Scito2M: A 2 Million, 30-Year Cross-disciplinary Dataset for Temporal Scientometric Analysis

Anonymous submission

Abstract

Understanding the creation, evolution, and dissemination of scientific knowledge is crucial for bridging diverse subject areas and addressing complex global challenges such as pandemics, climate change, and ethical AI. Scientometrics, the quantitative and qualitative study of scientific literature, provides valuable insights into these processes. We introduce *Scito2M*, a longitudinal scientometric dataset with over two million academic publications, providing comprehensive contents information and citation graphs to support cross-disciplinary analyses. Using *Scito2M*, we conduct a temporal study spanning over 30 years to explore key questions in scientometrics: the evolution of academic terminology, citation patterns, and interdisciplinary knowledge exchange. Our findings reveal critical insights, such as disparities in epistemic cultures, knowledge production modes, and citation practices. For example, rapidly developing, application-driven fields like LLMs exhibit significantly shorter citation age (2.48 years) compared to traditional theoretical disciplines like oral history (9.71 years). Our code and data are available at https://anonymous.4open.science/r/Scito2M/.

Introduction

Scientific advances are crucial for addressing global challenges such as pandemics, energy security, climate change, social justice, and ethical AI. Tackling these issues requires a holistic understanding of how scientific knowledge evolves across disciplines. In this context, *scientometrics*–the qualitative and quantitative study of scientific literature–plays a pivotal role in uncovering the structure, dynamics, and evolution of research across fields (Donthu et al. 2021). By analyzing publication contents and citation networks, scientometrics offers insights into key topics, trends, and scholars, providing valuable perspectives for researchers and policymakers in decision-making in scientific advances and solving global challenges.

Challenges. Current scientometric studies face two major challenges: *1) Lack of Comprehensive Longitudinal Datasets.* While scientometric datasets are available, there is a scarcity of large-scale, longitudinal datasets that combine both content-level and citation-level information across multiple disciplines. *2) Limited Analytical Scope.* Despite extensive research in scientometrics–scrutinizing the creation (Gu, Meng, and Farrukh 2021), diffusion (Radev and Abu-Jbara 2012), and association (Leto et al. 2024) of academic knowledge–many studies focus on limited timespans (Koch et al. 2021; Zhang et al. 2022), venues (Ciotti et al. 2016; Jin et al. 2024b), or particular areas like natural language processing (Radev and Abu-Jbara 2012; Singh et al. 2023; Nguyen and Eger 2024) and human-computer interaction (Oppenlaender 2024). These gaps impedes a comprehensive understanding of critical issues, such as the *breadth* (topical diversity) and *depth* (long-term impact) of scientific knowledge exchange.

This Work. We present *Scito2M*, a large-scale **Sci**en**to**metric dataset comprising over **2 million** academic literature from arXiv[1], encompassing 30 years since its inception in 1991. *Scito2M* offers detailed metadata such as titles, abstracts, full-text[2], keywords, subject categories, and a comprehensive citation graph, making it a valuable resource for comprehensive scientometric analyses. *Scito2M* supports detailed, longitudinal analysis of scientific knowledge evolution and citation patterns, contributing to our understanding of how interdisciplinary research contributes to solving global challenges. Using this dataset, we conduct content and citation analyses to investigate scientific influence over 3 decades. Our key findings are:

- **Paradigm Shifts** (Shapere 1964). Scientific progress occurs through periodic *leaps* rather than linear knowledge accumulation. Recent paradigm shifts have shown a noticeable change from theoretical to applied research.
- **Terminology Prominence.** Machine learning-related terms have seen a marked rise in prominence, accounting for an average of only 0.31 words of the top 20 annual terms prior to 2010, but surging to 9.5 words from 2015.
- **Disciplinary Homophily** (Zhang et al. 2018). Citation networks display a strong tendency towards homophily, with intra-disciplinary citation accounting for over 91% of all citations.
- **Epistemic Cultures** (Cetina 2007). Different fields exhibit unique patterns in the production, validation, and citation of knowledge. Compared to applied research, basic research places greater emphasis on intra-disciplinary citations to maintain academic rigor and coherence.
- **Citation Amnesia** (Singh et al. 2023). Applied research

[1] https://arxiv.org/.

[2] To comply with Term of Usage for arXiv[3], we provide downloadable links instead of PDFs of paper e-prints.

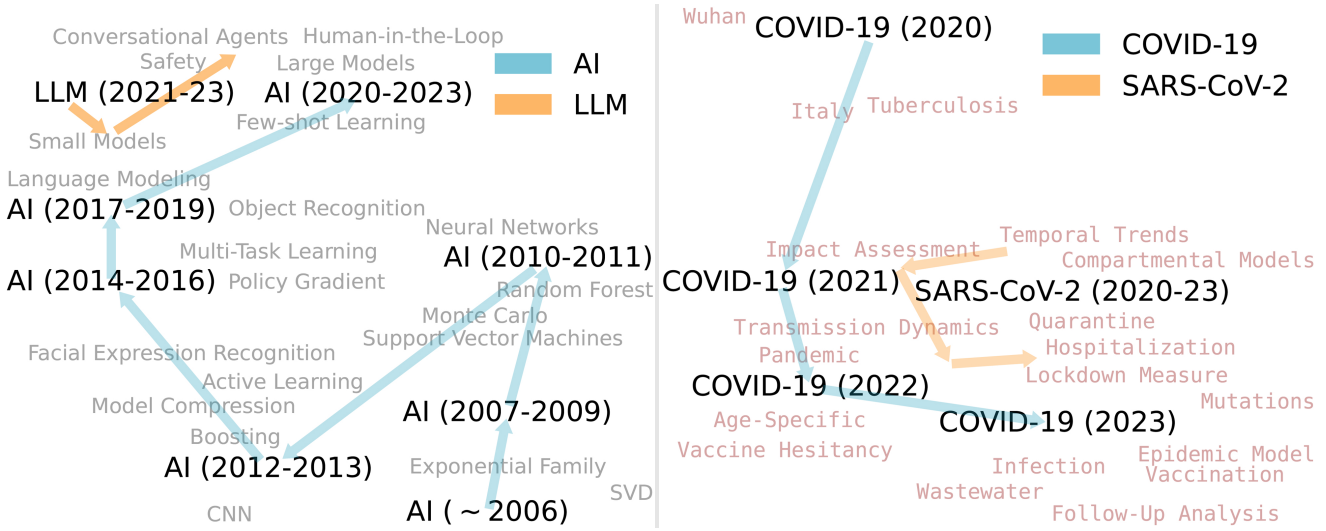

Figure 1: Keyword trajectories reflect critical paradigm shifts in AI and epidemiology research over time. Grey words represent t-SNE projections of keyword embeddings. (a) AI-related keywords; (b) COVID-related keywords.

exemplifies more citation amnesia and recency bias, favoring recent works while neglecting foundational, historical contributions. The median age of citation (AoC) for `LLM` research is 2.48 years, compared with 9.71 years for `Oral History`.

Contributions. Our contributions are three-fold:

- *Comprehensive Dataset.* We introduce *Scito2M*, an extensive dataset of over two million arXiv papers with detailed contents and citation data, offering a valuable resource for longitudinal scientometric analysis across multiple disciplines.
- *Analytic Tool Suite.* We provide a set of analysis and visualization tools for scientometric research, allowing researchers to understand the evolution of scientific terminology and citation.
- *Extensive Longitudinal Insights.* We perform a 30-year longitudinal analysis of academic literature, offering new insights into how scientific knowledge is created and shared over time.

The *Scito2M* Dataset

We select arXiv as the data source since it has been a standard for disseminating preprints. Meanwhile, the permanence of arXiv papers ensures the integrity of the citation relations. Detailed motivation is in Appendix.

Content Retrieval. We retrieved all papers published on arXiv from its establishment in 1991 to June 2024 that are under Creative Commons (CC) licenses using the arXiv API[4]. The features of the papers include titles, abstracts, arXiv categories, comments, publishing and last updating timestamps, and full texts. We carefully curated *Scito2M* to ensure representation across disciplines. Each paper is categorized into 8 subjects according to arXiv Category Taxonomy[5].

[4]https://info.arxiv.org/help/api/index.html

[5]https://arxiv.org/category_taxonomy

Feature	Statistics
#Papers	2,118,385
Time Span	1991 – 2024
#Groups	8
#Categories	156
Avg. #Categories	1.98 ± 1.05
Avg. Length (Title)	10.58 ± 4.07
Avg. Length (Abstract)	146.04 ± 53.70
Avg. #Keywords (Title)	3.06 ± 0.30
Avg. #Keywords (Abstract)	14.62 ± 1.41

Table 1: Statistics of the dataset. `Avg. Length (Title/Abstract)` are the average numbers of words in the titles & abstracts, respectively. `Avg. #Keywords (Title/Abstract)` are the average numbers of LLM-extracted keywords from each paper's title and abstract.

Citation Retrieval. As arXiv does not provide citation information, for each arXiv paper, we find the corresponding entry on semantic scholar, and retrieve the citation relations, publication venues, and author information using the semantic scholar API (Kinney et al. 2023), which allow us to analyze the citation relations among papers from different subject areas.

