NetSci 2022

Exploration vs Exploitation of Scientific Fields

Chakresh Kumar Singh

@chakresh_iitgn

https://chakreshiitgn.github.io/

LEARNING
PLANET
INSTITUTE

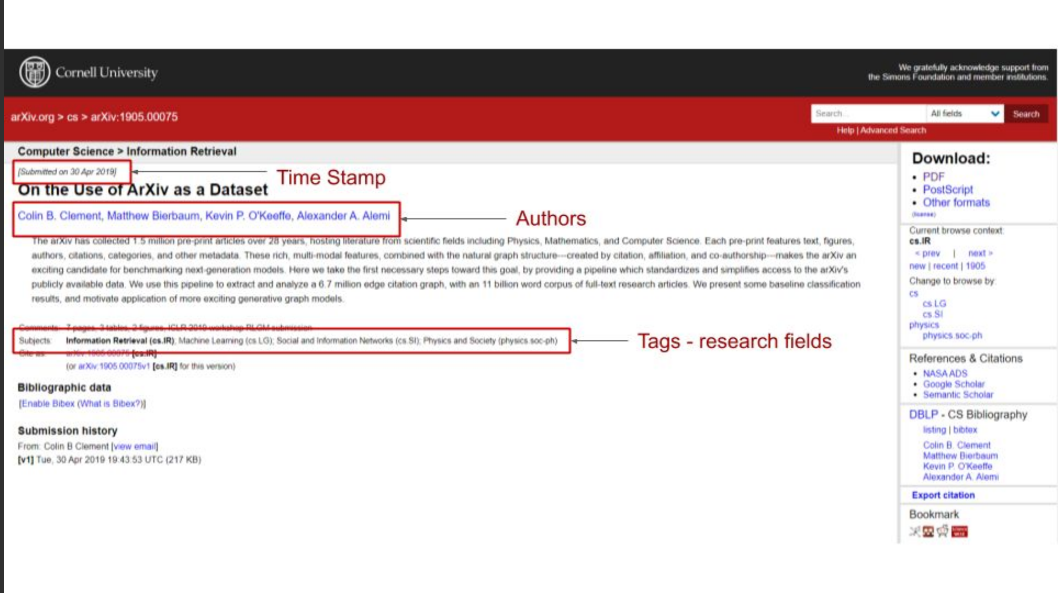
What are the **mechanisms** underlying the knowledge discovery process?

- Rise and Fall of Research Fields
- **Exploration vs Exploitation**
- Research trajectory

arXiv data

- First open-access **preprint repository**
- **30 years** (1986-2018)
- **175 field tags (**Physics, Maths, Computer Science, Finance / Economy, Biology)
- **1.5M articles**
- **50k authors** mapped to unique ORCID ID’s .

Collected using the arXiv API



Cornell University

We gratefully acknowledge support from the Simons Foundation and member institutions.

arXiv.org > cs > arXiv:1905.00075

Computer Science > Information Retrieval

[Submitted on 30 Apr 2019]

Time Stamp

On the Use of ArXiv as a Dataset

Colin B. Clement, Matthew Bierbaum, Kevin P. O'Keeffe, Alexander A. Alemi

Authors

The arXiv has collected 1.5 million pre-print articles over 28 years, hosting literature from scientific fields including Physics, Mathematics, and Computer Science. Each pre-print features text, figures, authors, citations, categories, and other metadata. These rich, multi-modal features, combined with the natural graph structure---created by citation, affiliation, and co-authorship---makes the arXiv an exciting candidate for benchmarking next-generation models. Here we take the first necessary steps toward this goal, by providing a pipeline which standardizes and simplifies access to the arXiv's publicly available data. We use this pipeline to extract and analyze a 6.7 million edge citation graph, with an 11 billion word corpus of full-text research articles. We present some baseline classification results, and motivate application of more exciting generative graph models.

Comments: 7 pages, 3 tables, 2 figures, ICLR 2019 workshop RLGM submission
Subjects: **Information Retrieval (cs.IR)**; Machine Learning (cs.LG); Social and Information Networks (cs.SI); Physics and Society (physics.soc-ph)
Cite as: arXiv:1905.00075 **[cs.IR]**
(or arXiv:1905.00075v1 **[cs.IR]** for this version)

Tags - research fields

Bibliographic data

[Enable Bibex (What is Bibex?)]

Submission history

From: Colin B Clement [view email]
[v1] Tue, 30 Apr 2019 19:43:53 UTC (217 KB)

Download:

- PDF
- PostScript
- Other formats

(license)

