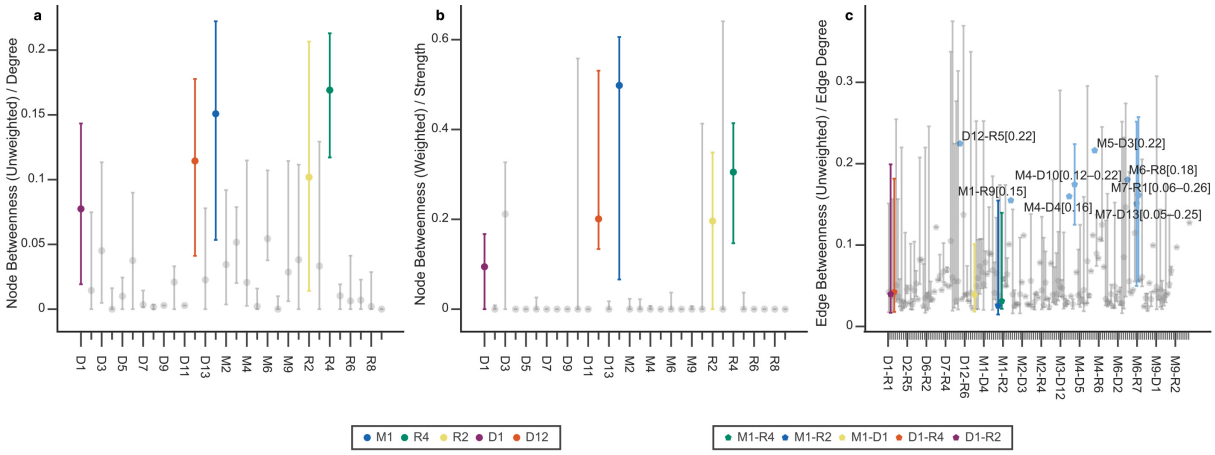


Figure 5: **Median betweenness-degree/count ratios.** Each point shows the period-median ratio for a node or pair. With the same color codes as above, the error bars are the range (min-max) of the ratio across six periods. Those that are not highlighted but have ratios in the 95$^{\text{th}}$ percentile are colored light blue. **a. Unweighted Node Capacitance (Eq. 8).** Given that weights are ignored in calculating degrees, unweighted node betweenness is utilized for the ratio. The highlighted ones, which are the five most frequent nodes in the literature, not only have high degrees but also high unweighted node capacitance. **b. Weighted Node Capacitance (Eq. 9).** Given that weights are considered in calculating strengths, weighted node betweenness is used for the ratio. Among the highlighted nodes, $\mathcal{D}_{12}$ is less frequently used in the literature than $\mathcal{D}_1$ but has a higher weighted node capacitance. **c. Unweighted Edge Capacitance (Eq. 10).** Note that some edges only appear once in the network, hence no error bar. We consider the unweighted edge betweenness and edge degree. At the edge level, the five most popularly used elements have low median efficiency, unlike the other two panels. The highest median capacitance is performed by $\mathcal{D}_{12}$-$\mathcal{R}_5$ (*No dataset–Demographic Differences in Mobility (Race, Gender, etc.)*).

In terms of the unweighted edge capacitance (Eq. 10) in Fig. 5 **c**, the five most frequent pairs are not the ones with the highest capacitance. This means that the most popular pairs do not bridge different edges beyond their connectivity. Yet, it is noticeable that the edges with top 10% capacitance (in light blue) consist of two nodes that differ by at least 8 in their node-degree rankings. The most considerable difference is by $\mathcal{M}_1$-$\mathcal{R}_9$, where the first node has the highest degree, and the other has the lowest degree. The fact that no edges appear to have distinctly high unweighted edge capacitance suggests that the network is well connected, which is likely the case for the small corpus this paper focuses on. Within these edges, we further inspect the top two edges, i.e., $\mathcal{D}_{12}$-$\mathcal{R}_5$ and $\mathcal{M}_5$-$\mathcal{D}_3$. As they appear only once in a periodic network, we examine the impact of adding this edge to the periodic network before their appearance. Note that no changes will be made in the unweighted capacitance when the same triplet is added to the periodic network where the edge already exists.

For $\mathcal{D}_{12}$-$\mathcal{R}_5$, which appear as $\mathcal{M}_3$-$\mathcal{D}_{12}$-$\mathcal{R}_5$ in T3 (2006-2010), the addition of this triplet change the unweighted edge capacitance of 17 edges. Note that, besides $\mathcal{D}_{12}$-$\mathcal{R}_5$, the two other edges of this triplet had already appeared in the previous period. Most of these changes indicate that adding $\mathcal{M}_3$-$\mathcal{D}_{12}$-$\mathcal{R}_5$ reduce edge capacitance across the T2 (2001-2005) network, while four edges – three of which are connected to $\mathcal{D}_{12}$ – exhibit a slight increase in capacitance. In contrast, $\mathcal{M}_5$-$\mathcal{D}_3$ first appear as part of the triplet $\mathcal{M}_5$-$\mathcal{D}_3$-$\mathcal{R}_4$ in T2 (2001-2005), with all three edges making their first appearance in the network. In this case, adding the triplet to the T1 (-2000) network reduces capacitance across all 15 affected edges.

This investigation suggests that edges with high unweighted edge capacitance bridge otherwise weakly connected areas of the network upon their introduction, thereby reducing many edges' capacitance. These two edges do not recur after their first appearance, suggesting that they are forgotten knowledge components in this field, despite their potential to connect different areas of the literature. In this sense, high-capacitance edges such as $\mathcal{M}_5$-$\mathcal{D}_3$ and $\mathcal{D}_{12}$-$\mathcal{R}_5$ could indicate *forgotten combinations* of knowledge components that may warrant qualitative investigation, or highlight essential knowledge linkages that are structurally meaningful within the field. In other words, unweighted edge capacitance reveals hidden combinations of knowledge components that may later become recurrent pathways in the literature or define the knowledge space of the field.

Two additional analyses in SM S4 can be applied. The first examines the relationship between degree and strength, as well as between triangle dispersion ratio (SM S2.4) and the number of unique triplets, thereby opening a discussion of the underlying structural configurations (Fig. S2). The second analysis incorporates the time-decayed appearance of nodes, pairs, and triangles. By allowing weights to decay over years in the case of absence, Fig. S3 can demonstrate when specific components resurge or progressively disappear over time.

4 Discussion

4.1 Interpretation of the field structure

Beyond the descriptive network statistics, the results also offer some reading of the development of the intergenerational wealth mobility literature. First, from T3 (2006–2010) onward, $\mathcal{M}_1$ (*Regression-based Measures*) is both highly frequent and consistently the node with the highest betweenness. This indicates that regression-based measures are not only widely used but also occupy a structurally central position by linking a broad range of data types and research question types. In this sense, regression-based measures, including intergenerational elasticities (Mazumder, 2005; Solon, 2002) and fixed-effects estimates (Chetty & Hendren, 2018), appear to form a persistent methodological backbone of the field.

Second, the change in the dominant triangle from $\mathcal{M}_1$-$\mathcal{D}_1$-$\mathcal{R}_4$ to $\mathcal{M}_1$-$\mathcal{D}_1$-$\mathcal{R}_2$ points toward a shift in the field's most recurrent configuration. A possible interpretation is that inheritance-centered combinations ($\mathcal{R}_4$) became relatively less dominant than combinations centered on empirical estimates and determinants of mobility ($\mathcal{R}_2$), which may be partially

attributable to the relatively limited availability of parent-child linked data. While the present framework cannot identify the causes of this shift, it makes the change in dominant configurations visible.

Third, node rankings are markedly more stable across periods than pair and triangle rankings. This suggests that the field has developed a relatively stable core set of measures, data types, and research question types, while the ways these components are combined continue to evolve. Continuity, therefore, appears more strongly at the level of components than at the level of their pairwise or triadic combinations.

