STAT 405/605 Presentation Group 11

Gabe Gress, Minyue Hu, Tina Dantono, Kristin West

11/17/2020

Introduction

- arXiv, hosted by Cornell University, is a web-based open-access archive for over 1.7M scholarly articles
- The documents in arXiv range from 1996 to 2020
- Our research debated whether different authors have different propensities for research topics over time
- We expected the final results to show that certain research fields are of greater interest over time

Introduction

- We also considered the following issues:
  - which year has the greatest number of published papers on average
  - the number of papers published in each arXiv category per year
  - the geographical distribution of the publishers and the types of papers they published
  - the universities to which the researchers belong,
  - the influencing factors on the papers (i.e. funding)
  - 6 categories: physics, computer science, math, statistics, economics, electrical engineering and systems science. Categories can also be combined.

Background

- arXiv collects research papers on the topics of physics, mathematics, computer science, quantitative biology, quantitative finance, statistics, electrical engineering and systems science, and economics.
- As of October 2016, the submission rate has exceeded 10,000 articles per month.
- We chose to study the literature information on the arXiv website because the existence of this database is one of the factors that created the open access movement in the scientific publishing industry.

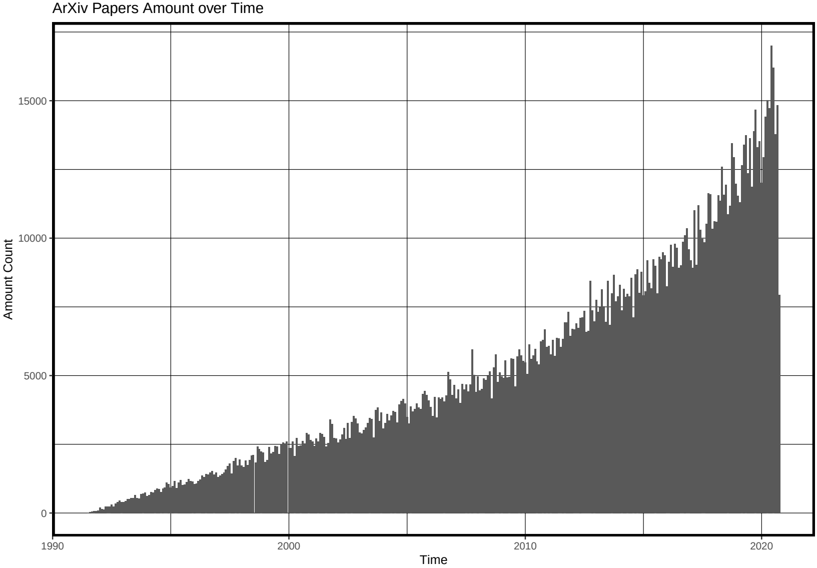
Why?

- arXiv accepts publications free of charge, meaning it is most likely to fairly represent all scientific contributions
- Most preprints are on arXiv, so it often can represent trends in science