Current browse context:
cs.IR
< prev | next >
new | recent | 1905

Change to browse by:
cs
cs.LG
cs.SI
physics
physics.soc-ph

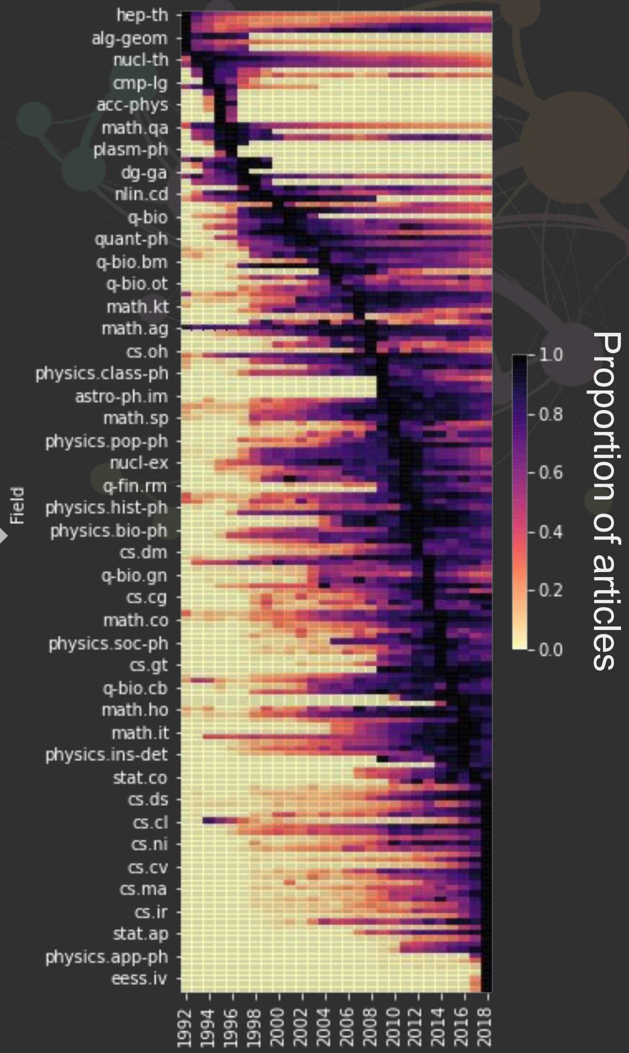
References & Citations

- NASA ADS
- Google Scholar
- Semantic Scholar

DBLP - CS Bibliography
listing | bibtex
Colin B. Clement
Matthew Bierbaum
Kevin P. O'Keeffe
Alexander A. Alemi

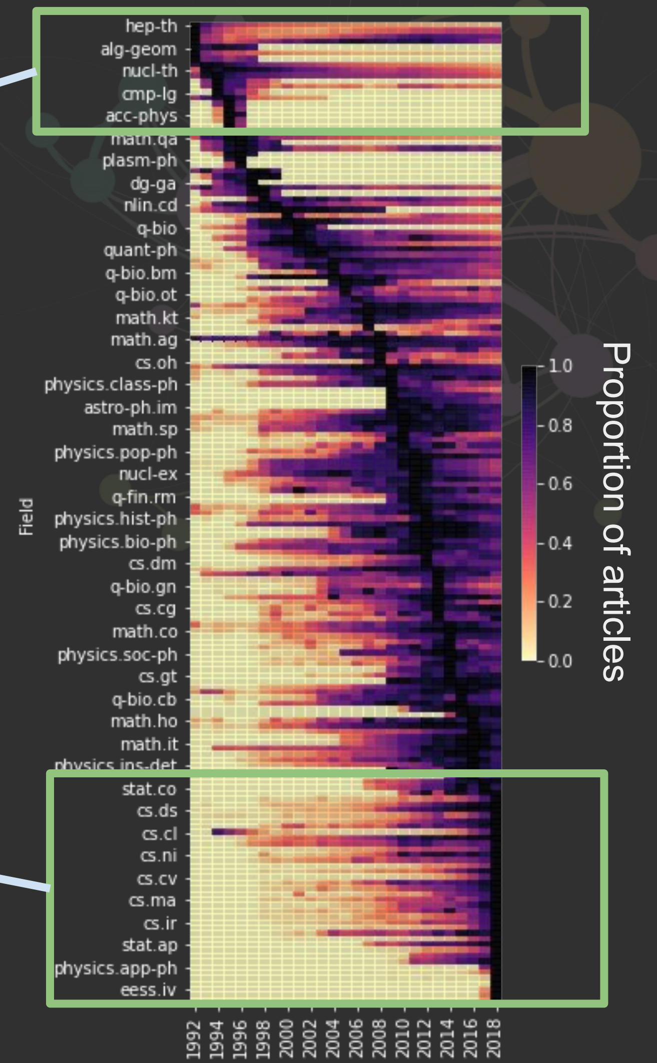
Export citation

Bookmark



Clement, C. B., Bierbaum, M., O'Keeffe, K. P., & Alemi, A. A. (2019). **On the Use of ArXiv as a Dataset**. *arXiv preprint arXiv:1905.00075*.

High Energy Physics, Accelerator Physics, Nuclear Physics, Algebraic Geometry



Computation and Language, Vision and Pattern Recognition, Data Structures, Computation, Applications, Applied Physics