Keywords Extraction. Titles and abstracts in academic publications are typically crafted to highlight their most significant contributions, offering a concise yet accurate summary of the key concepts (Krishnan et al. 2017). To enhance understanding of the paper contents, we extract keywords from each title and abstract using GPT-4o (OpenAI 2023), inspired by previous works showing that LLMs demonstrate holistic understandings of academic literature (Liang et al. 2024).

Statistics. The resulting dataset contains 2.1 million papers spanning 34 years, falling under 8 groups and 156 categories. The dataset statistics is in Table 1 and the number of papers & extract keywords per year is in Figure 7. The detailed breakdown of the arXiv taxonomy is in Table 4.

Diachronic Analysis of Terminology and Lexicons

Thomas Kuhn's Theory of Paradigm Shifts describes scientific progress as a series of periodic revolutions rather than a continuous, linear accumulation of knowledge (Shapere 1964). Over time, existing research paradigms may become inadequate for address emerging problems, prompting the exploration of new, more effective approaches. To trace such shifts, we conduct diachronic analysis to study the evolution of language, concepts, and trends in academic literature.

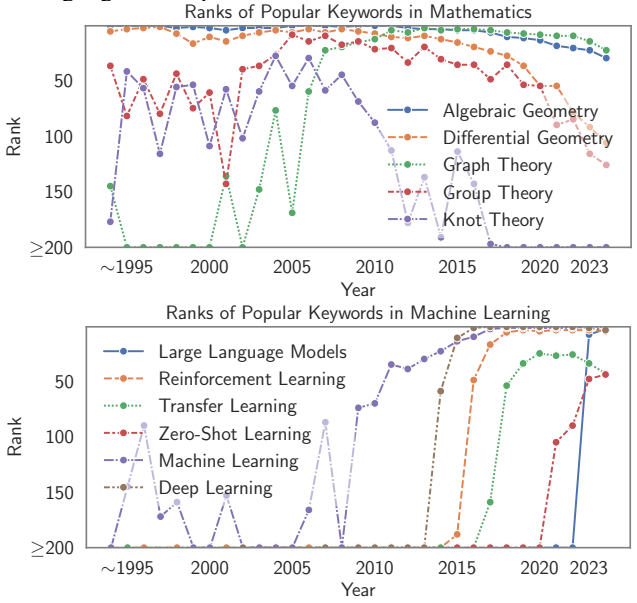

Figure 2: Evolution in the ranks of math and machine-learning terms among all keywords over time. Math keywords remain consistently popular but show a decline in the past decade, while ML keywords surged in prominence over the last ten years.

Macro-level Changes in Research Priority

To trace paradigm shifts, we divide the papers into temporal snapshots according to their publication timestamps, and rank keywords in each snapshot by frequency. Figure 2 & 4 illustrate how keyword prominence have changed shifts in algorithmic advances in recent years. From a macroscopic perspective, machine learning-related terms have significantly risen in prominence. Among the top 20 annual keywords, ML related keywords, accounting for an average of only 0.31 of the top 20 annual keywords prior to 10, but surging to an average of 9.5 from 2015 onward. In the 1990s, theoretical fields like `Quantum Field Theory` and `Particle Physics` were highly dominant and remain long-term popularity, as they were considered foundational research in advancing pure mathematics and theoretical physics. With the rise of computational technologies in the mid 2000s, data-driven methodologies such as Deep Learning and Reinforcement Learning constitute over 50% of the top-10 keywords in the dataset, surpassing foundational keywords in mathematics like `Algebraic Geometry` and `Differential Geometry`. Keywords like `Deep Learning` and `Reinforcement Learning` showed notable growth from the 2010s onward, while more recent advancements such as `Large Language Models` and `Zero-shot Learning` gaining attention in the 2020s. The top-ranked keywords in each time period (Table 3) confirms this shift and highlights the growing dominance of data-driven, AI-related research, which has overshadowed traditional theoretical fields in recent years.

Micro-level Shifts in Term Usage

From a microscopic perspective, shifts in meanings and usage of academic keywords offer insights into how a field matures and how research priorities adapt to emerging challenges or technological advancements. To trace such shifts, we analyze the co-occurrence patterns of terminology in title keywords of *Scito2M*, as titles offer a high-level summary of paper content.

Embedding Training. To capture *temporality* in terminology usage, we partition the papers into temporal snapshots. For each snapshot, we construct a keyword co-occurrence graph using the extracted keywords in *Scito2M* (Section). Each paper serves as a hyperedge connecting all keywords associated with its title. On average, each snapshot includes 9,028 keywords. We train a two-layer Graph Convolutional Network (GCN) (Kipf and Welling 2022) model with a link prediction objective on each snapshot to extract the high-order co-occurrence relations among keywords. To ensure quality of the embeddings, we filter the embeddings and keep phrases that appear $\geq$ 3 times, following Hamilton, Leskovec, and Jurafsky.

Temporal Embedding Alignment To compare word vectors across different time periods, we align them to the same embedding space using orthogonal Procrustes (Ten Berge 1977), which effectively preserve proximity of relevant terms (Hamilton, Leskovec, and Jurafsky 2016). Let $\mathbf{E}_t \in \mathbb{R}^{d\times|V|}$ be the word embedding matrix at year t, where d is the embedding dimension and V is the vocabulary, we align these embeddings by optimizing:

$$\mathbf{R}_t = \underset{\mathbf{Q}^\top\mathbf{Q}=\mathbf{I}}{\operatorname{argmin}} \|\mathbf{E}_t\mathbf{Q} - \mathbf{E}_{t+1}\|_F, \quad (1)$$

where $\mathbf{R}_t$ is the rotation matrix for orthogonal transformation that best aligns $\mathbf{E}_t$ to $\mathbf{E}_{t+1}$. $\mathbf{Q}$ is an orthogonal matrix that preserves the geometric properties of the embeddings. A *keyword trajectory* traces the movement of a keyword in the embedding space over time (Jin et al. 2024a). The trajectory of a keyword w converge to a word w' in the embedding space if w and w' frequently appear in similar contexts over a given time period. To interpret the trajectories, we project the keyword embeddings and trajectories into 2D space using t-SNE (Van der Maaten and Hinton 2008).

The keyword trajectory of `Artificial Intelligence` in Figure 1a illustrates important paradigm shifts within AI research over the past few decades. The research focus transitions from early theoretical foundations such as `Exponential Family` and `SVD` to core ML techniques like `Support Vector Machines` and `Monte Carlo` (2007 – 2009). With growing computational power and larger datasets, more complex models like `CNN` and `Boosting` (2012–2013) emerged, followed by advanced tasks like `Multi-Task Learning` and `Active Learning` (2014–2016). As

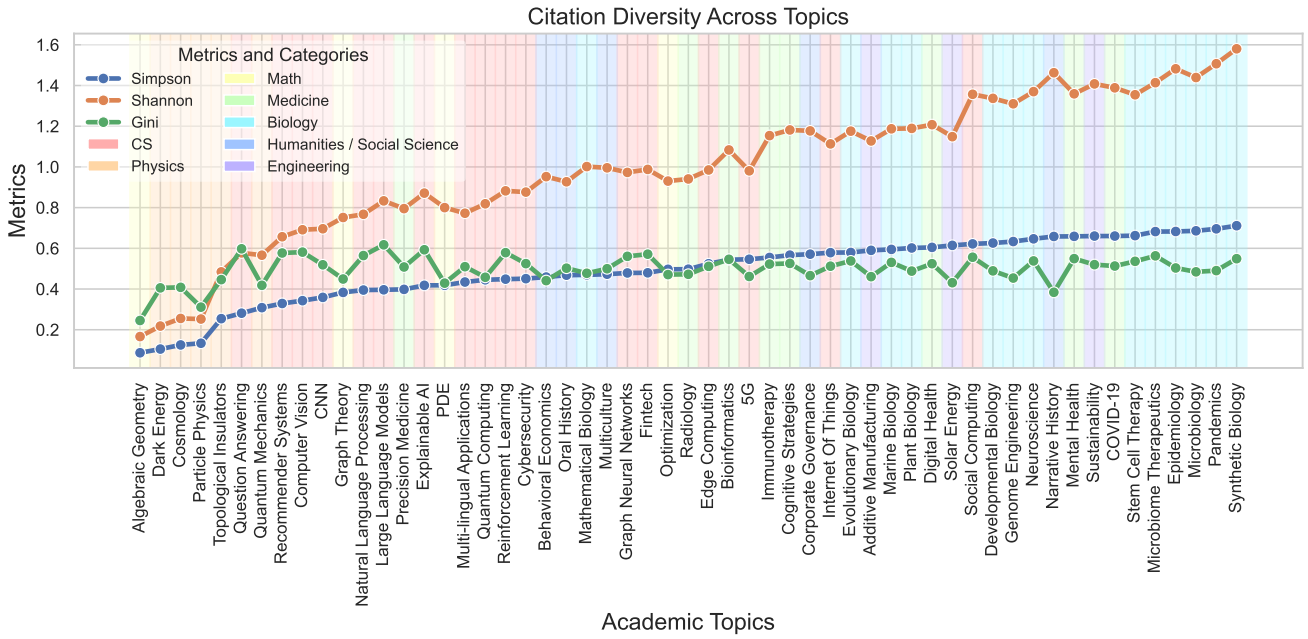

Figure 3: Citation diversity in terms of Simpson's Diversity Index, Shannon Diversity Index, and Gini Index. Topics are sorted according to Simpson's Diversity Index. The background is colored according to subject areas. Higher values for $\text{Simpson}(i), \text{Shannon}(i)$ and lower values for $\text{Gini}(i)$ imply greater diversity.

the field mature, attention turned to applications such as `Object Recognition` in 2017 and `Few-shot Learning` around 2020. Recent movements towards `Large Models` and `Human-in-the-Loop` reflect new challenges in scalability and human intervention.