Finally, the capacitance analysis highlights rare combinations whose structural role is disproportionate to their prevalence: $\mathcal{D}_{12}$-$\mathcal{R}_5$ and $\mathcal{M}_5$-$\mathcal{D}_3$. These cases should be interpreted cautiously. The metric, in its current form, does not, by itself, explain why a combination is rare, but it does identify combinations that may warrant closer qualitative investigation.

Taken together, these findings show that the framework not only recovers frequent components, but also reveals persistent structural backbones, changing dominant configurations, and rare but informative combinations in directly interpretable domain vocabulary.

4.2 Limitations

While the present analysis of the case study demonstrates the potential of the LLM-based content-level knowledge graph approach for revealing structural and temporal patterns in a research field, several limitations should be noted.

First, the curated corpus of wealth mobility studies was intentionally narrow and domain-specific, comprising 617 papers. While the small dataset may limit our ability to generalize insights from the literature, this focused selection ensured conceptual consistency across nodes, pairs, and triangles. Moreover, we were able to focus on exploratory analyses in an environment free of any field- or topic-related factors. This way, the present study highlights the potential of our LLM-empowered content-based knowledge graph for knowledge exploration.

Second, the curation process relied on LLM-assisted classification of measures, data types, and research question types using the paper's abstracts rather than the full texts. Although it was a deliberate choice aimed at avoiding systematic bias toward open-access publications, and our approach achieved internal consistency, utilizing the full texts would likely have increased the accuracy of label generation and assignment. Nevertheless, these constraints do not detract from the central contribution of this work, demonstrating how content-based graph modeling can uncover the evolving architecture and methodological trajectories of a scientific field.

Lastly, the framework tested in this paper ensures that each paper is represented by a single element for measures, data types, and research question types. This design choice allows the resulting knowledge graph to remain structurally simple and analytically transparent, but it necessarily abstracts away some within-paper complexity. In particular, measures may be challenging to reduce to a single element, as some papers employ multiple measures or combine them. As a pragmatic solution, we extracted three categories of measures per paper, ordered by priority, and used only the highest-priority category when forming the triplet. Future work could extend the framework to accommodate multiple measures per paper without sacrificing interpretability.

Building on these findings, future research could extend the present framework in several directions. Expanding the corpus beyond wealth mobility studies to incorporate diverse research topics and domains would enable comparative analyses of methodological convergence and divergence across fields, as well as of the influence of one field on another. Here, a more generalizable set of taxonomies can be designed. Moreover, while this paper's approach expressed each paper as a triplet, more than three components can be considered for other applications.

Overall, we have demonstrated that turning citation networks inside out, by focusing on the actual content of scientific papers rather than their citation relations, can yield meaningful insights into the field of intergenerational wealth mobility and open up opportunities for applications in other domains. For instance, the capacitance measures have helped us identify key knowledge components in the literature, as well as the forgotten combinations of knowledge components that bridge otherwise weakly connected parts of the network. All of which was made possible by recent advances in LLMs. More broadly, this approach suggests a shift in how we study and understand the structure of scientific knowledge, moving toward the organization of the ideas themselves.

5 Conclusion

This paper introduced a content-based knowledge graph framework that reconstructs the structure of a scientific field directly from the conceptual categories of its publications, rather than from citation relations. Contrary to keyword co-occurrence approaches, we leverage LLMs to generate and assign interpretable taxonomies and encode each paper as a triplet comprising a measure, a data type, and a research question type. Unlike high-dimensional word-embedding approaches, in which latent dimensions lack semantic grounding, every node in our graph corresponds to a named, domain-validated concept. The network structure is therefore directly readable and actionable for domain researchers without computational expertise.

Applied to 617 studies on intergenerational wealth mobility, the framework surfaces several concrete findings. First, regression-based measures have served as a persistent methodological backbone since 2006-2010, consistently appearing among the most frequent nodes, pairs, and triangles in all subsequent periods. Second, the dominant research configuration shifted from inheritance-focused questions (*Intergenerational Wealth Mobility and Inheritance*, $\mathcal{R}_4$) peaking at T4 (2011-2015), toward empirical estimation of determinants ($\mathcal{R}_2$) dominant from T5 (2016-2020) onward. This transition mirrors the broader empirical turn in the field. Third, the finding that node rankings are considerably more stable across periods than pair and triangle rankings reveals a field that is methodologically conservative but thematically dynamic. In other words, the conceptual vocabulary is frozen, but how those concepts are combined continues to evolve. Finally, the capacitance analysis identifies pairs that briefly bridged otherwise disconnected parts of the network before disappearing, which may be worth examining to determine whether they are underexplored pairs in the field.

More broadly, with this experiment, the paper demonstrates that focusing on the content of academic articles rather than their citation relations can recover meaningful structural signals in a controlled setting, and opens up a complementary perspective on how scientific fields organize and evolve over time.

6 Author Contribution

SK: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing - original draft, Writing - review & editing VH: Methodology, Validation, Writing - original draft, Writing – review & editing. VG: Conceptualization, Funding acquisition, Methodology, Project administration, Supervision, Writing - review & editing

7 Funding

VG acknowledges funding (IBOF-23078) from Vrije Universiteit Brussel (vub.be). The funder did not play any role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

8 Competing Interests

The authors have declared that no competing interests exist.

9 Data Availability Statement

The data and code necessary to reproduce the analyses are available on Zenodo at https://zenodo.org/records/18035391.

References

Aceves, P., & Evans, J. A. (2024). Mobilizing conceptual spaces: How word embedding models can inform measurement and theory within organization science. *Organization Science*, *35*(3), 788–814. https://doi.org/10.1287/orsc.2023.1686

Bail, Y.-l., Zhang, L.-x., Sun, M.-x., & Xu, X.-b. (2021). Status and path of intergenerational transmission of poverty in rural china: A human capital investment perspective. *Journal of Integrative Agriculture*, *20*(4), 1080–1091. https://doi.org/https://doi.org/10.1016/S2095-3119(20)63373-1

Brooks, T. A. (1986). Evidence of complex citer motivations. *Journal of the American Society for Information Science*, *37*(1), 34–36. https://doi.org/10.1002/(SICI)1097-4571(198601)37:1<34::AID-ASI5>3.0.CO;2-0

Chetty, R., & Hendren, N. (2018). The impacts of neighborhoods on intergenerational mobility ii: County-level estimates. *The Quarterly Journal of Economics*, *133*(3), 1163–1228. https://doi.org/10.1093/qje/qjy006

Chetty, R., Hendren, N., Kline, P., & Saez, E. (2014). Where is the land of opportunity? the geography of intergenerational mobility in the united states *. *The Quarterly Journal of Economics*, *129*(4), 1553–1623. https://doi.org/10.1093/qje/qju022

Cozzens, S. (1989). What do citations count? the rhetoric-first model. *Scientometrics*, *15*(5-6), 437–447. https://doi.org/10.1007/bf02017064

Everett, M. G., & Valente, T. W. (2016). Bridging, brokerage and betweenness. *Social Networks*, *44*, 202–208. https://doi.org/10.1016/j.socnet.2015.09.001

Garfield, E. (1955). Citation indexes for science: A new dimension in documentation through association of ideas. *Science*, *122*(3159), 108–111. https://doi.org/10.1126/science.122.3159.108

Grames, E. M., Stillman, A. N., Tingley, M. W., & Elphick, C. S. (2019). An automated approach to identifying search terms for systematic reviews using keyword co-occurrence networks. *Methods in Ecology and Evolution*, *10*(10), 1645–1654. https://doi.org/10.1111/2041-210X.13268

Greenberg, S. A. (2009). How citation distortions create unfounded authority: Analysis of a citation network. *Bmj*, *339*. https://doi.org/10.1136/bmj.b2680

Horbach, S. P., Aagaard, K., & Schneider, J. W. (2021). Meta-research: How problematic citing practices distort science. *MetaArXiv*. https://doi.org/10.31222/osf.io/aqyhg

Lan, J., Li, J., Wang, B., Liu, M., Wu, D., Wang, S., & Qin, B. (2025). Nlp-akg: Few-shot construction of nlp academic knowledge graph based on llm. *arXiv preprint arXiv:2502.14192*.