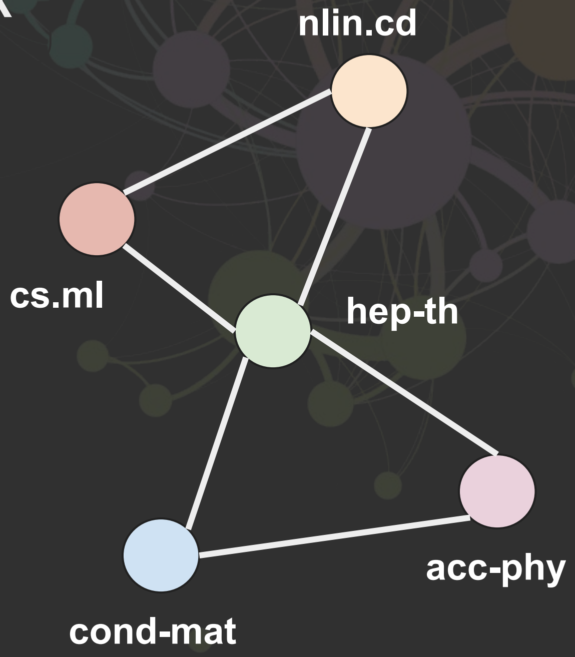
Building the Co-Tag Network

Article i

{ hep-th; nlin.cd; cs.ml }

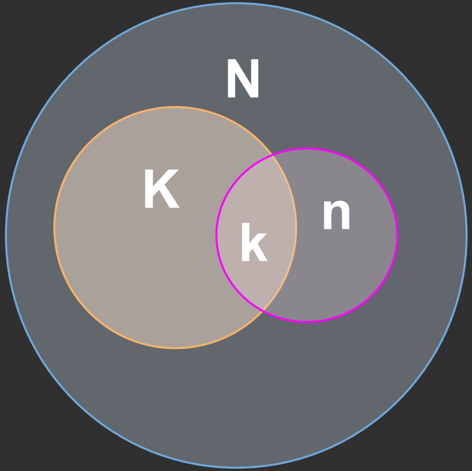
Article j

{ hep-th; cond-mat; acc-phy }



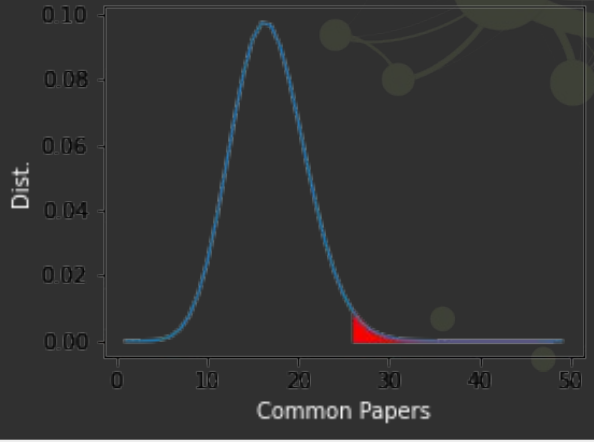
Edge Weight is defined as a function of common papers b/w two field tags

Edge Weight = $-log_{10}(p_{ij})$

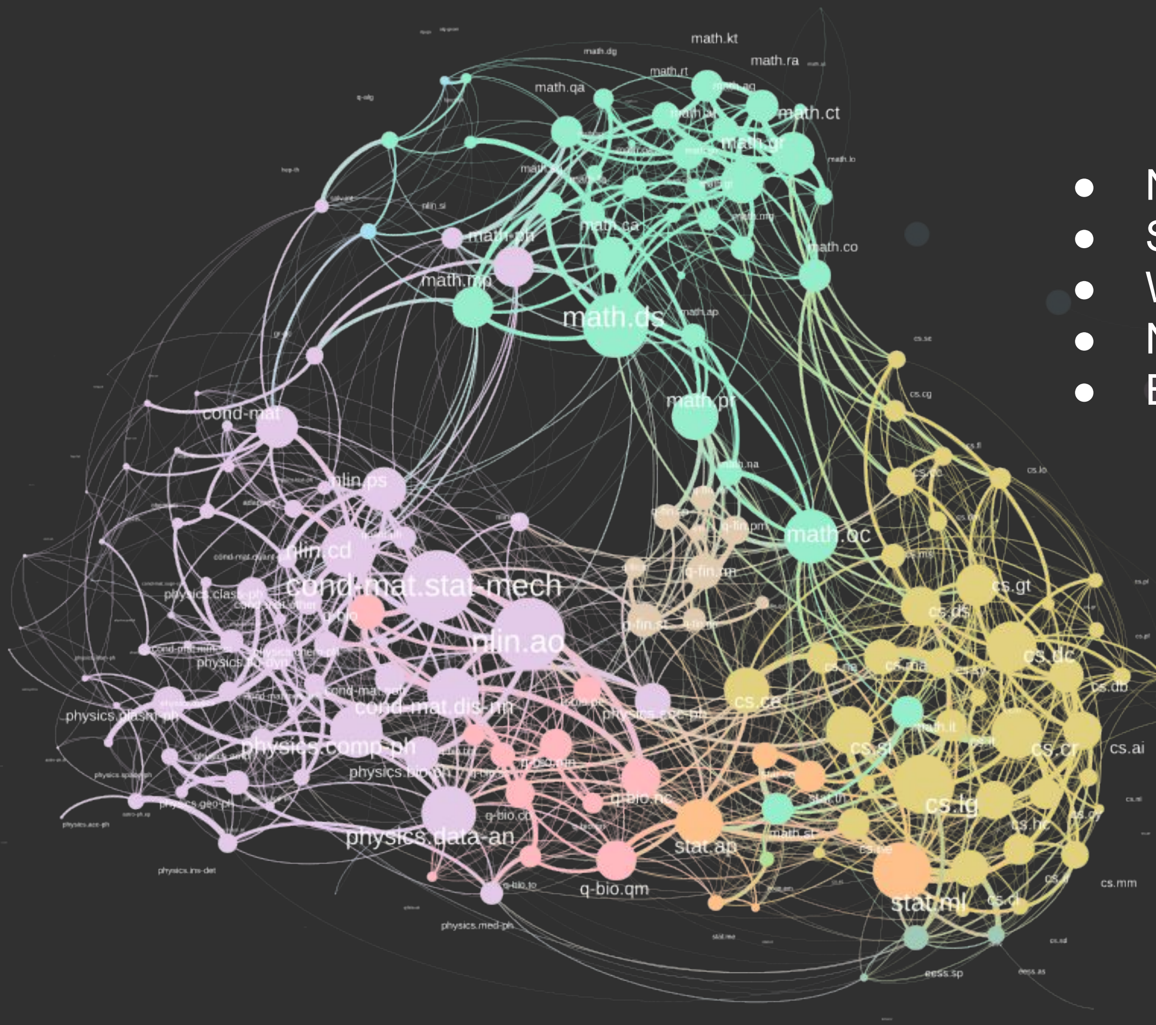


$$p_{ij} = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

- N - Total Articles
- K - Articles in field i
- n - Articles in field j
- k - common articles bw i and j



