



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

Yiqiao Jin, Yijia Xiao, Yiyang Wang,
Jindong Wang



WILLIAM
& MARY
CHARTERED 1693



 GitHub

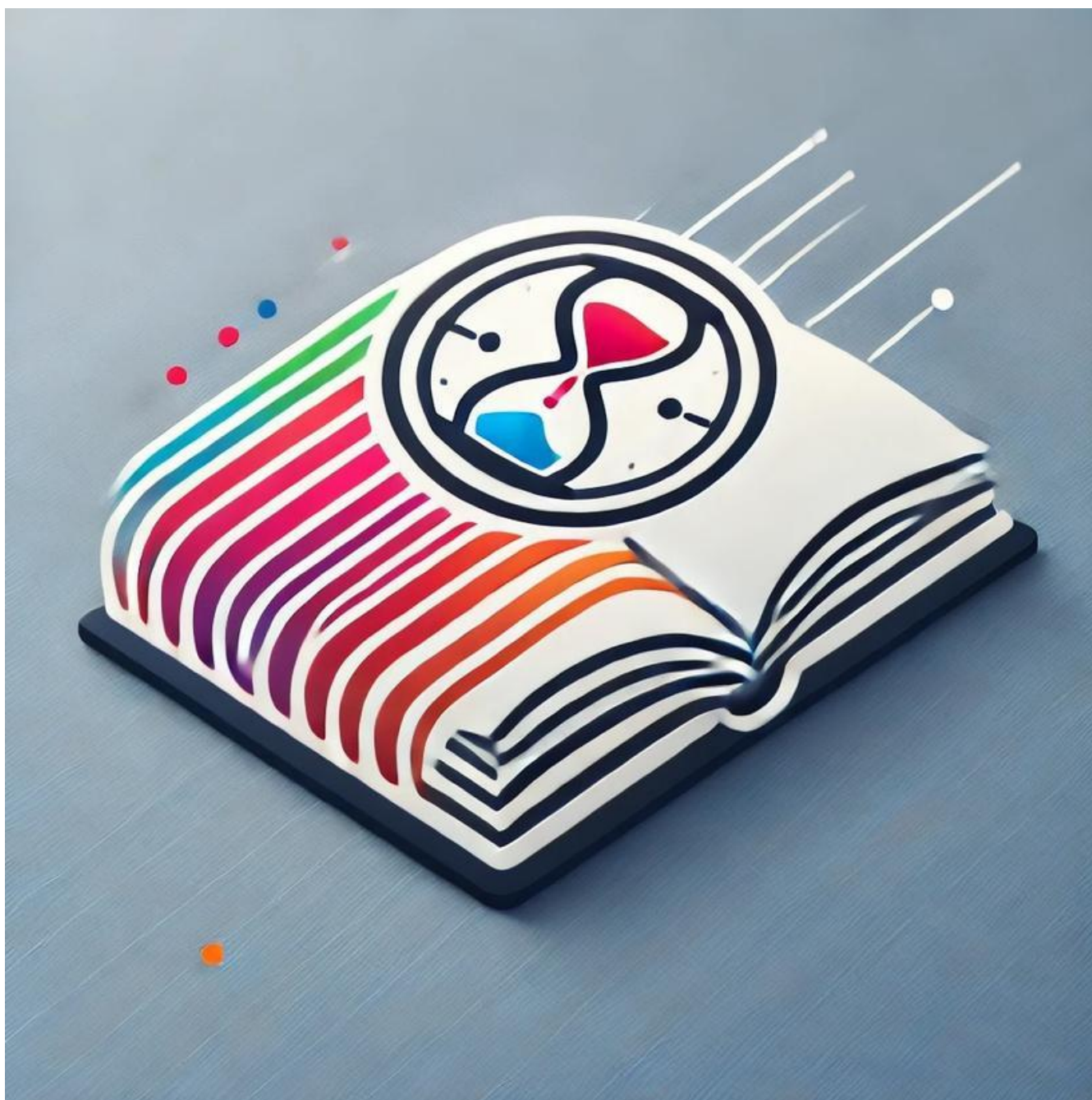


 HuggingFace



 Kaggle





Background

Scientific advances are crucial for addressing global challenges

Pandemics, energy security, climate change, social justice, and ethical AI.