Co-Evolution of Technology and Society. Technological advancements and societal changes mutually influence each other. The overlapping trajectories between `AI` and `LLM` show how AI research has increasingly incorporated large-scale models to address societal needs. The trajectory for `Large Language Models` (`LLM`), after its first occurrence around 2021, starts with technical topics like *language modeling* and is sometimes discussed in tandem with `Small Models`. As LLMs gained prominence, their associated terms evolved to reflect their broader academic adoption, incorporating keywords like *Human-in-the-Loop*, *Safety*, and *Large Models*, reflecting increased societal concerns and the need for responsible, interpretable, and user-centric AI development with human oversight as LLMs gained prominence. Notably, `LLM` represents a recent topic, thus remaining in a confined region of the embedding space filled with technical terminologies. while `AI` spans a wider scope, covering decades of research.

For each pair of keywords, the cosine similarity between their embeddings serves as a proxy for their lexical association (Pecina 2010) in academic discourse. Table 2 presents the phrases with the highest cosine similarity with `Machine Learning` across different time periods, highlighting the evolution of ML research focus from `Neural Networks` in the mid-1990s, `Machine Translation` in the early 2000s, followed by `Reinforcement Learning` in the 2010s, `Ethics` and `Scalability` concerns in the 2020s, and in recent years, `Large Language Models (LLMs)` and `Conversational Agents`. This reflects how machine learning research evolve in response to societal needs.

Near Synonymy and Lexical Choice. Near-synonymy refers to the relationship between words with similar meanings (Edmonds and Hirst 2002). For example, `SARS-CoV-2` refers to the virus or pathogen that causes the disease known as `COVID-19`. These terms are sometimes used interchangeably as **denominative variants** that refer to the same concept in non-technical contexts (Benítez Carrasco and León-Araúz 2023). Trajectories of the two keywords in Figure 1b demonstrate their lexical choices across different contexts. As a technical term, `SARS-CoV-2` is primarily used by biomedical professionals in academic contexts to ensure precision in scientific communication. The keyword follows a constrained trajectory, primarily associated with scientific discussions in epidemiology, virology, and healthcare as it intersect with keywords such as `Compartmental Models`, `Temporal Trends`, and `Hospitalization`. In contrast, `COVID-19` is a more widely recognized and accessible term that extends beyond medical contexts for clear communication across diverse fields such as politics, sociology, and economics. Its trajectory reflects broader societal concerns, with early associations including geographical keywords (`Wuhan`, `Italy`, and `India`) reflecting a continued efforts in tracking the outbreak. Later years (2021-22) shift towards public health measures like `Lockdowns` and `Quarantine`, reflecting the socio-political impacts of the pandemic.

Conclusion

We proposed *Scito2M*, a 2 million, 30-year dataset for longitudinal scientometric analysis, empowering researchers to better understand the creation, dissemination, and application of scientific knowledge. By leveraging the data, researchers can explore shifts in epistemic priorities, paradigm changes, and citation amnesia across multiple fields.

References

Abah, J. A. A. 2016. Recency Bias in the Era of Big Data: The Need to Strenthen the Status of History of Mathematics In Nigerian Schools. *Advances in Multidisciplinary Research Journal*, 2(4): 241–248.

Benítez Carrasco, V.; and León-Araúz, P. 2023. Denominative variation in the COVID-19 Open Research Dataset corpus. *Terminology*, 29(2): 252–305.

Bird, S.; Dale, R.; Dorr, B. J.; Gibson, B. R.; Joseph, M. T.; Kan, M.-Y.; Lee, D.; Powley, B.; Radev, D. R.; Tan, Y. F.; et al. 2008. The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics. In *LREC*.

Blondel, V. D.; Guillaume, J.-L.; Lambiotte, R.; and Lefebvre, E. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10): P10008.

Bollmann, M.; and Elliott, D. 2020. On forgetting to cite older papers: An analysis of the ACL Anthology. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7819–7827.

Bourdieu, P. 1983. The field of cultural production, or: The economic world reversed. *Poetics*, 12(4-5): 311–356.

Boyack, K. W.; Klavans, R.; and Börner, K. 2005. Mapping the backbone of science. *Scientometrics*, 64(3): 351–374.

Cetina, K. K. 2007. Culture in global knowledge societies: Knowledge cultures and epistemic cultures. *Interdisciplinary science reviews*, 32(4): 361–375.

Ciotti, V.; Bonaventura, M.; Nicosia, V.; Panzarasa, P.; and Latora, V. 2016. Homophily and missing links in citation networks. *EPJ Data Science*, 5: 1–14.

Donthu, N.; Kumar, S.; Mukherjee, D.; Pandey, N.; and Lim, W. M. 2021. How to conduct a bibliometric analysis: An overview and guidelines. *Journal of business research*, 133: 285–296.

Edmonds, P.; and Hirst, G. 2002. Near-synonymy and lexical choice. *Computational linguistics*, 28(2): 105–144.

Gao, S.; Hu, Y.; Janowicz, K.; and McKenzie, G. 2013. A spatiotemporal scientometrics framework for exploring the citation impact of publications and scientists. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 204–213.

Ghosh, S.; Das, D.; and Chakraborty, T. 2018. Determining sentiment in citation text and analyzing its impact on the proposed ranking index. In *Computational Linguistics and Intelligent Text Processing: 17th International Conference, CICLing 2016, Konya, Turkey, April 3–9, 2016, Revised Selected Papers, Part II 17*, 292–306. Springer.

Gibbons, M.; Limoges, C.; Scott, P.; Schwartzman, S.; and Nowotny, H. 1994. The new production of knowledge: The dynamics of science and research in contemporary societies.

Gu, Z.; Meng, F.; and Farrukh, M. 2021. Mapping the research on knowledge transfer: A scientometrics approach. *IEEE Access*, 9: 34647–34659.

Hamilton, W. L.; Leskovec, J.; and Jurafsky, D. 2016. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *ACL*, 1489–1501.

Jha, R.; Jbara, A.-A.; Qazvinian, V.; and Radev, D. R. 2017. NLP-driven citation analysis for scientometrics. *Natural Language Engineering*, 23(1): 93–130.

Jin, Y.; Lee, Y.-C.; Sharma, K.; Ye, M.; Sikka, K.; Divakaran, A.; and Kumar, S. 2023. Predicting Information Pathways Across Online Communities. In *KDD*.

Jin, Y.; Zhao, A.; Lee, Y.-C.; Ye, M.; Divakaran, A.; and Kumar, S. 2024a. Empowering Interdisciplinary Insights with Dynamic Graph Embedding Trajectories. *arXiv:2406.17963*.

Jin, Y.; Zhao, Q.; Wang, Y.; Chen, H.; Zhu, K.; Xiao, Y.; and Wang, J. 2024b. AgentReview: Exploring Peer Review Dynamics with LLM Agents. In *EMNLP*.

Kinney, R.; Anastasiades, C.; Authur, R.; Beltagy, I.; Bragg, J.; Buraczynski, A.; Cachola, I.; Candra, S.; Chandrasekhar, Y.; Cohan, A.; et al. 2023. The semantic scholar open data platform. *arXiv:2301.10140*.

Kipf, T. N.; and Welling, M. 2022. Semi-Supervised Classification with Graph Convolutional Networks. In *ICLR*.

Koch, B.; Denton, E.; Hanna, A.; and Foster, J. G. 2021. Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research. *NeurIPS*, 1.

Krishnan, A.; Sankar, A.; Zhi, S.; and Han, J. 2017. Unsupervised concept categorization and extraction from scientific document titles. In *CIKM*, 1339–1348.

Leto, A.; Roy, S.; Hoyle, A.; Acuna, D.; and Pacheco, M. L. 2024. A First Step towards Measuring Interdisciplinary Engagement in Scientific Publications: A Case Study on NLP+ CSS Research. In *ACL NLP+CSS Workshop*, 144–158.

Li, L.; Wang, Y.; Xu, R.; Wang, P.; Feng, X.; Kong, L.; and Liu, Q. 2024. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. In *ACL*.

Li, Y.; Xiong, H.; Kong, L.; Zhang, R.; Dou, D.; and Chen, G. 2022. Meta hierarchical reinforced learning to rank for recommendation: a comprehensive study in moocs. In *Joint European conference on machine learning and knowledge discovery in databases*, 302–317. Springer.