Note - Here lower p-values are more significant. We eliminate edges with p > 0.01



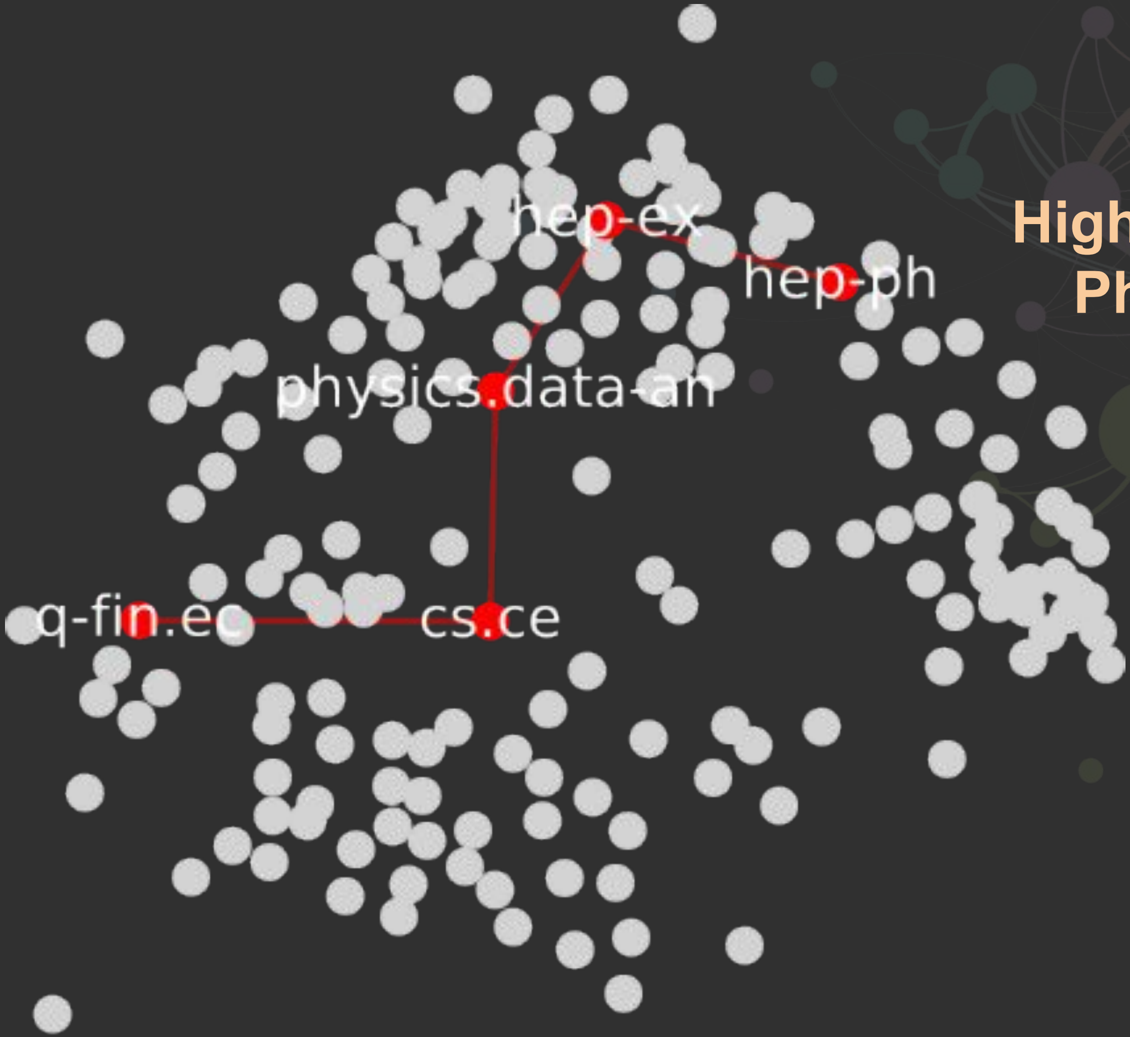
- Nodes are field tags
- Size is proportional to degree
- Weighted and undirected
- Node color is the main research area
- Edge thickness proportional to weight

