# Project CURA: Author Collaboration Network from Physics Papers

Yao Yuguang and Duc Khoi LE
Ecole Polytechnique
yuguang.yao@polytechnique.edu

May 2025

**Abstract**

In this project, we present a pipeline for constructing a collaboration graph from a dataset of physics research papers. Our goal is to identify co-authorship patterns and community structures in the research community. The process includes metadata extraction, author name normalization, graph construction with authors as nodes and collaborations as weighted edges, and community detection. This project lays the foundation for further analysis on scientific influence and interdisciplinary collaboration in the physics domain.

## 1 Introduction

Collaboration is a key driver of scientific progress. In theoretical and experimental physics, researchers often co-author papers, forming complex networks of cooperation. Project CURA aims to map these interactions by building a collaboration graph from a dataset of physics papers, using authors as nodes and co-authorships as edges.

This report presents the methodology used to collect, clean, and represent this data as a graph. It also outlines future plans for extending the project to include citation analysis and influence metrics.

## 2 Data Collection and Parsing

Our dataset consists of metadata extracted from physics papers hosted on repositories such as arXiv. Each paper provides information including the title, list of authors, submission date, and journal reference.

An example metadata entry is shown below:

```
Title: Discrete and Continuum Virasoro Constraints...
Authors: Waichi Ogura
arXiv ID: hep-th/9201018
```

We parse this information to extract:

- The list of authors for each paper
- Paper identifiers (e.g., arXiv IDs)
- Other metadata such as title and submission date (optional for this project phase)

## 3 Data Cleaning and Normalization

To ensure accurate graph construction, we implemented basic normalization:

- Standardized author name formats (e.g., ”J. Smith” vs ”John Smith”)
- Filtered out duplicate entries
- Removed papers with missing or corrupted metadata

While full name disambiguation is out of scope for this stage, future work may use author identifiers (e.g., ORCID) to improve precision.

## 4 Graph Construction

We represent the collaboration structure using a graph $G = (V, E)$, where:

- $V$: the set of all unique authors
- $E$: the set of edges where each edge $(u, v)$ indicates co-authorship
- Edge weights indicate the number of co-authored papers

The graph is implemented using the NetworkX Python library. An edge is added between every pair of authors who co-wrote a paper. For example, if three authors collaborated on one paper, three undirected edges are created: $(A, B), (A, C), (B, C)$.

## 5 Community Detection

To explore underlying structure, we apply community detection algorithms (e.g., Louvain method) to identify clusters of researchers. These clusters often correspond to research subfields or collaborative groups.

The modularity score is used to evaluate the quality of the detected communities.

## 6 Preliminary Results and Visualizations

Using a sample dataset, we successfully extracted a collaboration graph with thousands of nodes and edges. Below is a sample visualization of a subgraph (insert figure here if available).

Key observations:

- The largest connected component includes the most prolific authors
- Smaller communities often correspond to focused research topics

## 7 Future Work

This project can be extended in multiple directions:

- Integrating citation data to compute author influence (e.g., PageRank)
- Performing temporal analysis to study evolution of collaborations
- Improving author disambiguation using external identifiers
- Comparing community structures across disciplines (e.g., theory vs. experiment)

## 8 Conclusion

We have developed a working pipeline to extract, clean, and analyze co-authorship data from physics papers. The collaboration graph provides a powerful tool for understanding the structure and dynamics of research communities.