Scientific Research Trends: Uncovering Patterns of Global Innovation

Raman Pandey, Uzair Ahmed Shah, Jagadish Ishwar Patil
Department of Computer Engineering
Newton School of Technology, Pune, India
Email: raman.pandey@adypu.edu.in, uzair.shah@adypu.edu.in, jagadish.patil@adypu.edu.in

***Abstract*—This project performs a large-scale exploratory data analysis of scientific publications using the arXiv dataset from Cornell University, covering multiple decades of preprints across physics, computer science, and mathematics categories.[Library(2020)] The analysis focuses on publication volume growth, topical shifts between categories, authorship patterns, interdisciplinarity through category co-occurrence networks, and publication practices such as preprint versions and journal references.[Library(2020)] Without fitting any predictive or regression models, the study uses descriptive statistics and visualization to quantify how research output responds to macro shocks, how collaboration structures evolve, and how the role of journals versus preprints is changing.[Library(2020)] The resulting trends highlight a transition from physics-dominated experimentation to computation-driven research, an increase in team-based science, and growing reliance on preprint servers as a primary record of scientific activity.[Library(2020)]**

***Index Terms*—Data mining, arXiv dataset, bibliometrics, collaboration networks, scientific publishing.**

I. Introduction

Scientific publishing has expanded rapidly over the past three decades, driven by digital repositories like arXiv, advances in computation, and evolving funding priorities.[Library(2020)] Understanding how research output, subject categories, and collaboration structures evolve over time is important for funders, institutions, and early-career researchers when identifying emerging directions.[Library(2020)] This project applies exploratory data analysis (EDA) techniques to the arXiv dataset to uncover macro-level patterns in publication volume, topical growth, authorship behavior, interdisciplinarity, and journal adoption.[Library(2020)]

The main contributions of this work are:

- A longitudinal, purely descriptive analysis of yearly publication counts and growth regimes over several decades.[Library(2020)]
- Category-level exploration of rapidly growing areas, highlighting shifts from traditional high-energy and condensed-matter physics toward more computation-centric fields.[Library(2020)]
- Quantitative characterization of authors-per-paper and trends in team size using summary statistics and visualizations.[Library(2020)]
- Construction of a category co-occurrence network to study interdisciplinarity within the physics-related portion of arXiv.[Library(2020)]
- Descriptive analysis of preprint versions and journal references to understand evolving publication practices.[Library(2020)]

II. Methodology

A. Dataset Description

The analysis uses the arXiv metadata dataset released on Kaggle by Cornell University, which covers submissions from 1985 onwards across categories such as high-energy physics, quantum physics, condensed matter, astrophysics, mathematics, and several computer science areas.[Library(2020)] Each record includes an identifier, title, authors, abstract, submission date, primary and secondary categories, version information, and an optional journal reference field.[Library(2020)]

B. Preprocessing

Raw JSONL files were loaded into data frames, with submission dates converted to integer years for temporal aggregation.[Library(2020)] Subject category labels were kept in arXiv's standard notation and multi-category papers were expanded so each category participates in co-occurrence counts.[Library(2020)] Author strings were split into lists to compute authors-per-paper statistics without full author disambiguation, and basic text normalization of titles and abstracts (lowercasing, tokenization, and stopword removal) enabled keyword counting after dropping malformed records.[Library(2020)]

C. Feature Engineering

Temporal features included yearly paper counts and simple year-over-year growth measures at both global and category levels.[Library(2020)] Authorship features captured the mean and median number of authors per paper per year and distributions within selected periods.[Library(2020)] Interdisciplinarity features were derived by building an undirected graph where nodes represent categories and weighted edges connect categories that co-occur on the same paper, with weights equal to co-occurrence counts.[Library(2020)] Additional features summarized version numbers and the fraction of papers with non-empty journal references by year.[Library(2020)]

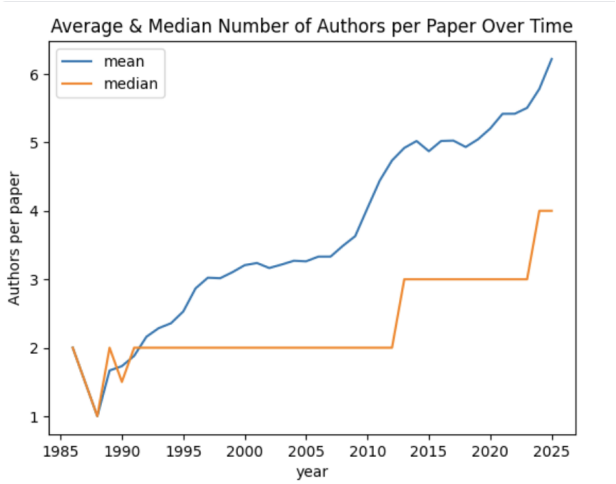

Fig. 1. Average (blue) and median (orange) number of authors per paper over time.