Cognitive Distance

$$CD_{ij} = min(\textstyle\sum_e \frac{1}{W_e})$$

e : edge on shortest path

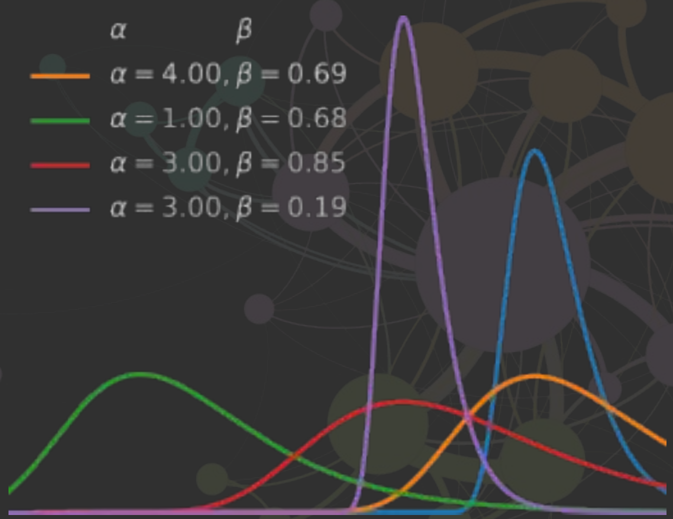
High Energy Physics

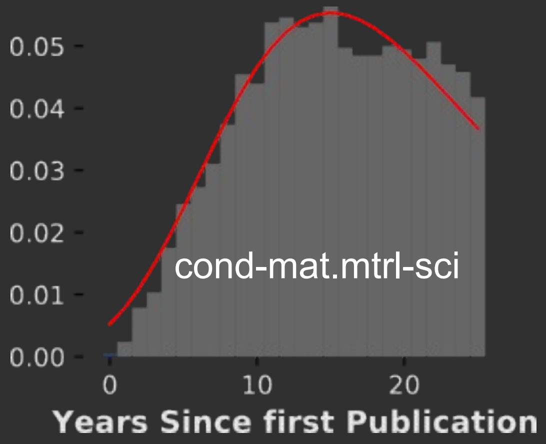


Economics

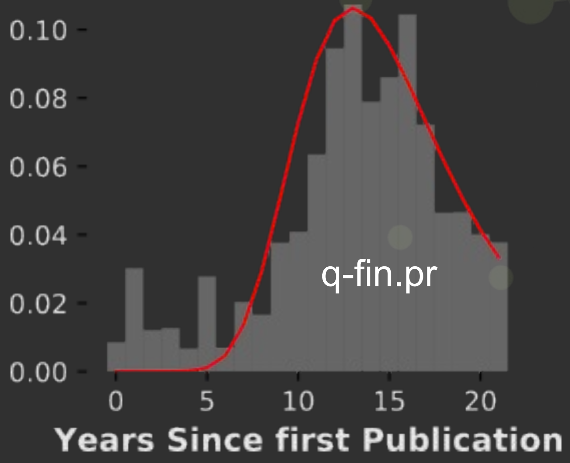
Gumbel distribution function

$$G = \frac{1}{\beta} e^{\frac{-(x-\alpha)}{\beta}} e^{-e^{\frac{-(x-\alpha)}{\beta}}}$$