Lu, Y., & Wang, J. (2025). Karma: Leveraging multi-agent llms for automated knowledge graph enrichment. *arXiv preprint arXiv:2502.06472*.

MacRoberts, M. H., & MacRoberts, B. R. (1989). Problems of citation analysis: A critical review. *Journal of the American Society for Information Science*, *40*(5), 342–349. https://doi.org/10.1002/(SICI)1097-4571(198909)40:5<342::AID-ASI7>3.0.CO;2-U

Mazumder, B. (2005). Fortunate sons: New estimates of intergenerational mobility in the united states using social security earnings data. *The Review of Economics and Statistics*, *87*(2), 235–255. https://doi.org/10.1162/0034653053970249

Moravcsik, M. J., & Murugesan, P. (1975). Some results on the function and quality of citations. *Social studies of science*, *5*(1), 86–92. https://doi.org/10.1177/030631277500500106

Nigel Gilbert, G. (1977). Referencing as persuasion. *Social studies of science*, *7*(1), 113–122. https://doi.org/10.1177/030631277700700112

Price, D. J. D. S. (1965). Networks of scientific papers: The pattern of bibliographic references indicates the nature of the scientific research front. *Science*, *149*(3683), 510–515. https://doi.org/10.1126/science.149.3683.510

Priem, J., Piwowar, H., & Orr, R. (2022). Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833*.

Radhakrishnan, S., Erbis, S., Isaacs, J. A., & Kamarthi, S. (2017). Novel keyword co-occurrence network-based methods to foster systematic reviews of scientific literature. *PloS one*, *12*(3), e0172778. https://doi.org/10.1371/journal.pone.0172778

Radicchi, F., Fortunato, S., & Vespignani, A. (2012). Citation networks. In A. Scharnhorst, K. Börner, & P. van den Besselaar (Eds.), *Models of science dynamics: Encounters between complexity theory and information sciences* (pp. 233–257). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-23068-4_7

Reason, T., Benbow, E., Langham, J., Gimblett, A., Klijn, S. L., & Malcolm, B. (2024). Artificial intelligence to automate network meta-analyses: Four case studies to evaluate the potential application of large language models. *PharmacoEconomics-Open*, *8*(2), 205–220. https://doi.org/10.1007/s41669-024-00476-9

Sampagnaro, G. (2023). Keyword occurrences and journal specialization. *Scientometrics*, *128*(10), 5629–5645. https://doi.org/10.1007/s11192-023-04815-1

Seltmann, S., Kormanyos, E., & Doll, H. C. (2025). Leveraging large language models to extract data citations. *IFC Bulletins chapters*, *64*. https://www.bis.org/ifc/publ/ifcb64_27.pdf

Shibayama, S., Yin, D., & Matsumoto, K. (2021). Measuring novelty in science with word embedding. *PLOS ONE*, *16*(7), 1–16. https://doi.org/10.1371/journal.pone.0254034

Singh, A., D'Arcy, M., Cohan, A., Downey, D., & Feldman, S. (2022). Scirepeval: A multi-format benchmark for scientific document representations. *Conference on Empirical Methods in Natural Language Processing*. https://api.semanticscholar.org/CorpusID:254018137

Solon, G. (2002). Cross-country differences in intergenerational earnings mobility. *Journal of Economic Perspectives*, *16*(3), 59–66. https://doi.org/10.1257/089533002760278712

Tosi, M. D. L., & dos Reis, J. C. (2022). Understanding the evolution of a scientific field by clustering and visualizing knowledge graphs. *Journal of Information Science*, *48*(1), 71–89. https://doi.org/10.1177/0165551520937915

Tosi, M. D. L., & dos Reis, J. C. (2021). Scikgraph: A knowledge graph approach to structure a scientific field. *Journal of Informetrics*, *15*(1), 101109. https://doi.org/10.1016/j.joi.2020.101109

Valenzuela, M., Ha, V., & Etzioni, O. (2015). Identifying meaningful citations. *Workshops at the twenty-ninth AAAI conference on artificial intelligence*. https://cdn.aaai.org/ocs/ws/ws0121/10185-46008-1-PB.pdf

Wang, X. J., Guo, S. Z., Jin, L., & Chen, M. (2017). Cascading failures mechanism based on betweenness-degree ratio distribution with different connecting preferences. *International Journal of Modern Physics C*, *28*(04), 1750052. https://doi.org/10.1142/S0129183117500528

Wu, S., Ma, X., Luo, D., Li, L., Shi, X., Chang, X., Lin, X., Luo, R., Pei, C., Du, C., et al. (2025). Automated literature research and review-generation method based on large language models. *National Science Review*, *12*(6), nwaf169. https://doi.org/10.1093/nsr/nwaf169

Yin, D., Wu, Z., Yokota, K., Matsumoto, K., & Shibayama, S. (2023). Identify novel elements of knowledge with word embedding. *PLOS ONE*, *18*(6), 1–16. https://doi.org/10.1371/journal.pone.0284567

You, T., Yoon, J., Kwon, O.-H., & Jung, W.-S. (2021). Tracing the evolution of physics with a keyword co-occurrence network. *Journal of the Korean Physical Society*, *78*(3), 236–243. https://doi.org/10.1007/s40042-020-00051-5

Yuan, C., Li, G., Kamarthi, S., Jin, X., & Moghaddam, M. (2022). Trends in intelligent manufacturing research: A keyword co-occurrence network based review. *Journal of Intelligent Manufacturing*, *33*(2), 425–439. https://doi.org/10.1007/s10845-021-01885-x

Supplementary Material

S1 LLM Prompting Protocols for Preprocessing

S1.1 Type

System input.

> You're a helpful assistant who can recognize whether this paper url studies intergenerational mobility in wealth, income, or earnings empirically with real-life datasets or theoretically by delving into the models to measure the mobility. This is the abstract of the paper: `[ABSTRACT]`

User input.

> If the study is empirical, return the term "empirical".
> If the study is theoretical, return the term "theoretical".
> If the study is both empirical and theoretical, return the term "both".
> If the study is not at all about intergenerational wealth mobility, return the term: "false".
> "'plaintext

S1.2 Relevance

System input.

> Extract relevant data.

User input.

> Consider this academic paper in the Social Sciences domain:
>
> - Title: `[TITLE]`
> - Year: `[YEAR]`
> - Abstract: `ABSTRACT]`
> - DOI: `[DOI]`
>
> Use this information to answer the following questions. You can try to access the full paper using the DOI provided.:
>
> Q1. Does the paper discuss inter/multigenerational wealth/income/earning mobility? Yes or No.
>
> Q1_1. If 'yes' in Q1, give me a sentence or two about what the paper says about inter/multigenerational wealth/income/earning mobility. If 'no' in Q1, what is the main focus of the paper?

Q2. If 'yes' in Q1, does this paper mention some measures to study inter/multigenerational wealth/income/earning mobility? Yes or No.

Q2_1. If 'yes' in Q2, what are the measures mentioned in the paper? If 'no' in Q2, return 'N/A'.

Q2_2. If 'no' in Q2, give me a reason. If 'yes' in Q2, return 'N/A'.

Q3. If 'yes' in Q2, does this paper empirically study inter/multigenerational wealth/income/earning mobility by applying the measures to some real data? Yes or No.