Scientometrics helps track this evolution via publication contents and citation networks.



Challenges

Limited Analytical Scope

- Studies investigated creation [1], diffusion [2], and association [3] of academic knowledge.
- However, many focus on limited timespan [4, 5], venues [6] or particular areas like NLP [7, 8, 9].

Lack of Comprehensive Longitudinal Datasets

- **Ready-to-use datasets that cover both content-level and citation-level**

information across multiple disciplines.

[1] Mapping the research on knowledge transfer: A scientometrics approach.

[2] Rediscovering ACL discoveries through the lens of ACL anthology network citing sentences.

[3] A First Step towards Measuring Interdisciplinary Engagement in Scientific Publications: A Case Study on NLP+ CSS Research.

[4] Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research.

[5] Investigating fairness disparities in peer review: A language model enhanced approach.


[6] Homophily and missing links in citation networks.


[7] Is there really a Citation Age Bias in NLP?

[8] Rediscovering ACL discoveries through the lens of ACL anthology network citing sentences

This Work



 **Dataset:** SciEvo, a continuously updating, ready-to-use dataset of over 2 million academic papers, with detailed metadata such as titles, abstracts, full-text, keywords, subject categories, and citation graphs

 **Analysis:** Longitudinal analysis of scientific terminology, citation patterns, and interdisciplinary exchanges.



Dataset Comparison

Dataset	Size	Text	Metadata	Citation	Tags	Analytic Tools	Disciplines
Li et al. 2024b	1,540	✓	✓	×	×	×	CS
Jurgens et al. 2018	20,000	✓	✓	×	✓	✓	NLP
Chen et al. 2021	0.2M	✓	✓	×	×	×	CS
Li et al. 2022b	0.4M	✓	✓	×	×	×	Physics, Math, CS
Clement et al. 2019	1.5M	✓	✓	×	×	×	multiple
Ginev et al. 2020	1.6M	✓	✓	×	×	×	multiple
Roy et al. 2024	1.7M	✓	✓	×	×	×	multiple
Ours	2.1M	✓	✓	✓	✓	✓	multiple

SciEvo offers the most comprehensive coverage across features, disciplines, and dataset sizes, surpassing other datasets in terms of breadth and depth.