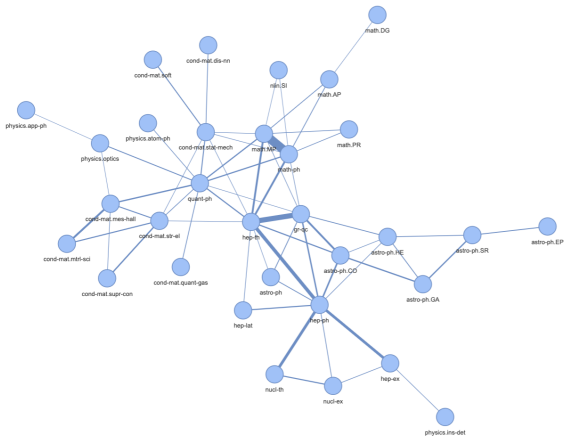

Fig. 2. Category co-occurrence network for different arXiv categories.

D. Analysis Approach

The project is strictly exploratory: no regression, classification, clustering, or forecasting models were trained.[Library(2020)] Pandas was used for aggregation, Matplotlib and Seaborn for visualizing time series and distributions, and NetworkX for constructing and plotting category co-occurrence networks.[Library(2020)] Interpretations are guided directly by these descriptive plots and statistics together with contextual knowledge of major scientific and economic events.[Library(2020)]

III. Results

A. Yearly Publication Dynamics

The yearly number of submissions shows clear long-term growth with recognizable phases, including a slow-down around the late-2000s financial crisis, a rebound as large experimental projects came online, a plateau in 2020 aligned with COVID-related disruptions, and renewed growth in subsequent years.[Library(2020)] These shifts illustrate how both macroeconomic conditions and scientific breakthroughs influence submission behavior on arXiv.[Library(2020)]

B. Authorship Patterns

The mean number of authors per paper increases steadily from roughly two in the late 1980s to more than six by 2025, while the median remains at two for many years before stepping to three and then four authors in the most recent period.[Library(2020)] This divergence between mean and median indicates a coexistence of very large collaborations with a broader trend toward larger typical teams, moving beyond the classic advisor–student pattern.[Library(2020)]

C. Interdisciplinarity and Category Network

The category co-occurrence network built from physics-related submissions reveals a dense core in which high-energy physics, quantum physics, general relativity, condensed matter, and several mathematics categories are tightly connected.[Library(2020)] Thicker edges between particular category pairs correspond to frequent co-labeling on the same papers, highlighting interdisciplinary regions where methods and problems are shared across subfields.[Library(2020)]

D. Preprints, Versions, and Journals

Descriptive statistics on version counts show that most arXiv entries appear only as a first version, with steadily fewer papers updated to second and higher versions.[Library(2020)] The yearly fraction of submissions that carry a journal reference decreases over time, reflecting a combination of publication lag and a cultural shift in which many authors treat arXiv itself as their main dissemination and archiving channel.[Library(2020)]

IV. Discussion

The exploratory results indicate that research within the arXiv ecosystem is becoming more collaborative, more interconnected across subfields, and more dependent on open preprint infrastructure.[Library(2020)] Rising authors-per-paper and the structure of the category network both point toward increasing specialization and coordination across multiple research groups.[Library(2020)] At the same time, the decline in journal references suggests a gradual decoupling between preprint dissemination and traditional journal publication, raising questions about peer review, curation, and long-term preservation of the scientific record.[Library(2020)]

V. Conclusion

This project shows that straightforward exploratory analysis of the Kaggle arXiv dataset can uncover meaningful, interpretable patterns in the evolution of physics-related and allied research, including growth in publication volume, changes in collaboration size, and shifts in interdisciplinarity.[Library(2020)] These trends highlight the shift toward computation-heavy, team-based science and preprint-first dissemination, creating both opportunities for

faster progress and challenges for evaluation and credit assignment.[Library(2020)]

REFERENCES

[Library(2020)] C. U. Library, "arxiv dataset," https://www.kaggle.com/datasets/Cornell-University/arxiv, 2020, accessed: 2025-12-06.

APPENDIX

A. Repository Link

The complete analysis pipeline, Jupyter notebook, and generated visualizations are available in the public repository:[1]

https://github.com/raman976/Scientific-Research-Trends

B. Repository Structure

- **project.ipynb**: Main notebook containing data loading, preprocessing, feature engineering, and all plots used in this paper.[Library(2020)]
- **README.md**: Instructions for setting up the environment, running the notebook, and regenerating all statistics and figures from the raw Kaggle dataset.

[1] https://github.com/raman976/Scientific-Research-Trends/blob/main/project.ipynb