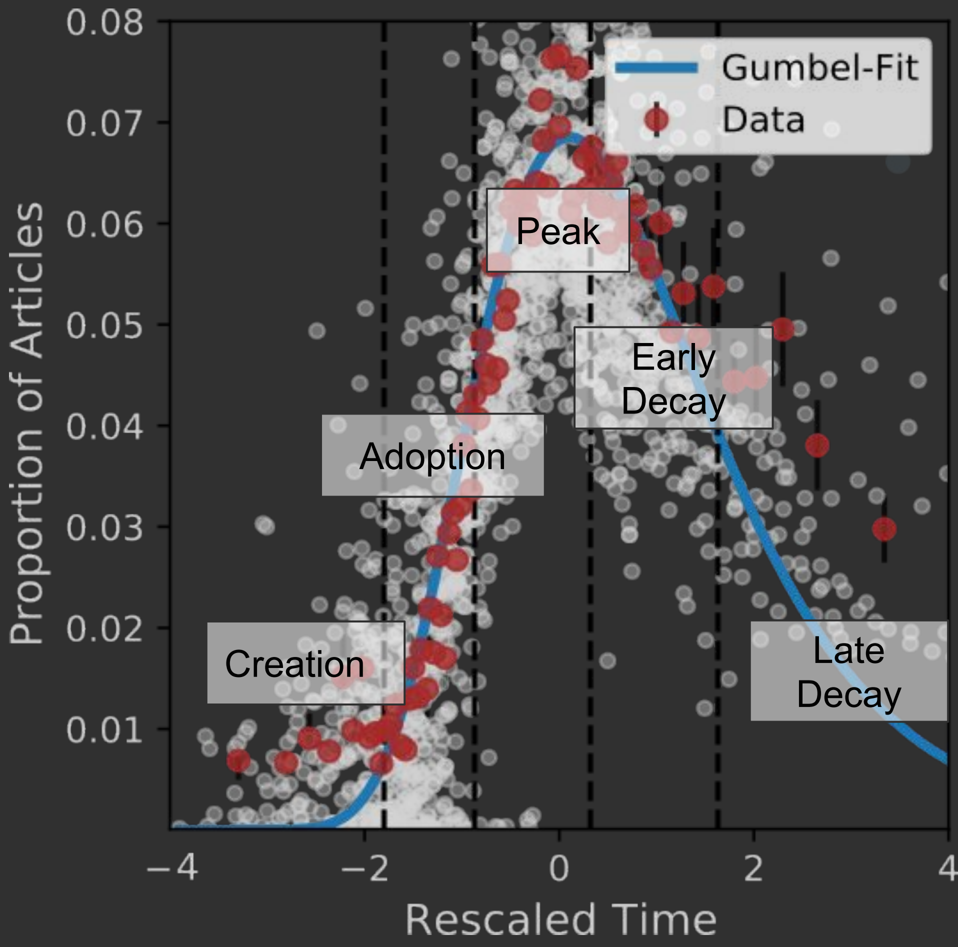




Material Sciences



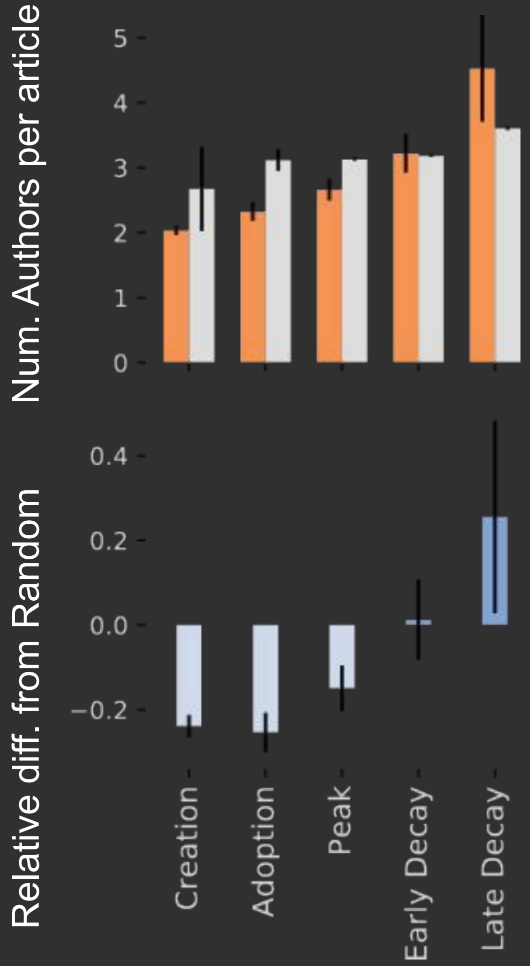
Quantitative Finance (Pricing)