Metrics

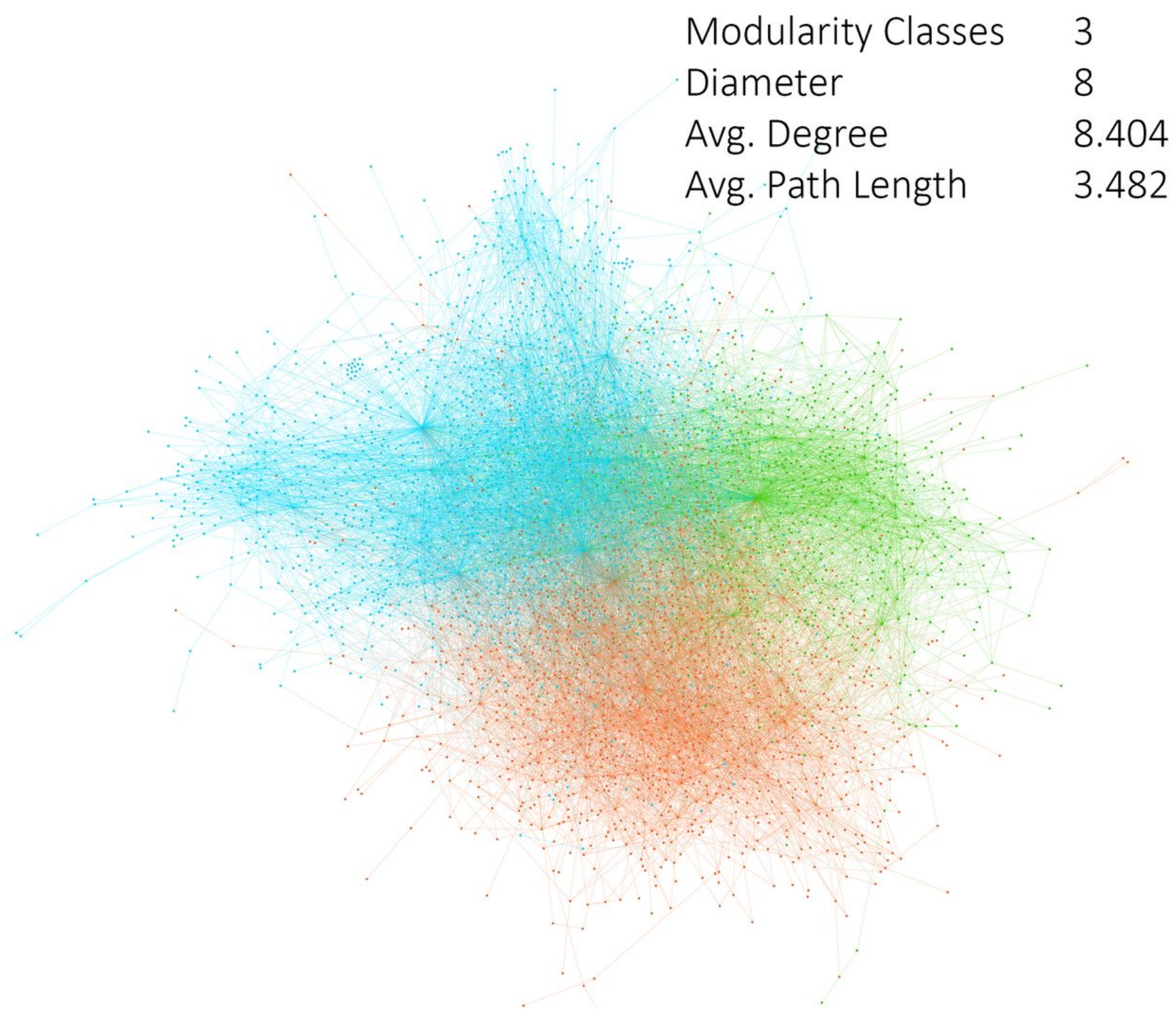
Topical Diversity

- 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.