Li, Y.; Xiong, H.; Kong, L.; Zhang, R.; Xu, F.; Chen, G.; and Li, M. 2023. MHRR: MOOCs Recommender Service With Meta Hierarchical Reinforced Ranking. *IEEE Transactions on Services Computing*.

Li, Z.; Chang, Y.; and Le, X. 2024. Simulating Expert Discussions with Multi-agent for Enhanced Scientific Problem Solving. In *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, 243–256.

Liang, W.; Zhang, Y.; Cao, H.; Wang, B.; Ding, D. Y.; Yang, X.; Vodrahalli, K.; He, S.; Smith, D. S.; Yin, Y.; et al. 2024. Can large language models provide useful feedback on research papers? A large-scale empirical analysis. *NEJM AI*, 1(8): AIoa2400196.

Nguyen, H.; and Eger, S. 2024. Is there really a Citation Age Bias in NLP? *arXiv:2401.03545*.

OpenAI. 2023. GPT-4 Technical Report. *Arxiv Preprint*, arXiv:2303.08774.

Oppenlaender, J. 2024. Past, Present, and Future of Citation Practices in HCI. *arXiv:2405.16526*.

Page, S. E.; et al. 2006. Path dependence. *Quarterly Journal of Political Science*, 1(1): 87–115.

Pecina, P. 2010. Lexical association measures and collocation extraction. *Language resources and evaluation*, 44: 137–158.

Radev, D.; and Abu-Jbara, A. 2012. Rediscovering ACL discoveries through the lens of ACL anthology network citing sentences. In *Proceedings of the ACL-2012 Special workshop on rediscovering 50 years of discoveries*, 1–12.

Ram, A. 1991. A theory of questions and question asking. *Journal of the Learning Sciences*, 1(3-4): 273–318.

Rogers, E. M.; Singhal, A.; and Quinlan, M. M. 2014. Diffusion of innovations. In *An integrated approach to communication theory and research*, 432–448. Routledge.

Schwarz, N.; Bless, H.; Strack, F.; Klumpp, G.; Rittenauer-Schatka, H.; and Simons, A. 1991. Ease of retrieval as information: Another look at the availability heuristic. *Journal of Personality and Social psychology*, 61(2): 195.

Shapere, D. 1964. The structure of scientific revolutions. *The Philosophical Review*, 73(3): 383–394.

Sim, Y.; Smith, N. A.; and Smith, D. A. 2012. Discovering factions in the computational linguistics community. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, 22–32.

Şimşek, Ö.; and Jensen, D. 2008. Navigating networks by using homophily and degree. *Proceedings of the National Academy of Sciences*, 105(35): 12758–12762.

Singh, J.; Rungta, M.; Yang, D.; and Mohammad, S. 2023. Forgotten Knowledge: Examining the Citational Amnesia in NLP. In *ACL*, 6192–6208.

Ten Berge, J. M. 1977. Orthogonal Procrustes rotation for two or more matrices. *Psychometrika*, 42: 267–276.

Thilakaratne, M.; Falkner, K.; and Atapattu, T. 2018. Automatic detection of cross-disciplinary knowledge associations. In *Proceedings of ACL 2018, Student Research Workshop*, 45–51.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Wang, Y.; Guo, Q.; Yao, W.; Zhang, H.; Zhang, X.; Wu, Z.; Zhang, M.; Dai, X.; Zhang, M.; Wen, Q.; et al. 2024. AutoSurvey: Large Language Models Can Automatically Write Surveys. *arXiv:2406.10252*.

Xiong, H.; Bian, J.; Li, Y.; Li, X.; Du, M.; Wang, S.; Yin, D.; and Helal, S. 2024. When Search Engine Services meet Large Language Models: Visions and Challenges. *IEEE Transactions on Services Computing*.

Zhang, C.; Bu, Y.; Ding, Y.; and Xu, J. 2018. Understanding scientific collaboration: Homophily, transitivity, and preferential attachment. *Journal of the Association for Information Science and Technology*, 69(1): 72–86.

Zhang, J.; Zhang, H.; Deng, Z.; and Roth, D. 2022. Investigating fairness disparities in peer review: A language model enhanced approach. *arXiv:2211.06398*.

Experimental Details

Metrics for Topical Diversity

Metrics. To quantify topical diversity, we measure the proportion of references associated with each academic field. Let $R(i)$ denote the set of references for a paper i. Each reference j is associated with one or more subject areas denoted by $F(i, j)$. Here, we use the Field of Study attribute in semantic scholar (Kinney et al. 2023). If a reference belongs to multiple fields, each subject is credited with an equal fraction of the reference's contribution. The total contribution of a subject s to paper i is:

$$C(i, s) = \sum_{j=1}^{|R(i)|} \frac{\delta_s(j)}{|F(i, j)|} \tag{2}$$

where $\delta_s(j)$ is an indicator function that equals 1 if subject s is in $F(i, j)$, and 0 otherwise. Topical diversity is then measured using three well-established indices: Simpson's Diversity Index ($\mathrm{Simpson}(i)$), Shannon's Diversity Index ($\mathrm{Shannon}(i)$), and Gini Index ($\mathrm{Gini}(i)$):

$$\mathrm{Simpson}(i) = 1 - \sum_{j=1}^{k_i} C(i, j)^2, \tag{3}$$

$$\mathrm{Shannon}(i) = -\sum_{j=1}^{k_i} C(i, j) \log C(i, j), \tag{4}$$

$$\mathrm{Gini}(i) = \frac{\sum_{s=1}^{k_i} \sum_{t=1}^{k_i} |C(i, s) - C(i, t)|}{2k_i \sum_{s=1}^{k_i} C(i, s)}. \tag{5}$$

$\mathrm{Simpson}(i)$ measures the probability that two randomly selected references belong to different subjects. $\mathrm{Shannon}(i)$ quantifies the uncertainty in predicting the subject area of a randomly selected reference. $\mathrm{Gini}(i)$ assesses the inequality in the distribution of references across subject areas.

Hyperparameters

For keyword trajectory generation (Section), we train the Graph Convolutional Networks (GCN) model for 50 epochs using the Adam optimizer with an initial learning rate of 0.01, along with a linear decay learning rate scheduler. We apply a 1:1 negative sampling ratio to balance positive and negative edges. For t-SNE visualization, we set the perplexity to 30 and run the optimization for a maximum of 1000 iterations.

Usage of AI Assistants

We use GPT-4o to improve the writing of our manuscript.

Data Release Plan

We plan to release our dataset on Zenodo to ensure long-term access. To comply with arXiv's policy prohibiting third-party hosting of e-prints, we provide scripts for downloading the PDF e-prints directly from arXiv.

Year	Closest Keywords to `Machine Learning`
$\sim$ 1995	Information Retrieval, Classification, Neural Networks, POS Tagging
2000	Ensembles, Logic, Optimization, Machine Translation
2005	Regularization, Support Vector Machines, Data Mining, Reinforcement Learning
2010	Image Classification, Kernel Methods, Reinforcement Learning, Transfer Learning
2015	Object Recognition, Question Answering, Image Generation
2020	Ethics, Scalability, Model Explanation, Post-hoc
2022	Data-driven Analysis, Performance Improvement, Medical Imaging, Cross-lingual
2024	General AI, Large Language Models, Conversational Agents, Retrieval-augmented Generation

Table 2: Closest keywords to `Machine Learning` in the embedding space, reflecting the shifting focus of academic discourse.

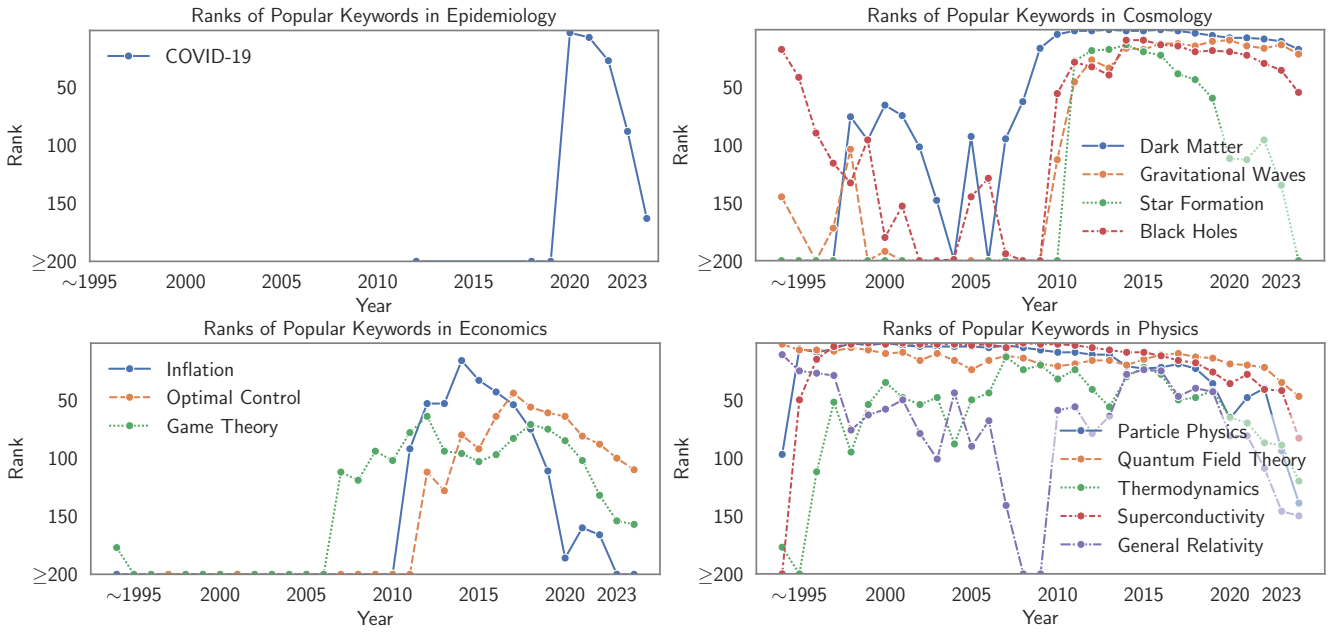

Figure 4: Ranks of mathematics-related (top) and machine-learning-related (bottom) keywords.