Q4. If 'No' in Q3, does this paper theoretically study the measures of inter/multigenerational wealth/income/earning mobility? Yes or No.

Answer with 'Yes' or 'No', or if irrelevant, 'N/A', for each main question (Q1, Q2, Q3, Q4). For subquestions (Q1_1, Q2_1, Q2_2), provide a brief explanation. Don't hallucinate.

S2 LLM Prompting Protocols for Labeling

To ensure transparency and reproducibility, we report the prompting protocol used for the LLM-assisted classification procedure. The workflow consisted of two stages. In the first stage, the model generated representative taxonomies from the full corpus of abstracts. In the second stage, each abstract was classified individually using the predefined category sets derived from the taxonomy-generation stage. The full prompt templates are reported below in a platform-independent format. Placeholders such as `[ABSTRACT_BATCH]` and `[ABSTRACT_i]` indicate where the actual abstract text was inserted at inference time. `[ABSTRACT_BATCH]` includes cleaned 742 abstracts.

Unless otherwise noted, the model used for these tasks was `o3-mini`. The prompts were designed to return structured outputs in either JSON format or a constrained schema for downstream analysis.

S2.1 Stage 1: Taxonomy generation from the abstract corpus.

Purpose. To identify representative categories of mobility {measures, datasets, research-questions} from a corpus of abstracts.

System input.

You are a research assistant in the Social Sciences domain. I have a list of abstracts of papers that studied intergenerational or multigenerational wealth, income, or earnings mobility. I want you to categorize the {measures, datasets, research questions} used in these papers based on the abstracts.

User input A1: Measure taxonomy generation.

Abstracts:

`[ABSTRACT_BATCH]`

> First, identify the representative categories of the measures used in the papers based on the abstracts. Here, "measures" refers to an estimate from a model or an equation that quantifies intergenerational or multigenerational wealth, income, or earnings mobility. Then, return the categories and the lists of paper IDs that belong to each category in JSON format. For the IDs that do not belong to any category, return them in a separate category named "Others".

User input A2: Dataset taxonomy generation.

> Abstracts:
>
> `[ABSTRACT_BATCH]`
>
> First, identify the representative categories of datasets used in the papers based on the abstracts. Then, return the categories and the lists of paper IDs that belong to each category in JSON format. For the IDs corresponding to papers without datasets, return them in a separate category named "No dataset". For the IDs that do not belong to any category, return them in a separate category named "Others".

User input A3: Research-question taxonomy generation.

> Abstracts:
>
> `[ABSTRACT_BATCH]`
>
> First, identify the representative categories of research questions that the papers aimed to answer based on the abstracts. Then, return the categories and the lists of paper IDs that belong to each category in JSON format. For the IDs that do not belong to any category, return them in a separate category named "Others".

Expected output. A JSON object mapping category names to lists of paper IDs.

S2.2 Stage 2: Paper-level category assignment.

Purpose. To assign an individual paper abstract to a predefined {measure, dataset, research-question} category.

System input.

> You are a research assistant in the Social Sciences domain. I have an abstract of a paper that studied intergenerational or multigenerational wealth, income, or earnings mobility. I want you to categorize the {measures, datasets, research questions} used in the paper based on the abstract.

User input B1: Measure category assignment.

> Abstract:
>
> `[ABSTRACT_i]`

First, identify the measure or measures used in the paper based on the abstract. Here, "measures" refers to an estimate from a model or an equation that quantifies intergenerational or multigenerational wealth, income, or earnings mobility.

Does it belong to any of the following categories?

1. Regression-based Measures
2. Rank-based Measures
3. Transition Matrix / Probability Measures
4. Absolute Mobility Measures
5. Multigenerational Measures
6. Decomposition / Structural Approaches
7. Non-parametric Approaches
8. Others

Return the category as a number, and if possible, also return the exact name of the measure. For example, if it belongs to category 1, also return the measure name such as "Intergenerational Elasticity (IGE)". Do not repeat the category label in the returned measure name. If the measure does not belong to any category, return category 8 and the exact name of the measure.

User input B2: Dataset category assignment.

Abstract:

`[ABSTRACT_i]`

First, identify the type or types of datasets used in the paper based on the abstract.

Does it belong to any of the following categories?

1. Panel/Longitudinal Surveys
2. Administrative/Registry Data
3. National Survey Data
4. Opportunity Atlas
5. Natural/Experimental Data
6. Linked Administrative Data
7. International Panel Data
8. Rich List Data
9. University/Institution Data
10. Pseudo-Panel/Household Budget Survey
11. Archival/Historical Data

12. Big Data
13. No dataset
14. Others

Return the category as a number, and if possible, also return the exact dataset name. For example, if it belongs to category 1, also return the dataset name. Do not repeat the category label in the returned dataset name. If the dataset does not belong to any category, return category 14 and the exact dataset name. If the paper does not use any dataset, return category 13 and `N/A` as the dataset name.

User input B3: Research-question category assignment.

Abstract:

`[ABSTRACT_i]`

First, identify the type of research question that the paper aimed to answer based on the abstract.

Does it belong to any of the following categories?

1. Measurement and Methodological Advances
2. Empirical Estimates and Determinants
3. Policy, Institutional, and Geographic Impacts
4. Intergenerational Wealth Mobility and Inheritance
5. Demographic Differences in Mobility (Race, Gender, etc.)
6. Mobility and Non-Income Outcomes (Health, Wellbeing, etc.)
7. Theoretical and Structural Models
8. Perceptions of Mobility and Attitudes
9. Others

Return the category as a number, and if possible, also return the exact research question. If the research question does not belong to any category, return category 9 and the exact research question.

Structured output schema.

```
{
  "category": <integer>,
  "research_question": "<string>"
}
```

Implementation note. In the taxonomy-generation stage, the model was prompted with batches of abstracts and asked to induce representative category sets. In the paper-level assignment stage, each abstract was processed individually using the fixed category definitions shown above. This two-stage design separated category discovery from category assignment and allowed the final coding protocol to be applied consistently across papers.

S2.3 Extra: Multiple category assignment.

Considering that one paper may belong to multiple categories, we also tried assigning papers to multiple measure categories as a test. Note that the following prompt sometimes yielded multiple entries as the second category.

> The following is the description of the measure(s) used in the paper.
>
> “‘json
>
> cat
>
> “‘
>
> This was the categorization used for the measures previously.
>
> 1. Regression-based Measures
> 2. Rank-based Measures
> 3. Transition Matrix / Probability Measures
> 4. Absolute Mobility Measures
> 5. Multigenerational Measures
> 6. Decomposition / Structural Approaches
> 7. Non-parametric Approaches
> 8. Others
>
> Now, based on the description of the measure(s), categorize them further.
> ’Unspecified’: If the description is not clear enough to categorize, return ’0’. Otherwise, ’1’.
> ’Relevance’: If the measure(s) are not used to study inter/multi-generational wealth/income/earning mobility, return ’0’. Otherwise, ’1’.
> ’Category_2’: If ’1’ in Q1 and there is more than one type of measure mentioned, return the category that the measure(s) belong to as a number. Otherwise, return empty.
> ’Measure_1’: If ’1’ in Q1, give me the keyword that best describes the measure(s). If there are multiple measures, return the most relevant one.
> ’Measure_2’: If ’1’ in Q1 and there is more than one type of measure mentioned, return the second most relevant keyword. Otherwise, return empty.
> Return the results in a JSON format for each id. Each id should have [’Unspecified’, ’Relevance’, ’Category_2’, ’Measure_1’, ’Measure_2’].
>
> Make sure to be consistent with the keywords. For example, if you gave ’Intergenerational Elasticity’ for id ’1’, give the same value to id ’5’ whose measure is described as ’IGE’ or ’IGE in income’.