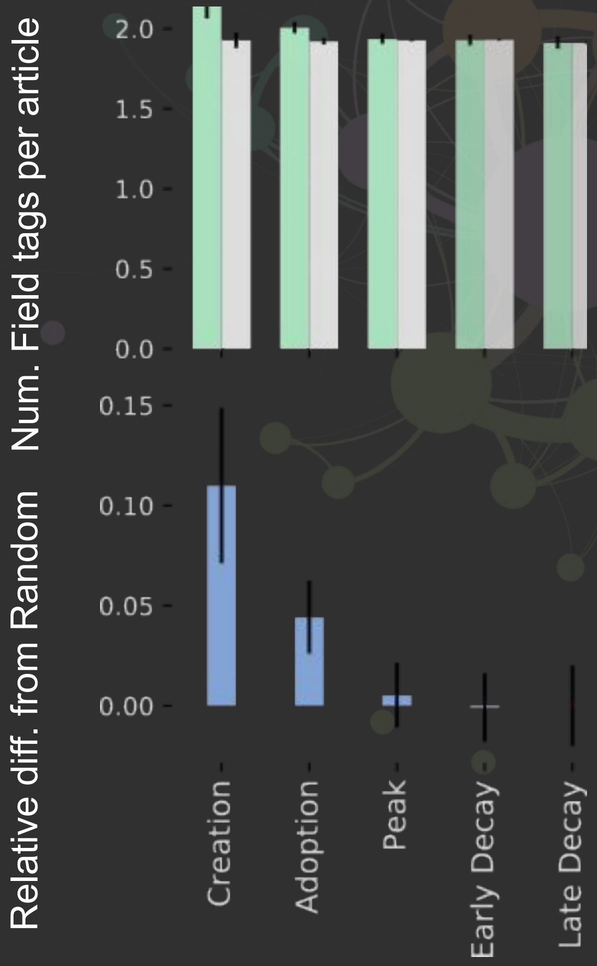


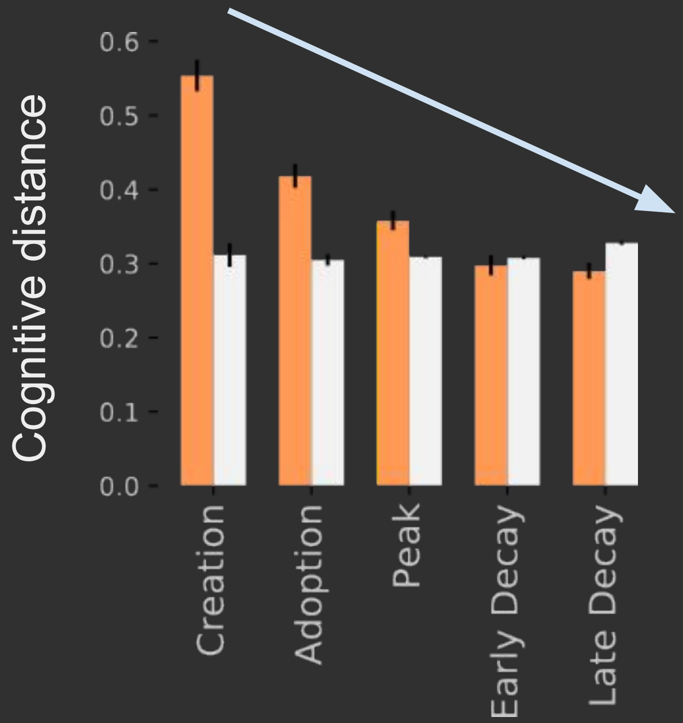
Rescaled time

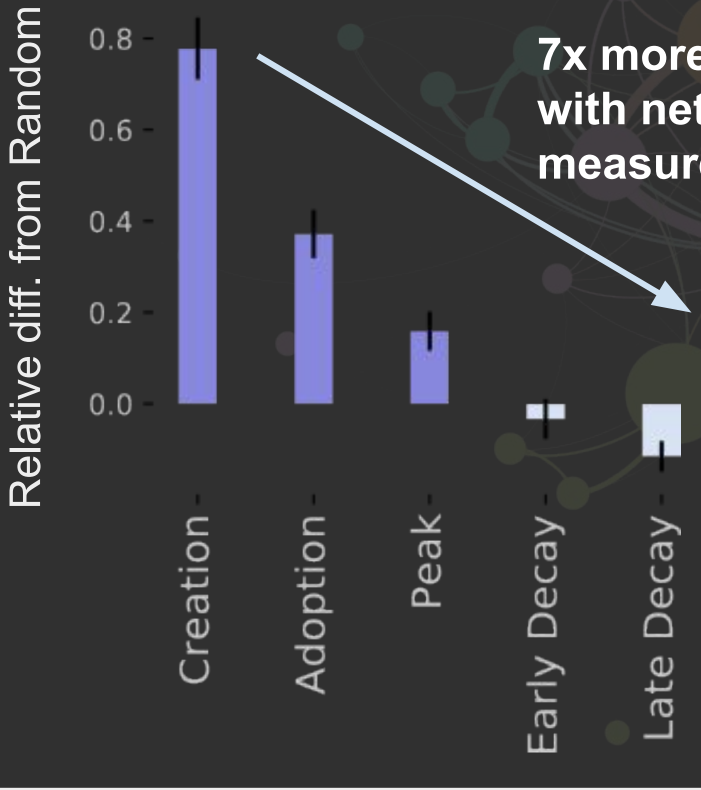
$$t' = \frac{t-\alpha}{\beta}$$

Field stages are defined at 2.5%, 16%,50% and 84% of the fit curve (blue). These number are borrowed from the **diffusion of innovation** literature


Num. Authors per article
5
4
3
2
1
0
Relative diff. from Random
0.4
0.2
0.0
−0.2
Creation
Adoption
Peak
Early Decay
Late Decay


Num. Field tags per article
2.0
1.5
1.0
0.5
0.0
Relative diff. from Random
0.15
0.10
0.05
0.00
Creation
Adoption
Peak
Early Decay
Late Decay





7x more enrichment with network based measure

Researchers in creation phase connect distant fields together (**explore**) whereas in later phases they focus on closely related fields (**exploit**)

- Scientific fields exhibit a universal rise and fall process allowing to define standardised stages of development

- Early stages are enriched with small teams of interdisciplinary authors

- **Network based cognitive distance is a strong marker of early innovation (exploration)**

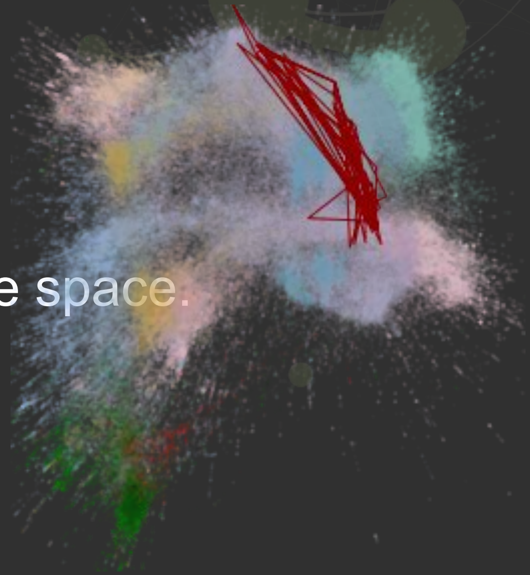
Singh, Chakresh, Emma Barme, Robert Ward, Liubov Tupikina, and Marc Santolini. "**Quantifying the rise and fall of scientific fields.**" *arXiv preprint arXiv:2107.03749* (2021). (In Review)