Limitations

We acknowledge the following limitations about *Scito2M*. First, while arXiv is widely adopted and represents a significant portion of academic literature, it may not fully reflect research published in local venues or non-English conferences that are less closely associated with arXiv, potentially leading to gaps in coverage. Second, citation retrieval relies on Semantic Scholar data, which, while generally robust, may occasionally lack metadata for very recent papers or those in specialized areas. Lastly, the arXiv taxonomy can sometimes group diverse subfields, complicating the task of distinguishing closely related research areas. Future enhancements can integrate additional data sources such as CrossRef[6] and Google Scholar[7] to enhance topical & citation coverage and subject classification. Expanding the temporal scope to include real-time updates would also ensure the dataset's ongoing relevance, supporting longitudinal studies on shifting research trends, especially in response to major global events or scientific breakthroughs. Future works can also expand the coverage of papers into broader subject areas, such as humanities.

Ethical Considerations

Compliance with Data Usage Policies. We are committed to ensuring that all data collection and analyses strictly comply with Terms of Use for arXiv API[8], Semantic Scholar API[9], and Ai2 Privacy Policy[10]. The dataset is used strictly for academic research purposes, focusing on scientometric analysis, and adhering to all data privacy and intellectual property guidelines. The dataset usage aligns with recommended academic use cases, focusing on the retrieval and analysis of scientometric data for research. We strictly follow ethical guidelines concerning access, data privacy, and intellectual property. Importantly, *Scito2M* should not

[6]https://www.crossref.org/

[7]https://scholar.google.com/

[8]https://info.arxiv.org/help/api/tou.html

[9]https://www.semanticscholar.org/product/api/license

[10]https://allenai.org/privacy-policy/2022-07-21

Time Period	Top 5 Most Mentioned Keywords
$\sim$ 1994	algebraic geometry, quantum field theory, quantum groups, quantum gravity, lattice qcd
1995 — 2004	superconductivity, algebraic geometry, lattice qcd, particle physics, quantum mechanics
2005 — 2014	algebraic geometry, superconductivity, quantum mechanics, dark matter, cosmology
2015 — 2017	deep learning, dark matter, graph theory, cosmology, optimization
2018 — 2019	deep learning, machine learning, neural networks, dark matter, reinforcement learning
2020 — 2021	deep learning, machine learning, neural networks, reinforcement learning, covid-19
2022 — 2023	deep learning, machine learning, neural networks, reinforcement learning, quantum computing
2024 $\sim$	machine learning, language models, large language models, deep learning, reinforcement learning

Table 3: Most frequently mentioned keywords in paper titles across different time periods. arXiv shows a noticeable shift in epistemic priority from mathematics and physics to computer science topics.

Subject	Category
Computer Science (cs)	cs.AI, cs.AR, cs.CC, cs.CE, cs.CG, cs.CL, cs.CR, cs.CV, cs.CY, cs.DB, cs.DC, cs.DL, cs.DM, cs.DS, cs.ET, cs.FL, cs.GL, cs.GR, cs.GT, cs.HC, cs.IR, cs.IT, cs.LG, cs.LO, cs.MA, cs.ML, cs.MM, cs.MS, cs.NA, cs.NE, cs.NI, cs.OH, cs.OS, cs.PF, cs.PL, cs.RO, cs.SC, cs.SD, cs.SE, cs.SI, cs.SY
Economics (econ)	econ.EM, econ.GN, econ.TH
Electrical Engineering and Systems Science (eess)	eess.AS, eess.IV, eess.SP, eess.SY
Mathematics (math)	math.AC, math.AG, math.AP, math.AT, math.CA, math.CO, math.CT, math.CV, math.DG, math.DS, math.FA, math.GM, math.GN, math.GR, math.GT, math.HO, math.IT, math.KT, math.LO, math.MG, math.MP, math.NA, math.NT, math.OA, math.OC, math.PR, math.QA, math.RA, math.RT, math.SG, math.SP, math.ST
Physics (physics)	astro-ph.CO, astro-ph.EP, astro-ph.GA, astro-ph.HE, astro-ph.IM, astro-ph.SR, cond-mat.dis-nn, cond-mat.mes-hall, cond-mat.mtrl-sci, cond-mat.other, cond-mat.quant-gas, cond-mat.soft, cond-mat.stat-mech, cond-mat.str-el, cond-mat.supr-con, gr-qc, hep-ex, hep-lat, hep-ph, hep-th, math-ph, nlin.AO, nlin.CD, nlin.CG, nlin.PS, nlin.SI, nucl-ex, nucl-th, physics.acc-ph, physics.ao-ph, physics.app-ph, physics.atm-clus, physics.atom-ph, physics.bio-ph, physics.chem-ph, physics.class-ph, physics.comp-ph, physics.data-an, physics.ed-ph, physics.flu-dyn, physics.gen-ph, physics.geo-ph, physics.hist-ph, physics.ins-det, physics.med-ph, physics.optics, physics.plasm-ph, physics.pop-ph, physics.soc-ph, physics.space-ph, quant-ph
Quantitative Biology (q-bio)	q-bio.BM, q-bio.CB, q-bio.GN, q-bio.MN, q-bio.NC, q-bio.OT, q-bio.PE, q-bio.QM, q-bio.SC, q-bio.TO
Quantitative Finance (q-fin)	q-fin.CP, q-fin.EC, q-fin.GN, q-fin.MF, q-fin.PM, q-fin.PR, q-fin.RM, q-fin.ST, q-fin.TR
Statistics (stat)	stat.AP, stat.CO, stat.ME, stat.ML, stat.OT, stat.TH

Table 4: The arXiv Taxonomy contains 8 major subject areas, including computer science (cs), economics (econ), electrical engineering and systems science (eess), mathematics (math), quantitative biology (q-bio), quantitative finance (q-fin), statistics (stat), and physics.

be used for unfair or harmful evaluations of individual researchers, especially those from underrepresented groups or early-career scholars. Ethical usage should prioritize promoting open science, transparency, fairness, and responsibility within the academic community.

Path Dependency. (Page et al. 2006) Our results show that academic research is often shaped by historical trajectories, where early dominant paradigms influence future directions (a concept known as *path dependency*) (Page et al. 2006). This can result in a *lock-in* effect, where unconventional ideas or new directions that diverge from established patterns are less likely to gain traction. This trend is reinforced by institutional practices, funding agencies, and high-impact journals, which tend to favor research that builds on well-established theories. Future investigations can explore how these historical and institutional factors limit the diversification of research.

Citation Patterns and Their Implications. Our findings in citation diversity have profound implications for research development and interdisciplinary collaboration. Citation **breadth**–the range of topics referenced–serves as an indicator of interdisciplinarity. A higher citation breadth suggests

