Authorship Networks: How Collaborator Citations Predict Researcher Success

Josef Liem, Frannie Cataldo, & An Adhikari

April 2025

1 Overview

In this project, we intend to investigate and compare the structure of collaborator networks across academic disciplines. In the field of information retrieval, collaborative practices can vary wildly given particular academic domains - from disciplines in the humanities, which tend toward fewer coauthors to methodologically technical domains, such as biochemistry or medicine, where intense collaboration is expected [2]. Understanding the structure and properties of these networks is essential in gauging intra- and interdisciplinary communication among prominent researchers, as well as potentially identifying researchers who collaborate as a result of working within growing sub-disciplines. Thus, understanding how researchers collaborate over time is critical in creating insights into impactful and developing areas of research, especially for network science which has only been considered a field of basic science research since 2005 [3].

In the paper "The Structure of Scientific Collaboration Networks" by M. E. J. Newman, the author investigates structures of scientific paper co-authorship social networks. Specifically, the author considers clustering coefficients, component sizes, and tau - a description of whether highly connected nodes or loosely connected nodes dominate the network, to describe the structure of medical and physics collaborations in published papers [4]. We will be using similar approaches measuring centrality, characteristics of the giant component, clustering, and tau to understand patterns of collaboration and communication in the fields of computer science, physics, and medicine. We will compare these networks against random graphs such that we can generate statistically significant insights in their structural distinctiveness. Additionally, among datasets that include annotations (psoriasis research and Google Scholar network datasets), we wish to relate the number of citations to the overall centrality. For the psoriasis research network, if time permits, we will also look at the evolution in previously mentioned co-authorship network properties over a series of decades.

2 Resources Needed

The first dataset, which was sourced by researchers from Google Scholar, contains annotations of job positions (professor/post-doc/student/unknown), total number of publications, and discipline (biology/sociology/CS)[1]. This network consists of 400K nodes and 1.2 million edges. The data is formatted as a separate list of annotations for given nodes, and pairs of nodes that form edges. Consequently, we will need to consider only the largest component, as well as probably use ADA high-performance-computing resources to run our analysis in Python using the networkx package. Josef is already aware of how to submit jobs to ADA, and can share scripts to run Jupyter Notebooks on the cluster via SSH tunneling. A link to the first dataset can be found here: `https://github.com/chenyang03/co-authorship-network`. Should the first network still be too large in terms of required compute, we have an alternative physics co-authorship dataset containing 19K nodes and 200K edges, though unannotated: `https://snap.stanford.edu/data/ca-AstroPh.html`.

Our second dataset consists of a psoriasis research co-authorship network. Though the psoriasis co-authorship network dataset does not include information on number of citations, the names of individual researchers and PMIDs are provided. The NIH provides a free iCite API with example scripts, which allows us to easily extract additional information on the number of citations for given papers, which we can aggregate or average their respective authors. No direct information is provided on the number of nodes/edges, though the publishers of this network specify nearly 100K paper/author combinations, in which a given row specifies the PMID/year/author combination. This dataset can be found at: `https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/4HZFWY`.

3 Work Plan

3.1 Data cleaning: Apr 20

We anticipate having cleaned the entirety of the data and properly configured ADA/notebook access by April 20th. That is, we should be able to already load in all network datasets and run some basic analyses over these networks. This will likely take a bit of time to familiarize with ADA and how to interact and run jobs.

3.2 Graph analysis : Apr 27

By April 27th, we should be able to have computed clustering coefficients, tau, and other useful parameters that might reveal the initial interesting insights on our specific datasets. These will serve as our initial findings, though we will still need to perform significance tests to confirm our hypotheses.

3.3 Comparisons: May 4

By May 4th, we should have produced our random graph comparisons for testing statistical significance of our observed properties in the actual networks.

3.4 Visualizations: May 7

Given the absolute size of some of these datasets, it is probable that we will be producing visualizations related to the properties of the dataset networks rather than the networks themselves.

4 Vision & Contingencies

If everything goes well in the project, we anticipate that we will have determined, to a statistical degree of significance, whether or not the collaboration graph of different disciplines varies. In addition, we anticipate some effective visualizations of the different network graphs to visually convey some of the findings and differences between the networks. Also, various models of random graphs will be used with their characteristics compared to those of the actual collaborator networks of the various disciplines, determining what characteristics of the collaborator graphs are a result of the unique aspects of paper coauthorship. In the event of a partial success, we anticipate that the structure of fewer coauthorship networks, likely the psoriasis network, will be analyzed and compared against fewer random-graph models. However, the analysis structure will remain the same. Collaboration graphs are compared with the generation of a random model with its features analyzed with some meaningful insights, hopefully derived.

5 Anticipated Learning

Through this project we will learn how to apply network analysis techniques to large-scale real-world data, including computational challenges such as working with networks of over hundred thousand nodes. Some of those techniques would involve using centrality measures to identify influential nodes, calculating clustering coefficients, and determining the power law exponent of the network degree distribution. Throughout the analysis process, we will familiarize ourselves further with the NetworkX package in Python. Ultimately, we will also find whether the computer science research community exhibits similar or different network properties compared to other educational domains and whether coauthorship networks exhibit structural properties that differ from the Erdos-Renyi random graphs.

References

[1] Yang Chen et al. "Building and Analyzing a Global Co-Authorship Network Using Google Scholar Data". In: (2017).

[2] Ying Ding, Schubert Foo, and Gobinda Chowdhury and. "A Bibliometric Analysis of Collaboration in the Field of Information Retrieval". In: *The International Information & Library Review* 30.4 (1998), pp. 367–376. DOI: 10.1080/10572317.1998.10762484. eprint: https://doi.org/10.1080/10572317.1998.10762484. URL: https://doi.org/10.1080/10572317.1998.10762484.

[3] Roland Molontay and Marcell Nagy. "Two Decades of Network Science as seen through the co-authorship network of network scientists". eng. In: *arXiv.org* (2020). ISSN: 2331-8422.

[4] M. E. J. Newman. "The structure of scientific collaboration networks". eng. In: *Proceedings of the National Academy of Sciences* 98.2 (2001), pp. 404–409. ISSN: 0027-8424.