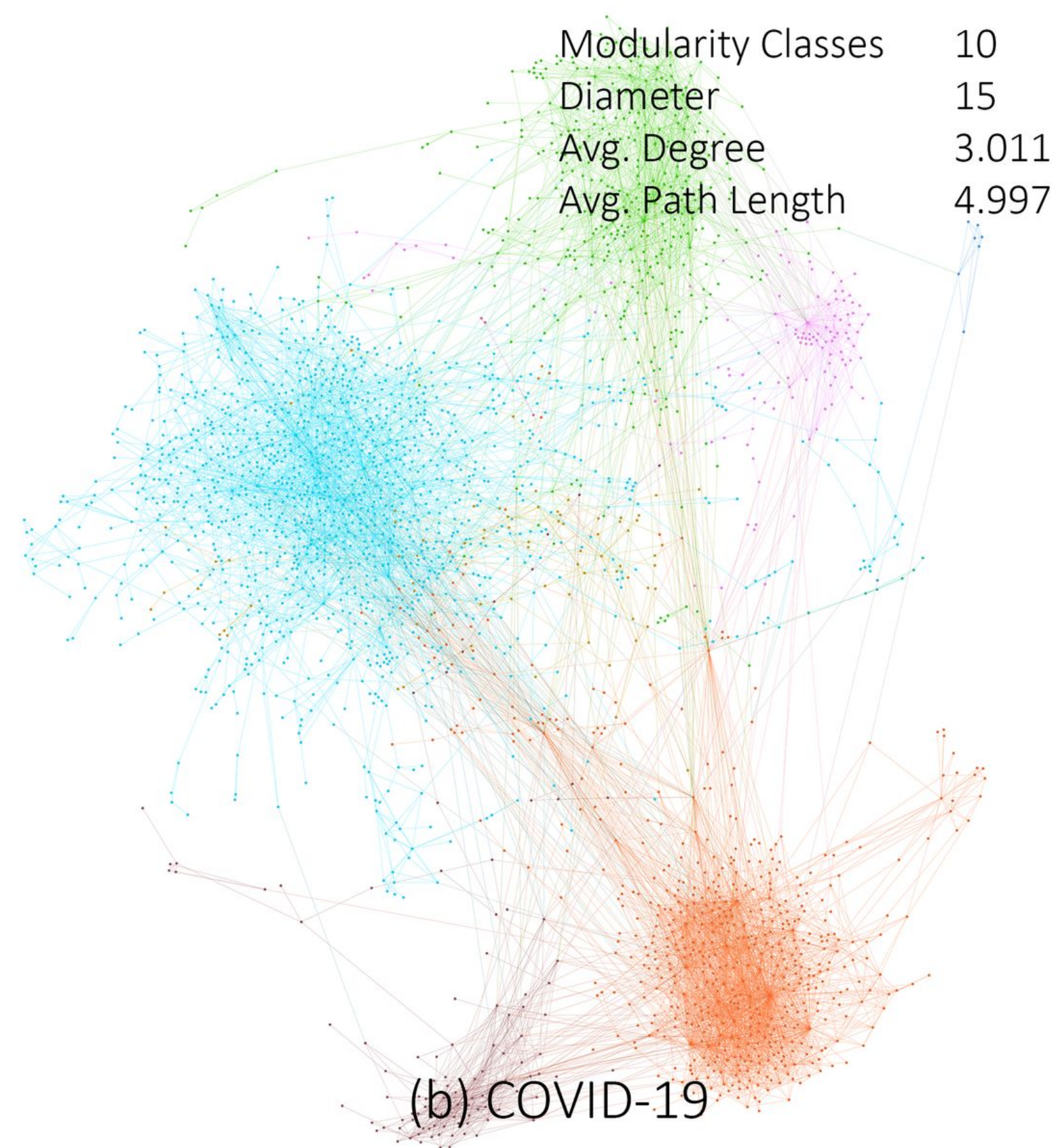
Temporal Diversity

- Reflects the **depth** of citations
- Measures citation distribution across different time periods
- Reflects the depth and profound influence over time