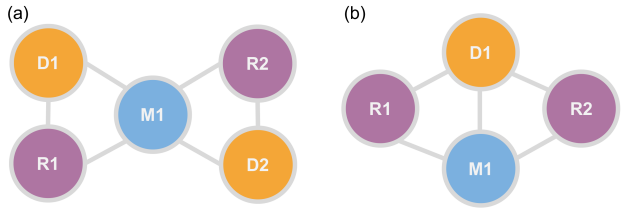
The empirical distribution of how often this multi-category facility was actually invoked is shown in the table below.

Across the 617 papers	Count	%
Only one measure category assigned	371	60.1%
Secondary measure same category as primary	164	26.6%
Secondary measure different category from primary	82	13.3%
Tertiary measure also assigned	7	1.1%

This shows that about 87% of the papers can be uniquely described by the primary measure.

S2.4 Additional Analytical Tools

The following are the additional analytical tools that can be applied to larger corpora.



Triangle Dispersion Ratio. Another question that can be asked about the graph is which nodes comprise triangles using similar edges and distinct edges, which we will call the *triangle dispersion ratio*:

$$\eta_{x,p} = \frac{k_{x,p}}{2T_{x,p}}, \tag{11}$$

where $T_{x,p}$ is the number of unique triplets (m, d, r) in period p such that x is one of m, d, r, and $k_{x,p}$ is its degree, i.e., the number of distinct partners x is connected to via the two opposite partitions in period p. As can be seen in the Figure above, if $x = \mathcal{M}_1$, two distinct triplets can either contribute to four edges around x (panel (a), the triplets do not have a shared edge), or to three edges around x (panel (b), the triplets have one shared edge in common). So, $2T_{x,p}$ is an upper bound for the degree $k_{x,p}$. Thus, the *triangle dispersion ratio* is bounded by 1, with $\eta_{x,p} = 1$ if and only if x consists of only distinct triangles where no edges are shared.

The higher the η is, the more connected the node x is to different nodes, with more distinct triangles, like in panel (a) in the illustration above, whereas the lower the η, the more often the edges connected with x are reused to create different triads, like in panel (b). In the case of high $T_{x,p}$, a node with a higher η can be considered a mixer node, one that is frequently used and broadly combined across the two partitions. In contrast, a node with a lower η is typically a node that is frequently used but mainly with a narrow, reused partner set. In the case of low $T_{x,p}$, a similar interpretation can be retrieved, but with the limitation that x participates in a few realized three-way contexts.

Time-decayed Appearance Weight. To capture the temporal persistence of key components – nodes, pairs, and triangles – we define a time-decayed weight δw_t that integrates past appearances with an exponential forgetting factor.

For each component (node, pair, or triangle) and each year t, the decayed weight evolves recursively as

$$\delta w_{t+1} = \begin{cases} \delta w_t \, e^{-\lambda} + c_{t+1}, & \text{if the component appears at year } t{+}1, \\ \delta w_t \, e^{-\lambda}, & \text{otherwise,} \end{cases} \tag{12}$$

where $\lambda > 0$ is the decay rate and c_{t+1} denotes the count of papers in which the component appears during year $t + 1$. The initial condition is

$$\delta w_0 = \begin{cases} 1, & \text{if the component first appears at year 0,} \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

This formula ensures that recent appearances contribute more strongly to the weight, while earlier observations decay exponentially over time:

$$\delta w_T = \sum_{t=0}^{T} c_t \, e^{-\lambda(T-t)}. \tag{14}$$

The parameter $\lambda = \ln(2)/5$ corresponds to a half-life of five years, implying that the contribution of a past appearance is halved after five years of inactivity. Note that we consider all years between the minimum and maximum years in the dataset (i.e., 1976-2025), and that the decay is applied annually (i.e., $T - t = 1$).

In the analysis, we examine two variants of time-decayed appearance weight: one where c_{t+1} is the count of papers (Eq.12) and another where $c_{t+1} = 1$ if it (re)appears in year $t + 1$.

S3 Unweighted Betweenness Centrality

Here, we also present the unweighted betweenness centrality at both the node and edge levels.

S4 Additional Analyses' Results

This section introduces the results of the additional analyses that can be applied to larger corpora. The conclusions drawn in the following subsections should not be generalized, given the limited size of our dataset. Nevertheless, we present these results to illustrate the kinds of analytical tools enabled by our content-based knowledge-graph approach.

S4.1 Degree, Strength, and Triangle Dispersion Relationships

Fig. S2 summarizes the relationships between strength and degree, and between the triangle dispersion ratio and the number of unique triplets. It further supports that the higher the degree – hence, the more diverse the node is connected to – the more popular it is (i.e., the higher the strength). Moreover, as the number of unique triplets increases for a node, the number of edges shared to form triangles increases. Understood with Fig. 4, this may hint that a few dominant nodes (*e.g.*, $\mathcal{M}_1, \mathcal{R}_4, \mathcal{R}_2$) form dominant triangles with many shared edges. In other words, when studying wealth mobility

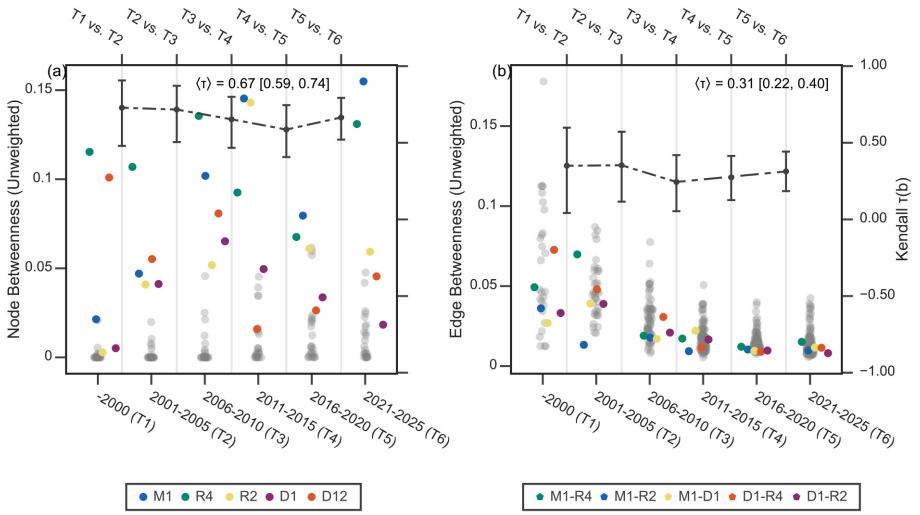


Figure S1: **Unweighted Betweenness Centrality.** The same color scheme and Kendall's τ_B calculation as in previous figures in the main text are used in this figure. Overall, unlike in weighted betweenness centrality, unweighted betweenness centrality does not show a large difference between the node and edge levels on their y-axes. **a. Unweighted Node Betweenness.** In the first three periods, $\mathcal{R}_4$ plays the most central node while in the last three periods, it is $\mathcal{M}_1$. Unlike the weighted betweenness in Fig. 4, $\mathcal{R}_2$ also plays a crucial role in T4, and the gap between the most central node and the rest is not as big. The rankings of the betweenness appear to be strongly correlated across periods. **b. Unweighted Edge Betweenness.** It is more obvious than in the weighted edge betweenness in Fig. 4 that the highlighted ones are not necessarily the ones with the high edge betweenness. Especially, in the last three periods, it appears that the popular pairs are not the ones that bridge areas of the network. The rankings of betweenness appear to be moderately similar across periods.

and its methodology, most literature applies regression-based measures since the period of 2006-2010, on different types of data, often in order to answer research questions about intergenerational wealth mobility and inheritance.