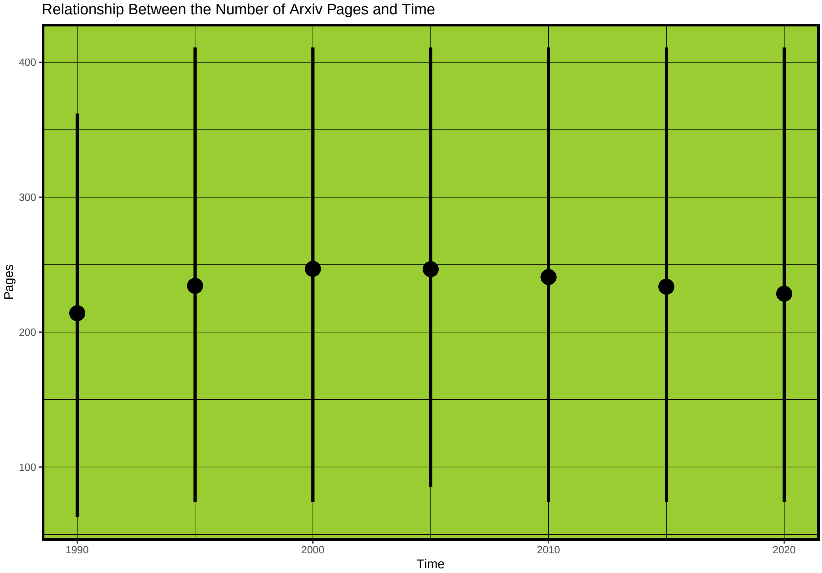
Secondary Dataset

- Universities & Colleges Dataset from the Department of Homeland Security

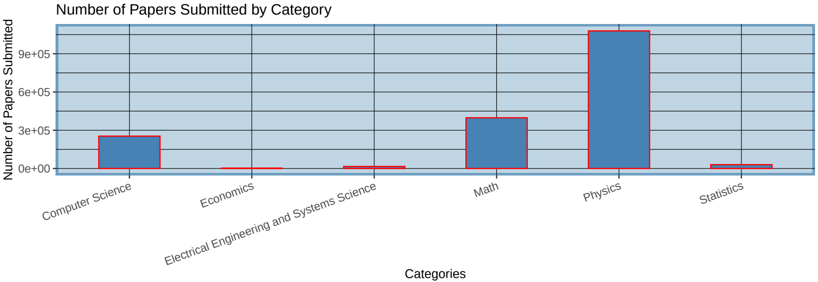
Growth of arXiv

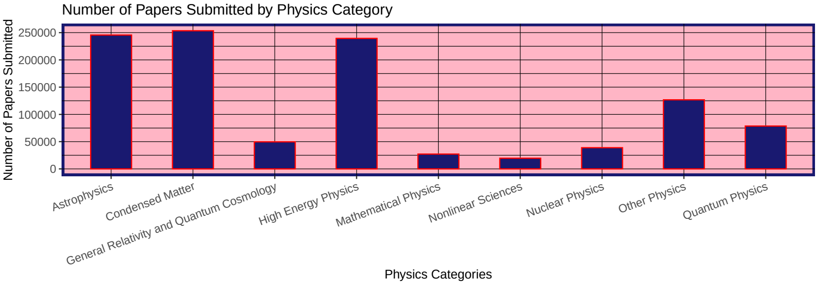


Trends in Format

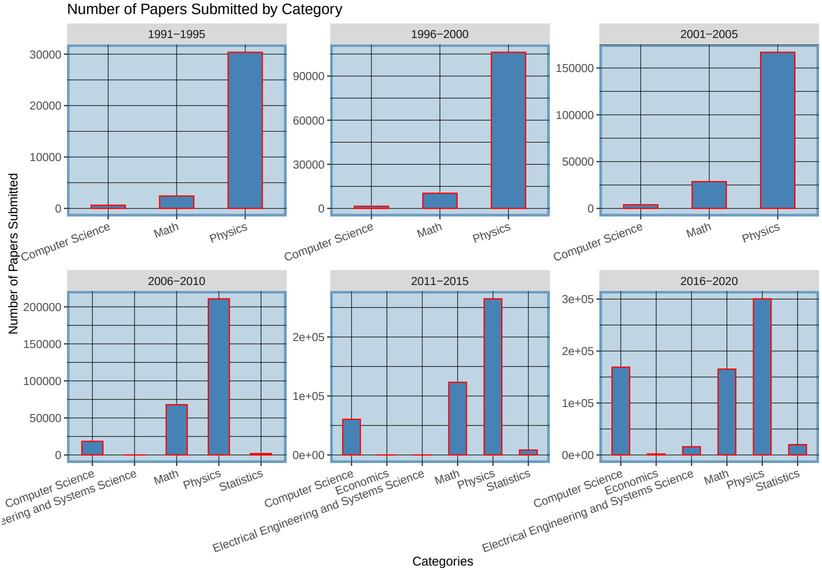


Overall Distribution

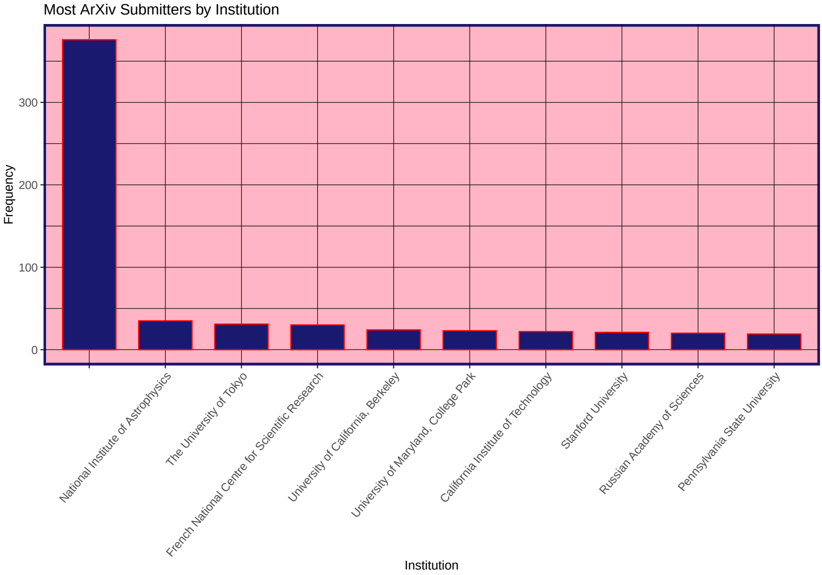




Distribution Over Time



Active Institutions in Science



Most ArXiv Submitters by Institution

Conclusion

- Graphs depicted are tip of the iceberg in terms of interesting trends
- Worldwide universities huge contributors of scientific journalism
- Physics originally dominated publishings, but seeing rapid growth in other fields

Future Work

- Expand correlation dataset between authors and respective universities
  - Use GIS data to identify active research in world
  - Observe global statistics about authors
- Correlate university funding with publications to observe effect of wealth
- Identify growths in key words in papers over time to identify latest topics of interest

Questions?