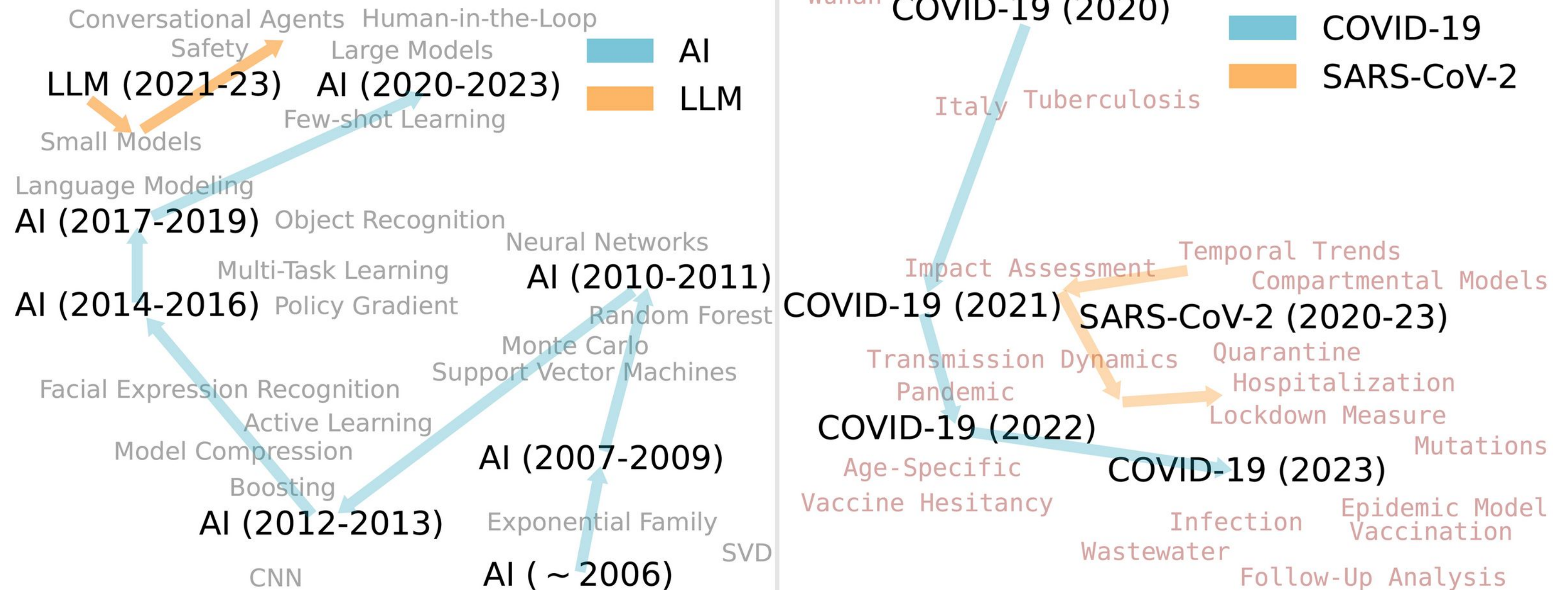


(a) Large Language Models



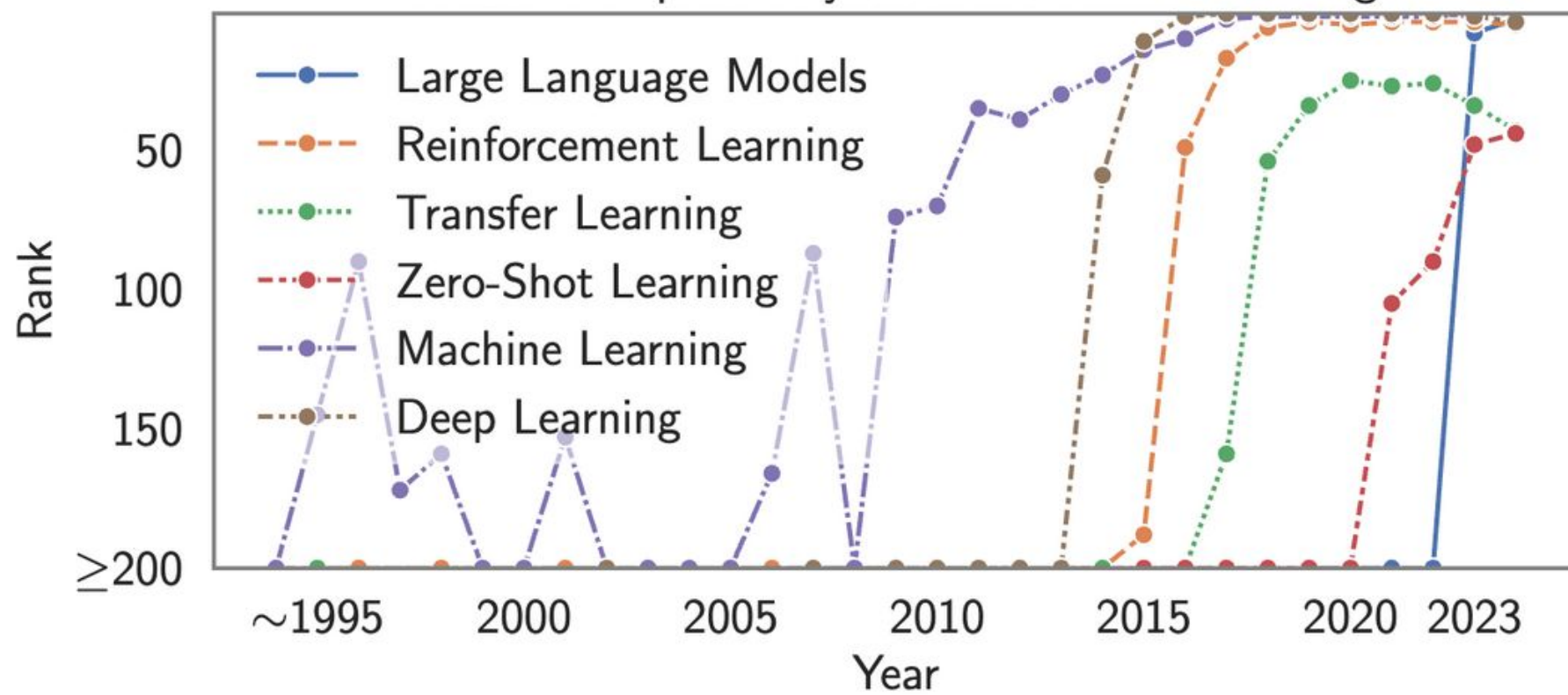