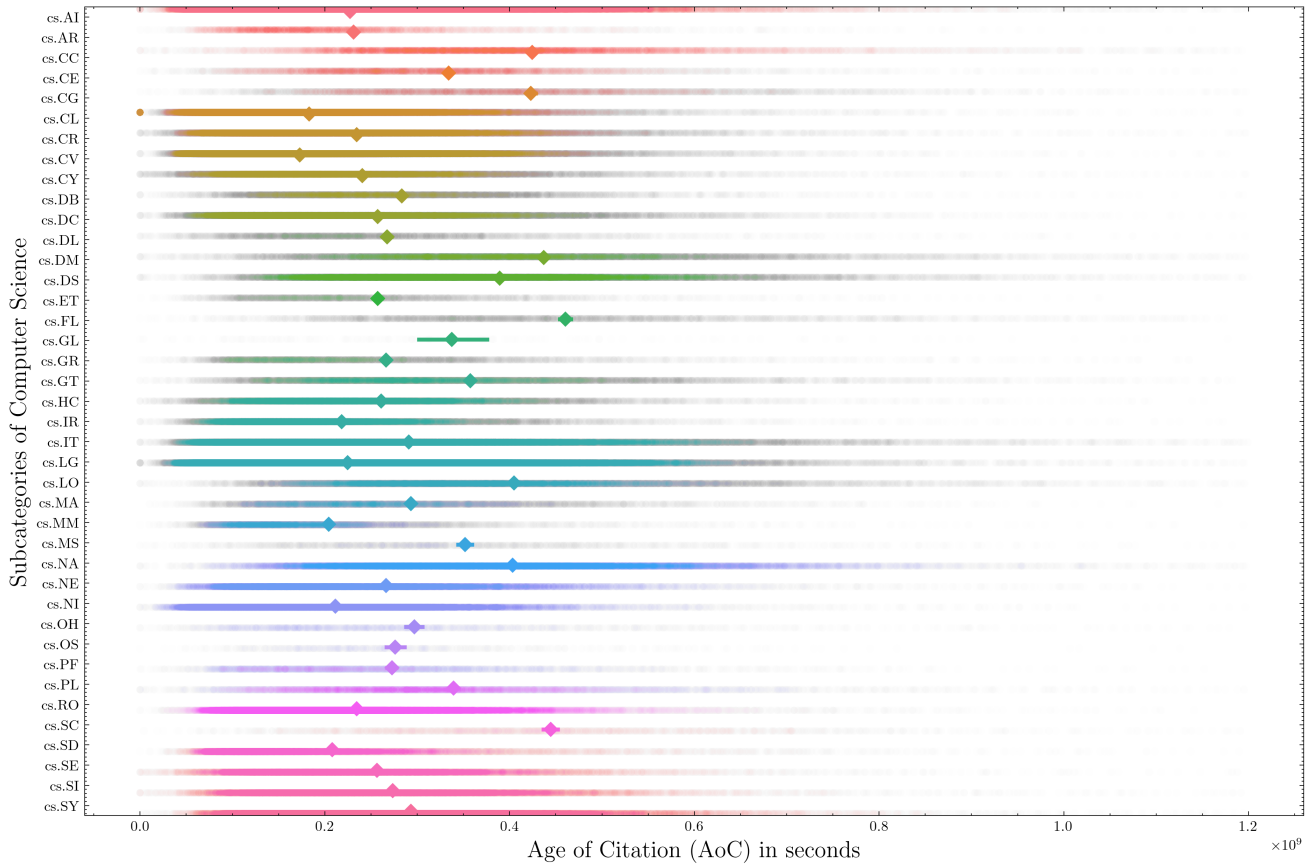

Figure 5: Age of Citation (AoC) of papers under different subfields in CS. The median AoC is marked in diamond shape.

that a field integrates diverse knowledge, which is crucial for addressing complex global challenges. Conversely, a low citation breadth indicates a focus on a narrow set of foundational theories or methods. While this specialization can lead to deep expertise, it may also risk intellectual isolation and hinder innovation.

In terms of citation **depth** – the temporal span of cited works – fields like natural language processing (NLP) demonstrate *citation amnesia*. Approximately 62% of citations point to works published within the last five years, and only about 17% are works older than ten years (Singh et al. 2023). Our research has highlighted the prevalence of this problem and underscored the need to strike a balance between maintaining relevance with recent work and preserving a strong connection to foundational theories.

Cognitive Load and Information Availability. The growing volume of publications poses a significant cognitive load on researchers, making it challenging to stay current with all relevant literature. To cope with this overload, researchers may rely on cognitive shortcuts, such as the **availability heuristic** (Schwarz et al. 1991), favoring recent or highly visible information over older, less prominent work. Future research should consider the impact of these cognitive biases on citation practices and the potential narrowing of scholarly discourse.

Motivation for Selecting arXiv

We select arXiv[11] as the data source due to the following reasons:

- *Widespread Adoption*: arXiv is the go-to platform for researchers, especially in technical fields such as physics, computer science, and mathematics, where it has become a standard for disseminating preprints.
- *Community Trust and Early Impact.* Unlike papers published in conferences or journals, which often involve lengthy peer review processes, arXiv allows researchers to share their findings quickly as preprints. This makes arXiv a critical resource for accessing the latest scientific developments in real-time and gauging when breakthrough technologies appear. arXiv papers often have significant early impact, as many researchers use the platform to gauge emerging trends and cite preprints in their work even before formal publication.
- *Multidisciplinary Coverage.* As shown in the arXiv Taxonomy in Figure 4, arXiv supports a wide array of fields, making it an ideal platform for cross-domain studies and the exploration of interdisciplinary trends.
- *Data Permanence and Integrity.* The permanence of papers on arXiv means that once they are part of the dataset,

[11]https://arxiv.org/

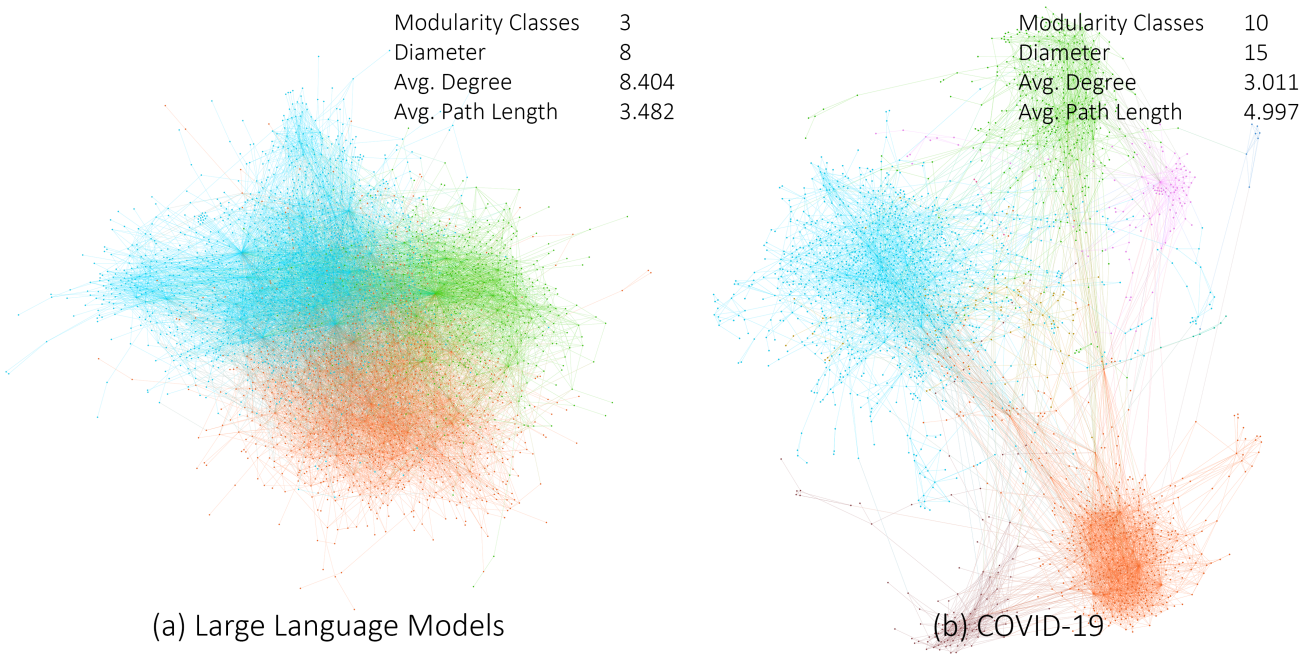

Figure 6: Citation graphs of LLM (left) and epidemiology (right) literature show distinct citation patterns. LLM publications is densely connected with three modularity classes forming a large cluster, suggesting a highly specialized, self-referential field. The epidemiology graph, with 10 modularity classes, highlights several distinct research subfields that reflect the interdisciplinary nature of the field.

they remain available indefinitely. This is unlike other platforms such as ResearchGate[12], which allow authors to remove or significantly modify their works. As arXiv papers cannot be deleted, citation networks based on arXiv papers remain intact and reliable. This is crucial for scientometric analysis that requires accurate and stable citation relationships.

- *Versioning* Authors can submit updated versions of their work, reflecting changes, corrections, or new findings. This versioning system provides a clear historical record of how a paper evolves without losing the original submission. In contrast, other platforms may not track versions as clearly or may allow full deletion, which can obscure the evolution of research and make it difficult to trace the development of scientific ideas over time.

We also acknowledge that the resulting dataset is mainly English-centric. Future works can incorporate non-English academic literature for a more holistic understanding of global academic landscape.

Broader Impact

The *Scito2M* provides a comprehensive resource for analyzing academic publications over the *past* 30 years, understanding the *present* spatio-temporal knowledge dissemination, and predicting *future* research trends and emerging fields.

Identifying Emerging Fields and Influential Papers. *Scito2M* can be used to identify emerging disciplines, cross-domain collaborations, and effects of global events on shifting academic focus. Researchers can also use early citation patterns provided in *Scito2M* to predict influential papers and topics, providing insights into the trajectory of scientific discoveries and intellectual influence.

Fine-grained Citation Analysis. By analyzing the text of citing sentences, researchers can perform fine-grained citation analysis to determine the purpose and sentiment of citations. This enables understanding of whether current works are more critical or supportive of new ideas.

Enhancing Research Discoverability and Knowledge Retrieval. With its structured abstracts, titles, and keywords, *Scito2M* serves as a rich resource for training and benchmarking NLP models in tasks such as text classification, summarization, and keyword extraction. This can significantly enhance research discoverability, support automated literature reviews, and contribute to more efficient knowledge retrieval. Future work can explore technological impacts on academic focus, facilitating the automatic discovery of scientific findings from existing literature (Li, Chang, and Le 2024).