S4.2 Resurgence

Fig. S3 illustrates the resurgence patterns of key components based on two types of time-decayed appearance weights: one without considering paper frequencies (top) and one incorporating them (bottom). In both cases, it becomes evident that $\mathcal{M}_1$, previously identified as the most dominant node in terms of frequency and betweenness centrality, is not the most persistently reappearing feature throughout 1976-2025. Instead, $\mathcal{R}_4$, which first appeared in 1976, maintains a continuous presence from 1997 onward and emerges as the most persistent node by 2000. When the number of papers is taken into account, $\mathcal{D}_{11}$ exhibits a sharp decline between 1980 and 1994, suggesting a temporary shift toward other data sources, followed by a dramatic resurgence around 1998, reaching nearly twice the maximum weight of all other nodes

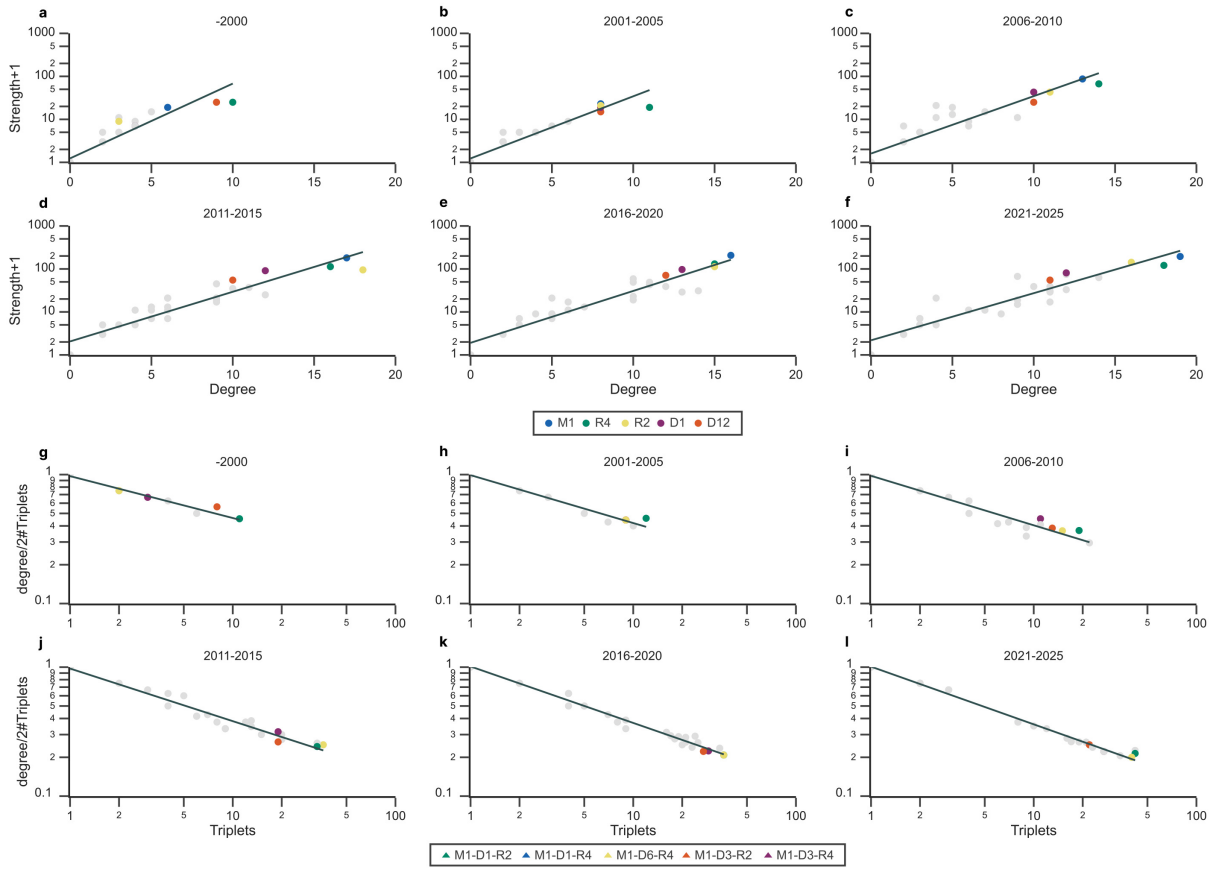


Figure S2: **Relationship between node degree and strength, and between triangle dispersion ratios and three-way co-occurrence patterns across periods.** Panels (a-f) show node degree versus strength $(+1)$, and panels (g-l) show Triangle Dispersion Ratio versus the number of distinct triangle co-occurrences for each period. Y-axes are logarithmically scaled to highlight proportional changes across orders of magnitude. The same coloring scheme as in Fig. 3 is applied. **a-f.** A clear positive association between degree and strength is observed: the higher a node's degree, the greater its strength, indicating that paper features connected to a wider range of other features also tend to co-occur more frequently in the literature. **g-l.** A clear negative association emerges between the Triangle Dispersion Ratio and the number of unique triplets: nodes involved in many triplets tend to reuse the same edges across them, forming natural clusters of densely interconnected features.

and remaining dominant thereafter. These patterns highlight that the most popular component in terms of frequency is not necessarily the one that appears most consistently across time.

When examining the pairs, particularly those incorporating paper counts, a clear divergence emerges for $\mathcal{M}_1$-$\mathcal{R}_4$, $\mathcal{M}_1$-$\mathcal{R}_2$, and $\mathcal{M}_1$-$\mathcal{D}_1$, which separates markedly from the rest. This indicates that these three pairs are not only the most frequent across the six periods, but also persistent over time. Finally, the two panels for triangles are identical, as

no year contains more than one paper representing the same triadic combination. In other words, within the wealth mobility literature, no two papers in the same year share exactly the same measure, data type, and research-question type.

S5 Robustness Check

S5.1 The Network Size Effect on the Measures

Given the small sample size in each period, we generated a random null model to assess whether the observed patterns in the counts and betweenness measures could be explained by network size. The randomization procedure preserves only the number of nodes per period while randomly assigning the three component labels. The core tripartite structure of the graph is maintained by ensuring that each paper receives one element of $\mathcal{M}, \mathcal{D}$, and $\mathcal{R}$. We then compared the observed Kendall's τ_B values across periods against the null node counts, pair counts, triangle counts, node betweenness, and edge betweenness. The number of null samples was determined adaptively using a sequential Monte Carlo procedure that continued until the Clopper-Pearson 95% CI narrowed below 0.01 or until a maximum of 250,000 samples was reached.

At the node level, all measures across adjacent periods were significantly different from the null Kendall's τ_B values at 1000 samples. At the edge level, pair counts and edge betweenness scores (both weighted and unweighted) did not significantly differ from the null Kendall's τ_B for the first two comparisons and reached the Clopper-Pearson 95% CI below 0.01 before reaching 1000 samples. For the triangle counts, their first three comparisons are not significantly different from the observed Kendall's τ_B, and Kendall's τ_B could not be computed.

Fig. S4 demonstrates the results of the node, pair, and triangle counts with the 1000 null samples. The high global mean of Kendall's τ_B observed for node counts in Fig. 3 contrasts with the corresponding random model. Despite the wider 95% CIs across periods, the global mean for pair counts (0.03) is substantially lower than the observed (0.52), with a bootstrap 95% CI $(0.01\text{-}0.04)$ which lies below the observed CI $(0.42\text{-}0.61)$. For the triangle counts, the global means cannot be directly compared since Kendall's τ_B values for the first two null comparisons were undefined due to insufficient overlap under the null model. However, the remaining null comparisons cluster around zero, whereas the observed values are consistently positive.

Overall, comparing the observed and null results suggests that part of the decrease in temporal stability (i.e,. lower $\langle\tau\rangle$) in tandem with the increase in the component number (i.e., $\tau_{\text{node}} > \tau_{\text{pair}} > \tau_{\text{triangle}}$) can be attributed to network size. Nevertheless, the consistently high similarity in node counts across periods appears to reflect genuine structural continuity in the literature rather than an artifact of size or randomness.