(b) COVID-19

Keyword Trajectories

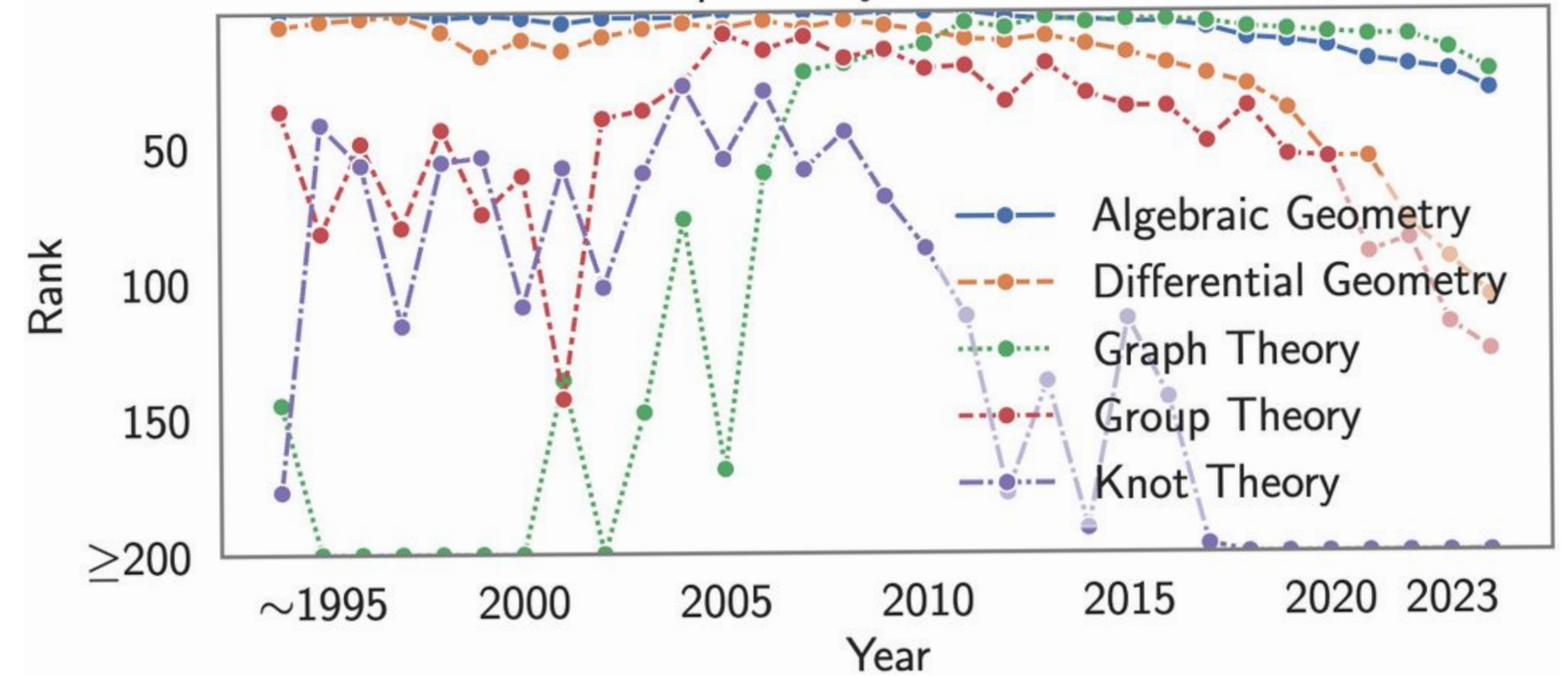