Studying Spatio-Temporal Knowledge Dissemination. The dataset enables the analysis of spatio-temporal dissemination patterns of scientific knowledge, providing insights into how research ideas originate in specific geographical or subject areas and spread globally over time.

Inspiring New Research Questions. According to the theory of questions and question asking, answers to existing questions often give rise to new ones (Ram 1991). By examining existing studies within the dataset, researchers can identify research gaps and inspire new questions and direc-

[12]https://www.researchgate.net/

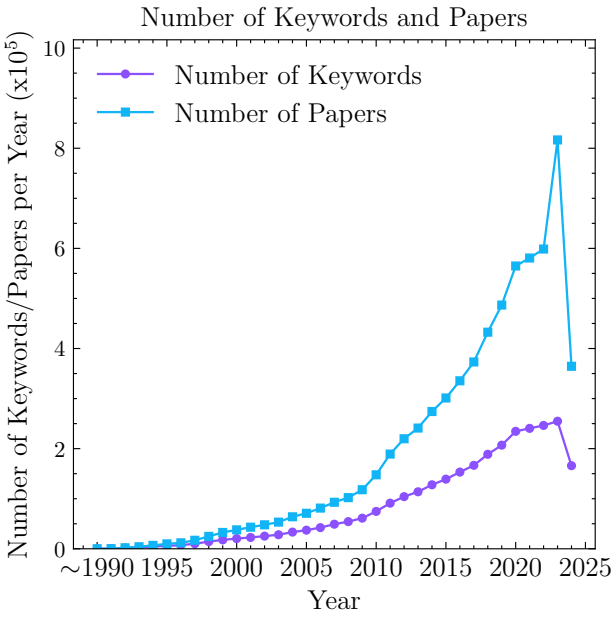

Figure 7: Number of papers per year *Scito2M* and their associated keywords. Note that the papers are collected up until June 2024.

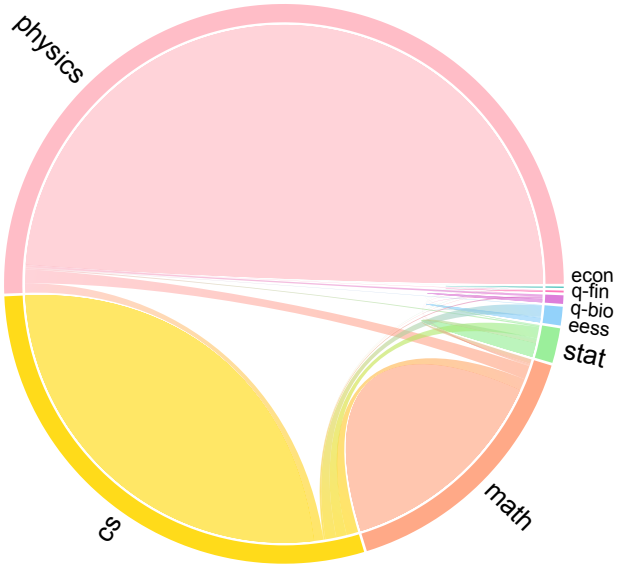

Figure 8: Literature in *Scito2M* exhibits higher intra-disciplinary than cross-disciplinary citations.

tions.

Automatic Survey Generation and Literature Retrieval. *Scito2M* can be utilized to automatically generate surveys (Wang et al. 2024), aiding in the synthesis of knowledge and identification of research gaps. It enables fast literature search and retrieval through keywords (Xiong et al. 2024), improving access to relevant research and facilitating scholarly communication.

Topical & Temporal Citation Dynamics

Citations form the backbone of scientific inquiry by connecting current research with prior contributions (Boyack, Klavans, and Börner 2005). The diversity of citations is essential as it fosters comprehensive understandings of research problems and encourages interdisciplinary collaboration. In this section, we analyze two key aspects of citations: topical diversity, which reflects the *breadth* of domains a paper engages with, and temporal diversity, which reveals the *depth* of references across time periods.

Topical Diversity

The topical diversity reflects the *breadth* of citation. A higher topical diversity suggests that the paper draws on insights from a broader range of disciplines, fostering cross-disciplinary innovation and perspectives. To measure this, we use three well-established metrics: Simpson's Diversity Index (Simpson(i)), Shannon's Diversity Index (Shannon(i)), and Gini Index (Gini(i)). Details can be found in Appendix .

Homophily (Zhang et al. 2018). Networks of academic knowledge sharing often exhibit *homophily*, where papers preferentially connect with works from their own discipline (Şimşek and Jensen 2008; Ciotti et al. 2016). The intra- and cross-disciplinary citation in Figure 8 show that over 91.0% citations occur within the same discipline. This *homophily* stems from researchers' tendency towards shared methods, expertise, language, and conceptual frameworks within an academic field.

Epistemic Culture refer to the distinct ways in which knowledge is produced, validated, and shared within different scientific disciplines (Cetina 2007). According to Cetina, global scientific knowledge production is characterized by *disunity*, where each subject area operates under its distinct epistemic culture. To examine these differences, We calculate citation diversity across academic topics (Figure 3), revealing how epistemic cultures shape the interdisciplinary scope and maturity of research fields. More theoretical or *pure* disciplines, such as `Algebraic Geometry` in mathematics and `Quantum Mechanics`, `Dark Energy`, and `Particle Physics` in physics, exhibit an epistemic culture with an internal focus. These fields rely on well-established, codified knowledge practices that prioritize internal citations, reinforcing disciplinary coherence while limiting interdisciplinary input from external fields. Conversely, emerging areas like `Large Language Models` (LLMs) and `Digital Health` demonstrate broader interdisciplinary citation patterns. As these areas are still developing their intellectual foundations, they draw on a wide range of research to shape their evolving epistemic frameworks.

Knowledge Production Modes and Field Theory. Gibbons et al. conceptualized two modes of knowledge production. Mode 1 refers to basic research driven by fundamental principles and theories, as seen in physics and math. Mode 2 is problem-oriented, usually associated with applied research that requires interdisciplinary collaboration, as exemplified by `Graph Neural Networks` (GNNs)

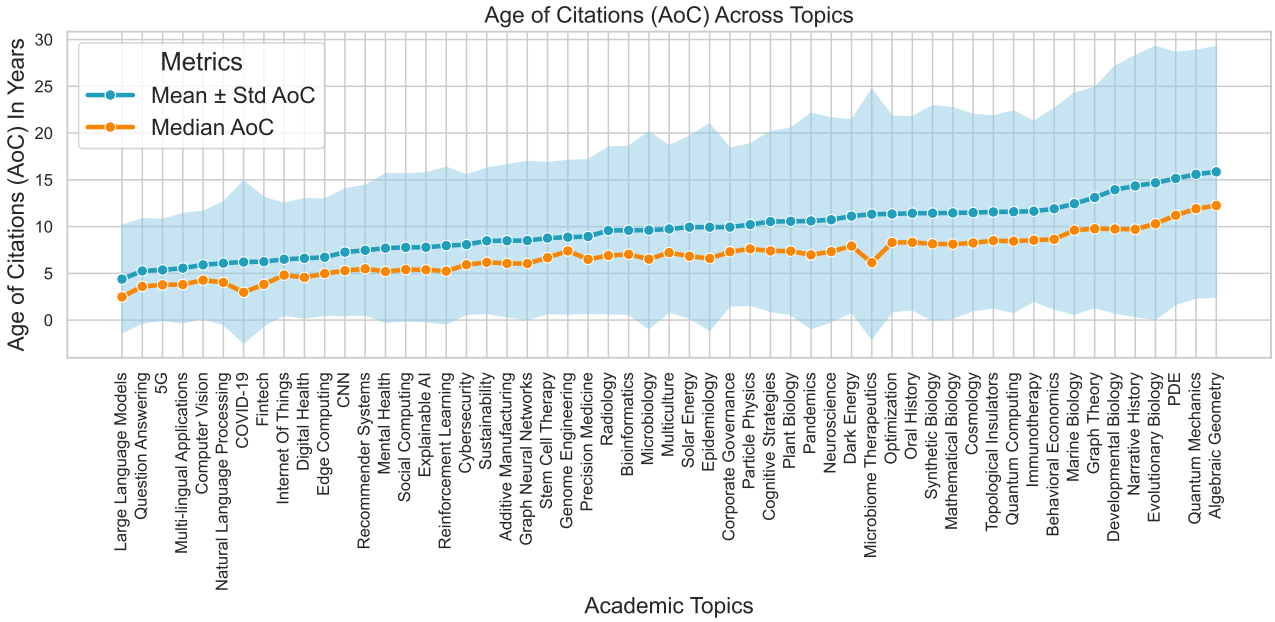

Figure 9: Age of Citation (AoC) of papers under each academic topic.