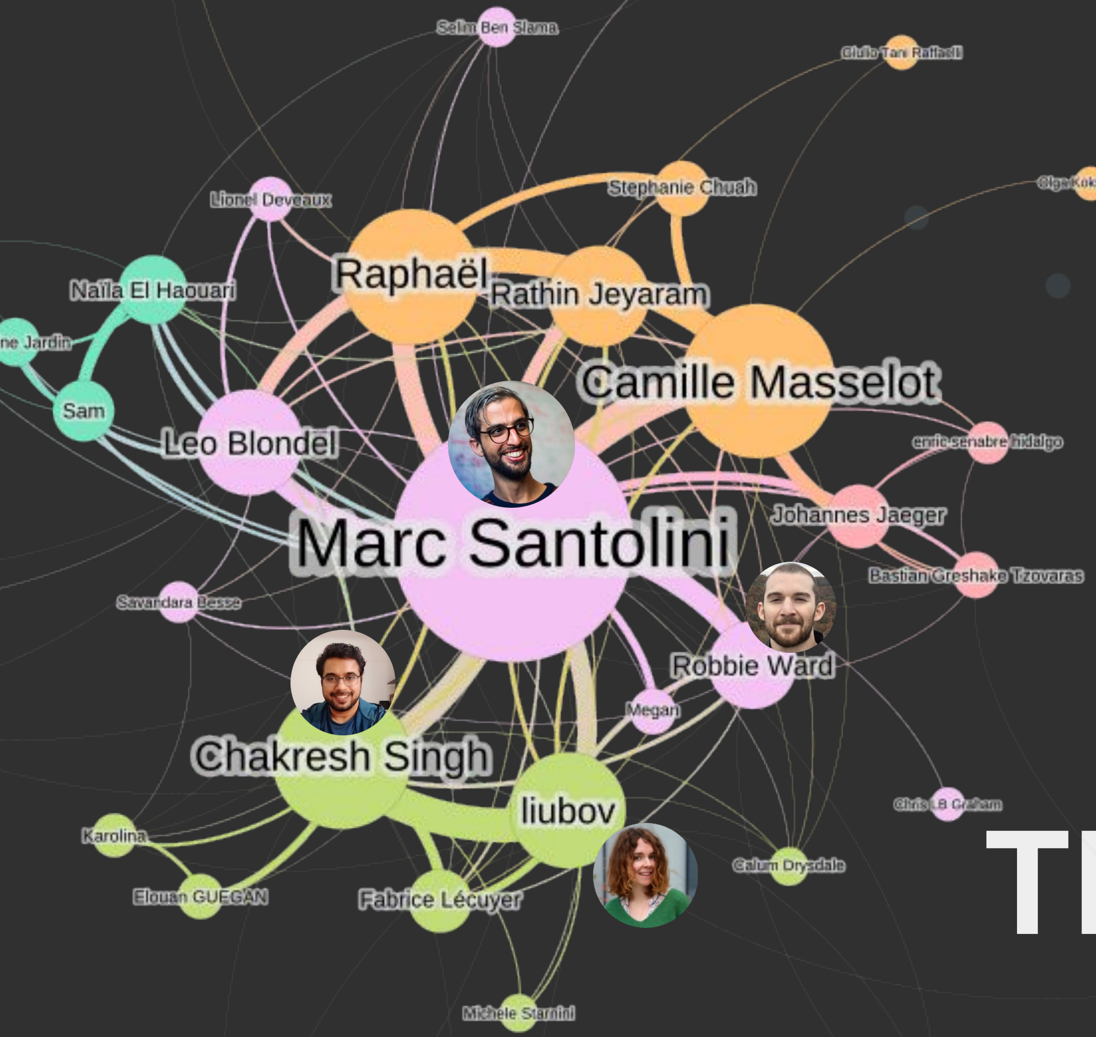
Limitations -

- We focus on fields that were unimodal and >10 yrs **(72)**
- We assume the co-tag network as a static network

Perspectives -

- **Knowledge foraging**:
  Researchers’ trajectories in the knowledge space.
- **Model** the knowledge discovery process
- **Large Scale** Data-sets



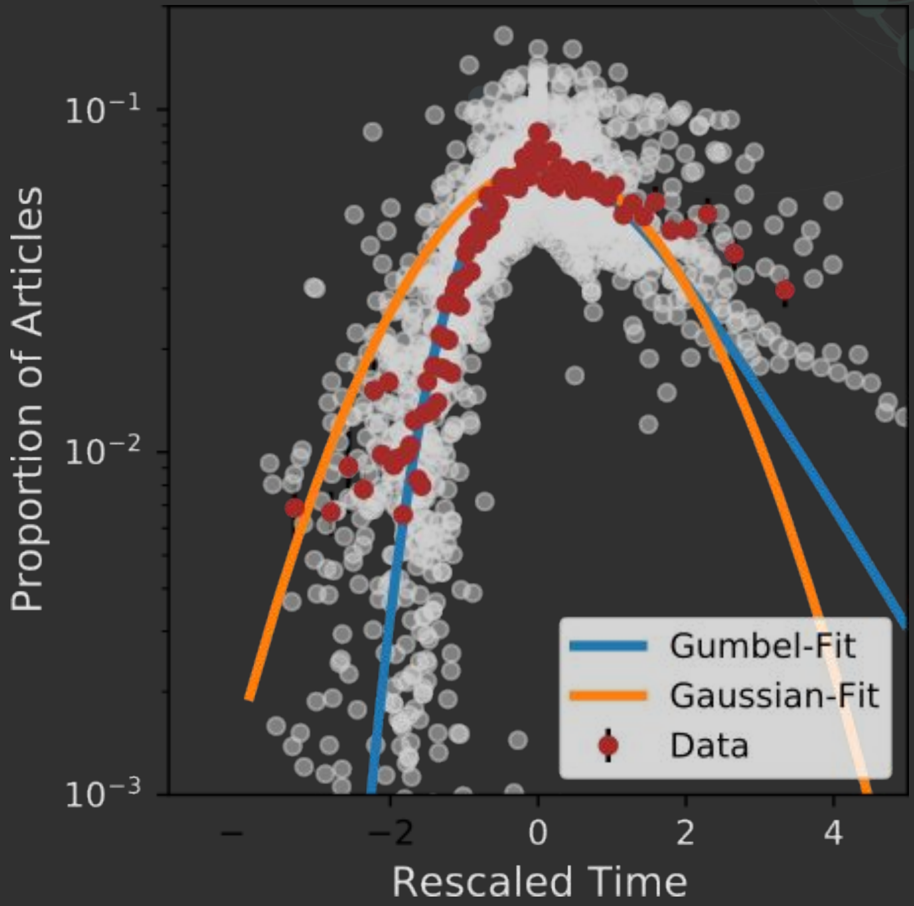


https://interactiondatalab.com/

@InteractionData

THANK YOU

Supplementary Slides


Proportion of Articles
10^{-1}
10^{-2}
10^{-3}
Rescaled Time
Gumbel-Fit
Gaussian-Fit
Data