Keyword Ranks

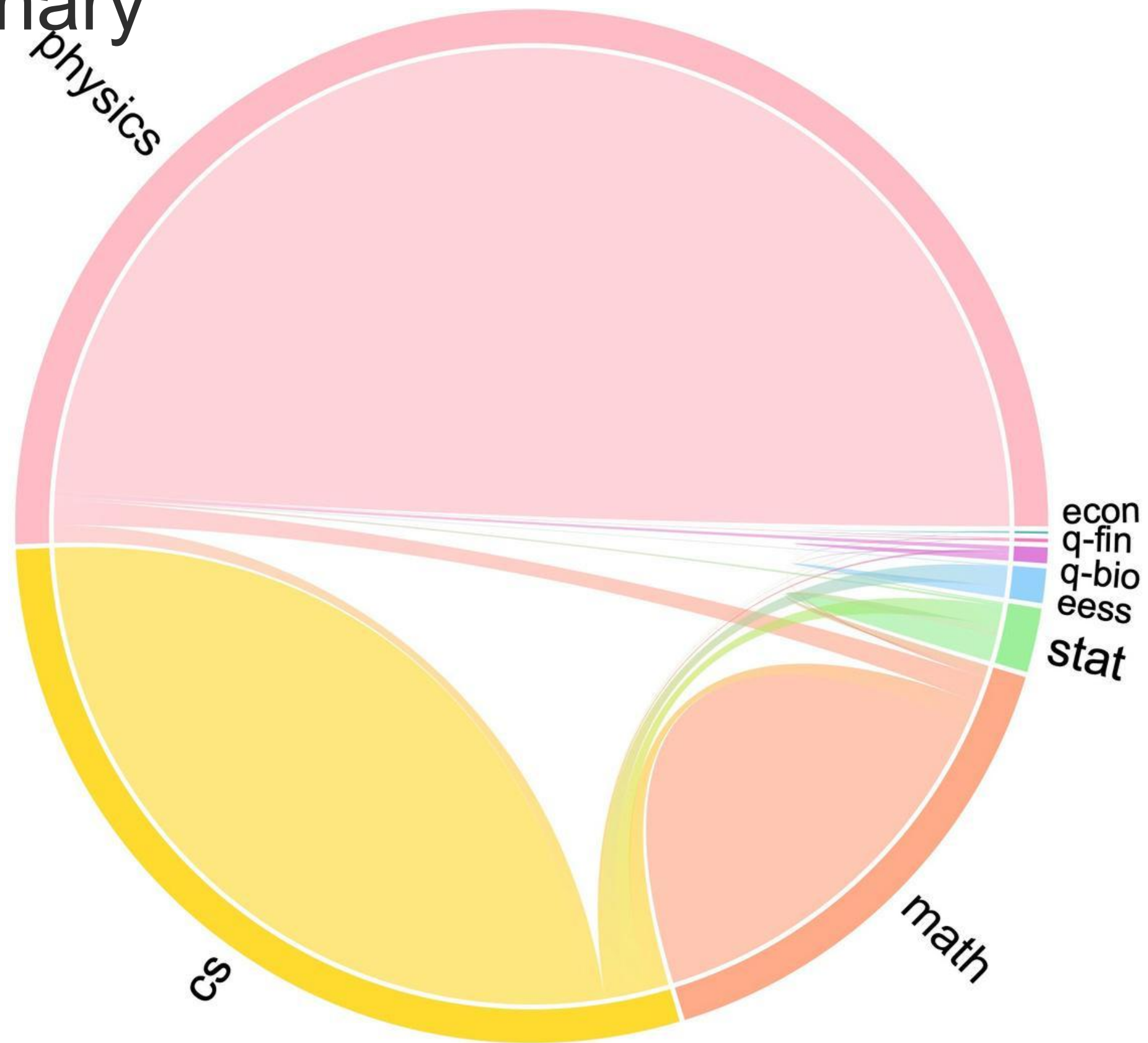
Ranks of Popular Keywords in Machine Learning