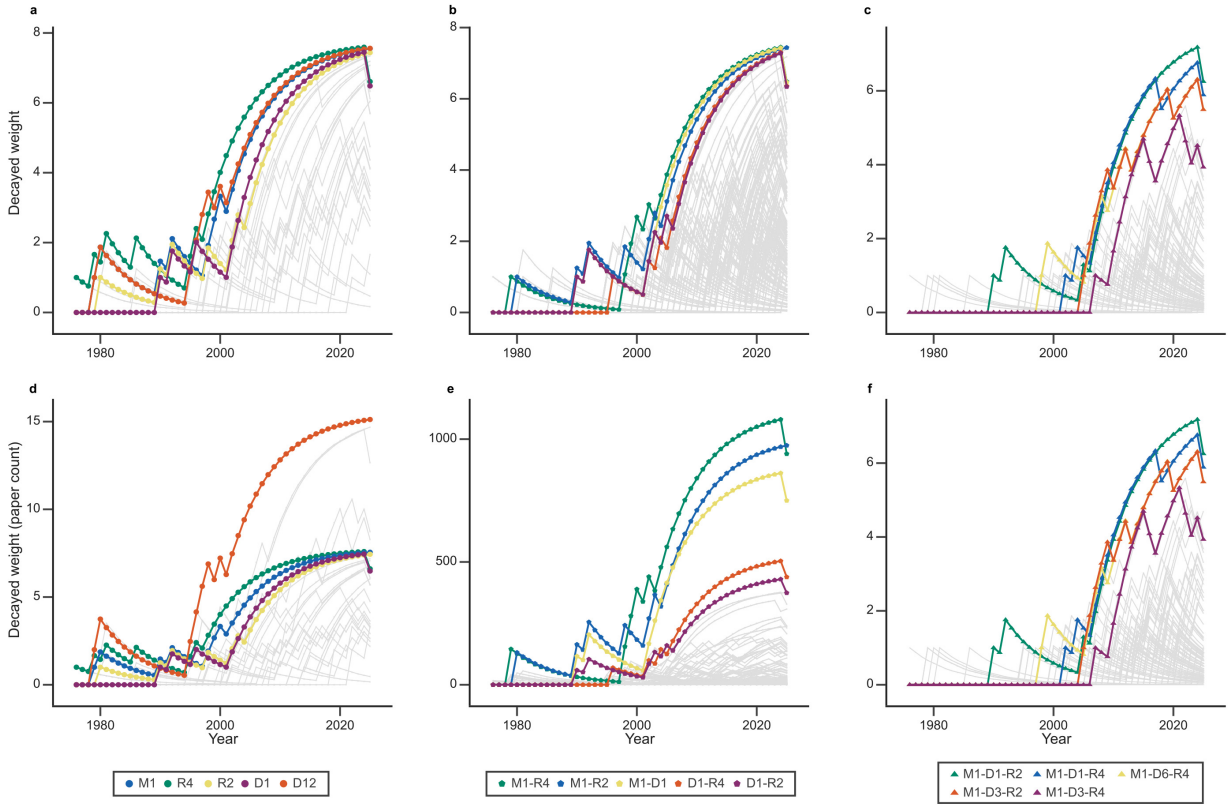


Figure S3: **Time-decayed appearance of nodes, pairs, and triangles with and without weighting by publication frequency over years.** The panels demonstrate the (re)appearance of nodes, pairs, and triangles by applying the time-decayed appearance weights (see S2.4). Panels (a-c) only consider whether a key component appeared in a given year, while panels (d-f) additionally account for the number of papers containing that component. **a&d.** When considering only the net appearance of nodes, $\mathcal{R}_4$ remains the dominant node over the years, whereas $\mathcal{D}_{11}$ becomes dominant once paper frequencies are incorporated. In both panels, $\mathcal{R}_4$ makes a sharp decline after peaking in 1980, followed by a pronounced resurgence after 1994. **b&e.** In both graphs, $\mathcal{M}_1$-$\mathcal{R}_4$ evolves into the dominant edge after a turning point in the year 1997, closely followed by $\mathcal{M}_1$-$\mathcal{D}_1$ and $\mathcal{M}_1$-$\mathcal{R}_2$. Given paper frequencies, the gap between these three widens. $\mathcal{M}_1$-$\mathcal{R}_2$ shows more frequent appearances in later years than $\mathcal{M}_1$-$\mathcal{D}_1$. Distinct shifts are observed in 2001 for $\mathcal{M}_1$-$\mathcal{D}_1$, in 1997 for $\mathcal{M}_1$-$\mathcal{R}_4$, and in 1989 for $\mathcal{M}_1$-$\mathcal{R}_2$, with the latter showing the most abrupt evolution overall. **c&f.**The two graphs are identical, as no papers from the same year share the same triplets. $\mathcal{M}_1$-$\mathcal{D}_3$-$\mathcal{R}_4$ appears latest among the highlighted triangles, whereas $\mathcal{M}_1$-$\mathcal{D}_1$-$\mathcal{R}_2$ emerges first in 1990. Compared to the nodes, where the gray traces closely follow the highlighted ones, edges and triangles (e,f) reveal stronger concentration, indicating that most papers focus on a limited set of dominant pairwise and triadic structures.

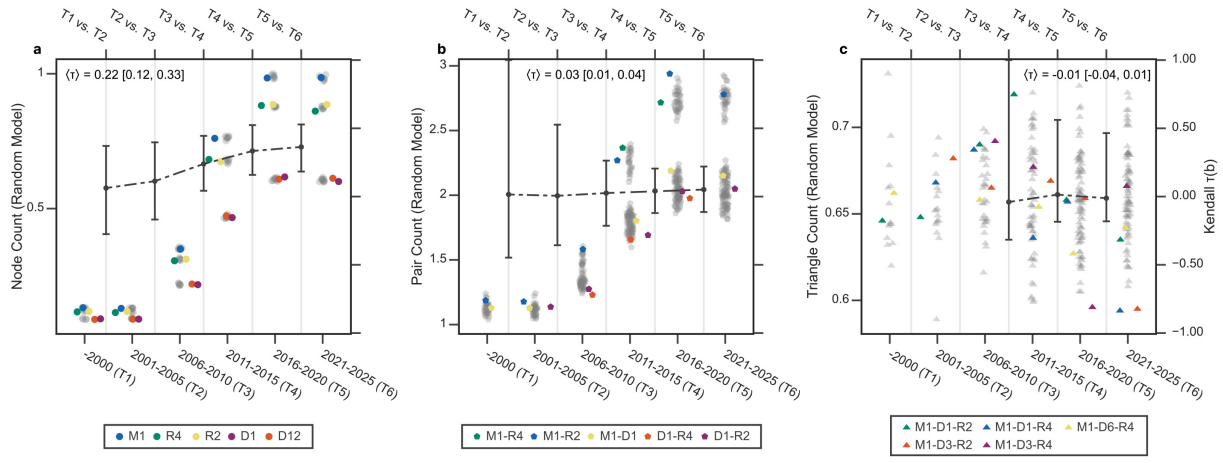


Figure S4: **Random null model for co-occurrences of components in the literature.** With the 1000 null samples, the node counts, pair counts, and triangle counts are plotted over six periods. At each period, the mean Kendall's τ_B over 1000 samples is plotted along with the Clopper-Pearson 95% CIs. $\langle\tau\rangle$ is the global mean of the sampled Kendall's τ_B, together with the bootstrap CI. Here, the bootstrap CI is used to achieve coherence with Fig. 3. Overall, across the panels, null distributions are significantly narrower and concentrated at substantially lower values than the observed ones. **a. Node counts.** In general, the period-wise mean Kendall's τ_B's and their corresponding CIs are lower than the observed ones in Fig. 3 **a**. $\mathcal{M}_1$ remains as one of the leading nodes as of T3, as in the observed result, but together with other non-highlighted nodes. **b. Pair counts.** Still, $\mathcal{M}_1$-$\mathcal{R}_4$ and $\mathcal{M}_1$-$\mathcal{R}_2$ appeared in the leading group over periods. However, the green, $\mathcal{M}_1$-$\mathcal{R}_4$, is no longer one of the leading nodes in every period. A flat trend is observed in the Monte Carlo procedure that continued until the Clopper-Pearson 95% CI narrowed below 0.01 or until a maximum of 250,000 samples were reached. The Kendall's τ_B values are concentrated around zero over periods, while the observed values were above zero in Fig. 3 **b**. **c. Triangle counts.** The first two comparisons had undefined τ_B values due to insufficient overlap under the null model. Compared to the observed distribution in Fig. 3 **c**, the null triangle counts are widely dispersed, and the dominance of the highlighted triangles has disappeared.