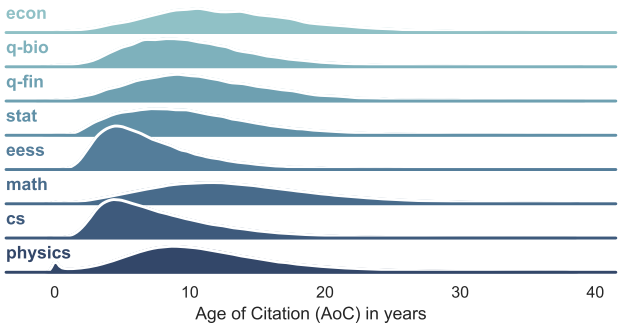

Figure 10: Age of Citation (AoC) across various arXiv subjects[13] shows distinct trends. `eess` and `cs` exhibit left-skewed distributions, indicating a preference towards recent citation. In contrast, disciplines such as `physics`, `math`, and `econ` demonstrate broader AoCs, reflecting their reliance on historical foundational research.

and `Quantum Computing`. These fields align with Mode 2, involving higher contributions from prior research (Simpson(i) $= 0.479/0.445$) compared to theory-centric topics like `Graph Theory` and `Quantum Mechanism` (Simpson(i) $= 0.383/0.308$). From the perspective of **field theory** (Bourdieu 1983), academic disciplines operate like competitive *social fields*, where scholars vie for intellectual capital. Established fields like `Graph Theory` have a highly structured system of knowledge production with a stable, widely accepted body of references. Scholars are expected to operate within a tightly regulated intellectual space, where contributions are expected to align closely with established theories. In contrast, emerging fields like `Graph Neural Networks` operate in less structured intellectual spaces, with scholars drawing upon a diverse array of disciplines, such as computer science, machine learning, and domain-specific fields like social network analysis and recommender systems (Li et al. 2022, 2023). Similarly, interdisciplinary topics such as `Social Computing` and `COVID-19` exhibit greater diversity across metrics as they integrate insights from multiple knowledge domains–sociology, epidemiology, computer science, and network science–to address complex, multi-faceted problems.

Temporal Diversity

Temporal diversity measures the distribution of citations across different time periods, reflecting the *depth* and profound influence of academic works over time. Examining temporal diversity can uncover issues like *citation amnesia* (Singh et al. 2023) and **Recency Bias** (Abah 2016), the tendency to prioritize recent contributions and overlook significant historical knowledge.

Pace of Innovation refers to the speed of technological advancement, which varies significantly across academic fields. We quantify this pace using the Age of Citation (AoC) (Singh et al. 2023), defined as the publication time gap between a paper and its cited works. `eess` (Electrical Engineering and Systems Science) and `cs` (Computer Science) exhibit right-skewed AoC distributions (Figure 10), indicating a rapid pace of development driven by recent research. This is especially evident in fast-evolving topics like `Natural Language Processing` and `Machine Learning` (Figure 9), which show median AoCs of 4.02 and 5.06 years. In contrast, humanities subjects like `Narrative History` and `Oral History` have significantly higher median AoCs of 8.31 and 9.71 years. Disciplines such as `econ` and `math` display relatively flat AoC distributions, suggesting a holistic development process that relies on both recent innovations and long-established foundational works.

Citation Network Structures. Figure 6 & 11 display citation graphs across various research topics, where nodes represent papers, edges represent citation relations, and colors

indicate communities identified via the Louvain Community Detection Algorithm (Blondel et al. 2008). The LLM literature (Figure 6a) displays a densely interconnected network with three major modularity classes forming a dominant cluster. This suggests a rapid and concentrated **diffusion of innovation** (Rogers, Singhal, and Quinlan 2014) as researchers converge on core methodologies, creating a relatively unified body of knowledge. The rapid innovation may be driven by strong technological momentum and the lower barriers to entry, though concerns about the oversight of social impacts remain. The epidemiology literature (Figure 6b) exhibits a fragmented structure with weak ties linking distinct cluster. Each cluster represents research on specific outbreaks or frameworks, often spanning multiple decades. While allowing for domain-specific innovations, it complicates the synthesis of knowledge across the broader field.

Related Works

Scientometrics plays a crucial role in understanding the structure and evolution of scientific research, providing insights into how knowledge is produced, disseminated, and consumed across academic communities. Previous scientometric studies explored aspects such as interdisciplinary knowledge associations (Leto et al. 2024; Thilakaratne, Falkner, and Atapattu 2018), academic knowledge diffusion (Jin et al. 2023, 2024a), concept extraction (Krishnan et al. 2017), academic community factions (Sim, Smith, and Smith 2012), figures (Li et al. 2024), research artifacts usage patterns (Koch et al. 2021), and estimation measures of research impact (Radev and Abu-Jbara 2012).

Citation analysis, in particular, is widely used to measure impact of researchers, publications, institutions, and venues across disciplines. While studies have investigated dimensions such as citation polarity (Ghosh, Das, and Chakraborty 2018; Radev and Abu-Jbara 2012), purpose (Jha et al. 2017), and influence (Gao et al. 2013), most existing works fail short in cross-disciplinary scope (Bollmann and Elliott 2020; Bird et al. 2008) and temporal depth (Koch et al. 2021; Zhang et al. 2022).

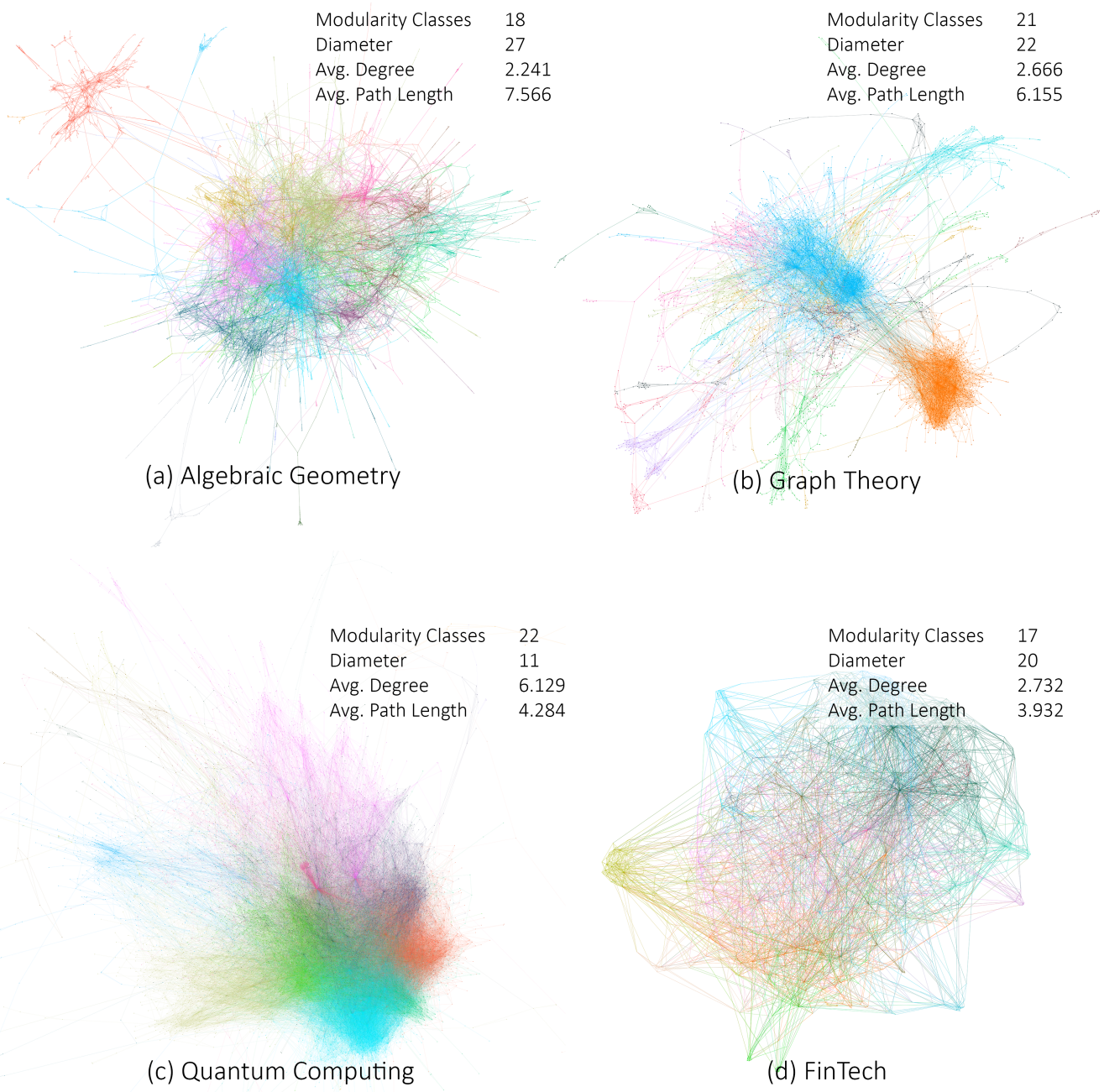

Figure 11: Citation graphs of a) Algebraic Geometry, b) Graph Theory, c) Quantum Computing, and d) FinTech. Graph Theory shows two relatively dominant, independent areas of focus, while Quantum Computing exhibits strong interconnections with diverse research communities.