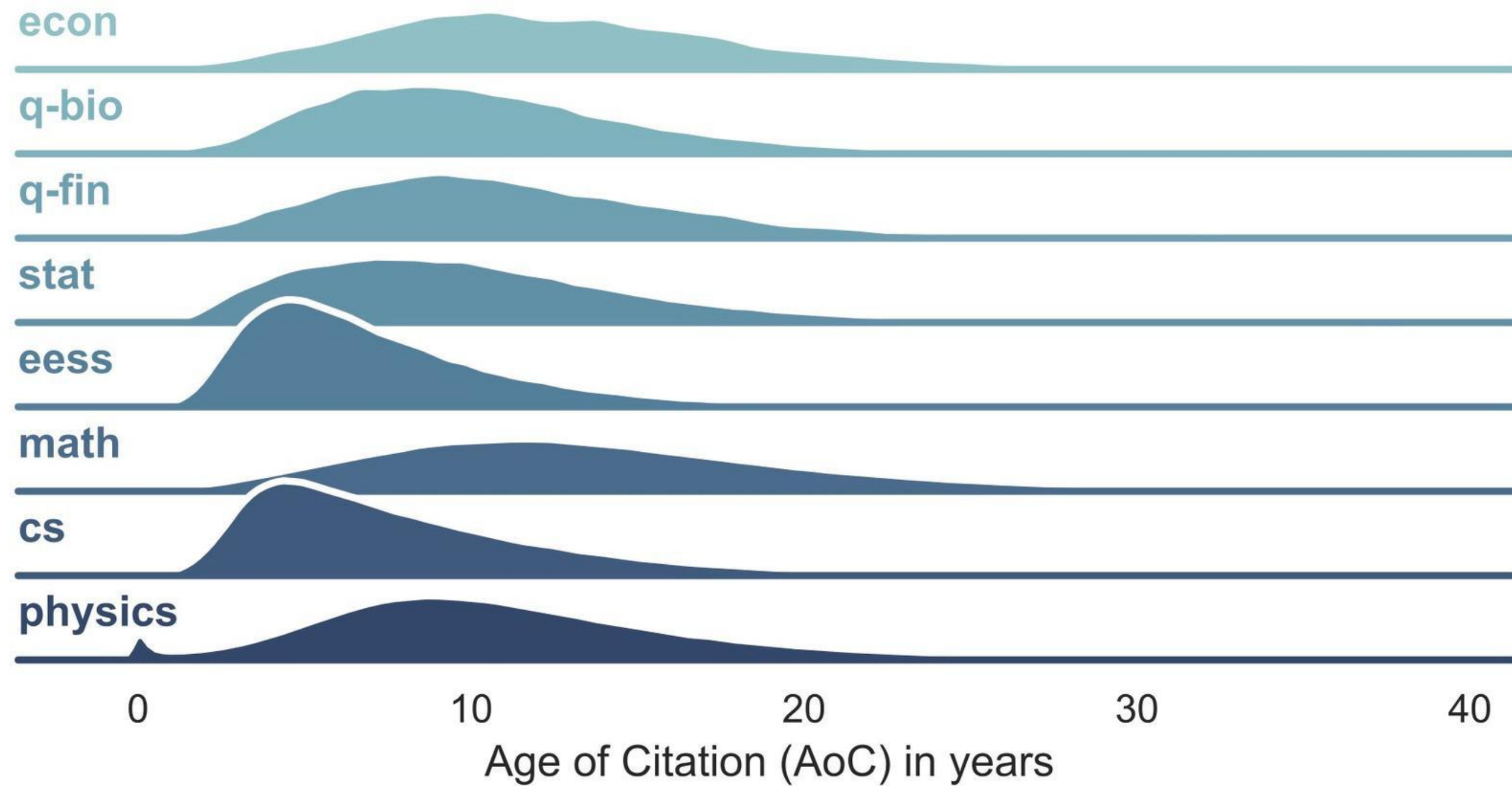
Ranks of Popular Keywords in Mathematics



Cross-disciplinary Citations



Temporal Diversity





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

Yiqiao Jin, Yijia Xiao, Yiyang Wang,
Jindong Wang



WILLIAM
& MARY
CHARTERED 1693



 GitHub



 HuggingFace



 Kaggle