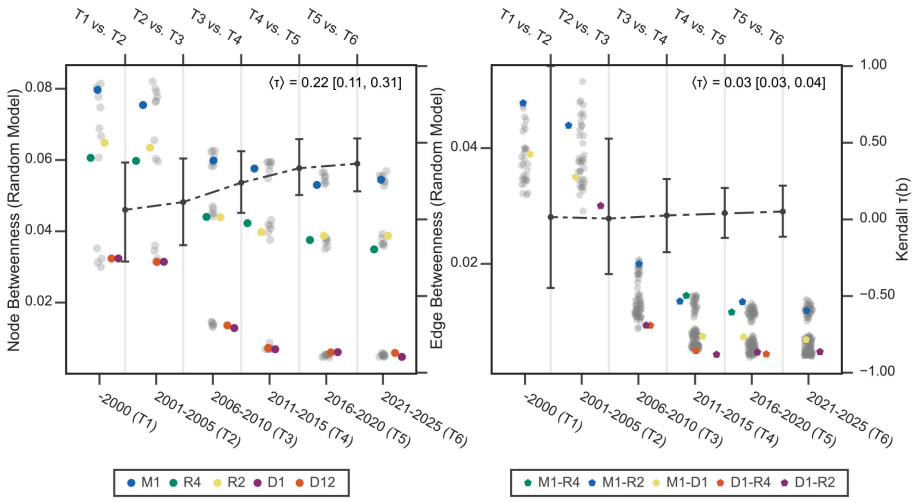


Figure S5: **Random null model for betweenness of components in the literature.** With the 1000 null samples, node and edge betweenness centrality scores are plotted over six periods. At each period, the mean Kendall's τ_B over 1000 samples is plotted, along with Clopper-Pearson 95% CIs. $\langle\tau\rangle$ is the global mean of the sampled Kendall's τ_B, together with the bootstrap CI. Here, the bootstrap CI is used to achieve coherence with Fig. 4. Overall, at both node and edge levels, null distributions are concentrated at substantially lower values than the observed ones. **a. Node betweenness.** In general, Kendall's τ_B stays between 0 and 0.5, and the global mean is much lower than in the observed one in Fig. 4. **b. Edge betweenness.** The first comparison shows a large variability across the null models, while it is narrower with the comparisons with later periods. Overall, the median τ_B value stays around zero across comparisons, while the observed ones ranged between 0.2 and 0.6.

S5.2 The Effect of Raw Pair Counts on the Measures

Additionally, we compare the observed and expected co-occurrence matrices to validate that the values are independent of the raw pair counts. We computed observed-vs-expected co-occurrence matrices for the three bipartite panes ($\mathcal{M} \times \mathcal{D}, \mathcal{D} \times \mathcal{R}, \mathcal{R} \times \mathcal{M}$), using the independence expectation, $E_{uv} = n \cdot P(u) \cdot P(v)$. We consider the ratio between the two:

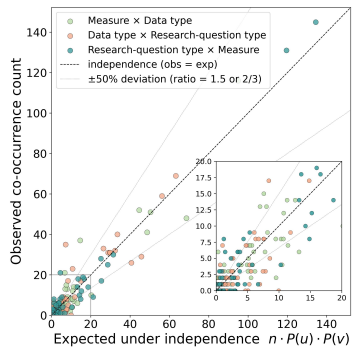
$$R_{uv} = w_{uv}/E_{uv}, \tag{15}$$

where w_{uv} is the observed counts of the edge connecting u and v nodes, and E_{uv} is the expected thereof. $R = 1$ denotes exact independence; $R > 1$ denotes over-representation beyond marginals; $R < 1$ denotes avoidance.

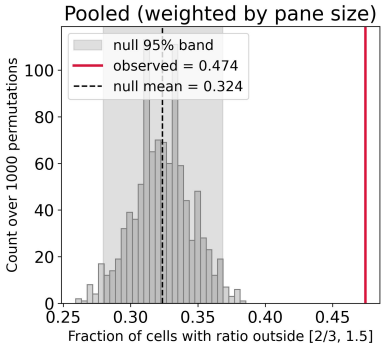
Fig. S6a below shows that the marginals are not a complete predictor for the observed counts. Pooling all 617 papers, we see that nearly half of all non-empty pair cells deviate substantively from independence. Across the three panes,

47.4% of cells (weighted by pane size, $N = 293$ non-empty cells) have observed/expected ratios outside $[2/3, 1.5]$. In other words, more than 50% away from the independence prediction. The per-pane fractions are 47.1%, 44.4%, and 52.8%, respectively. This is the structural signal that the paper's pair- and triangle-level analyses capture.

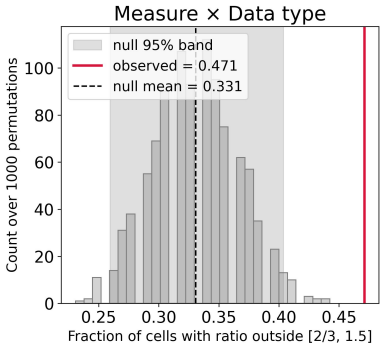
To check that this 47.4% is not itself produced by sampling noise in sparse cells, we built a null distribution by permuting the three label columns independently across the 617 papers and recomputing the deviant fraction over 1000 permutations. Under independence, the null mean is 32.4% (95% band [28.0%, 36.9%]). The observed 47.4% lies above the band in every pane (one-sided $p \leq 0.002$), and the same pattern holds per pane (Fig. S6b-e).



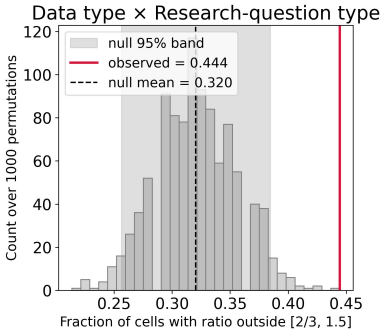
(a) Observed versus expected counts.



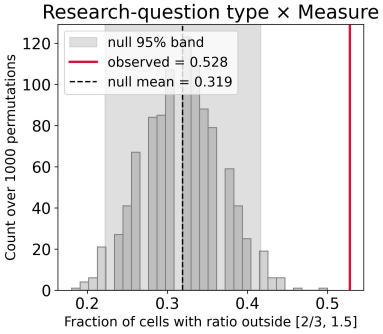
(b) Pooled panels.



(c) Measure × data type.



(d) Data type × research-question type.



(e) Research-question type × measure.

Figure S6: Observed versus expected co-occurrence patterns. Panel a compares observed and expected co-occurrence counts under independence, with the inset zooming into $[20, 20]$. Panels b-e show, for the observed data and 100 permuted datasets, the fraction of cells with observed-to-expected ratios outside $[2/3, 1.5]$.