EECS 6414 - Midterm Progress Report: Co-Author Network Analysis and Understanding the Impact of LLMs in Different Research Domains from Arxiv Data

Nafis Tahmid Chowdhury
York University
Toronto, Canada
ntahmid@yorku.ca

Shrishti Pathak
York University
Toronto, Canada
ss72@yorku.ca

Mohammed Saidul Islam
York University
Toronto, Canada
saidulis@yorku.ca

1 INTRODUCTION

1.1 Motivation

Over the past decade, there has been a noticeable increase in public interest regarding the intricate inter-connectedness of modern society. This phenomenon revolves around the concept of networks, which are patterns of interconnections among various entities. Networks have become a focal point in discussions spanning a wide array of topics. The diverse contexts in which networks are invoked underscore their significance in contemporary discourse [3].

In our work, we will focus on studying the arXiv dataset[1] comprising data of authors and their research of various domains. Studying the relatedness of authors taps into the heart of modern societal networks, offering a wealth of insights into the collaborative nature of scientific advancement. By mapping out the interconnections between researchers, we can uncover the often invisible webs of knowledge transfer, intellectual influence, and collective progress. In an era where interdisciplinary research is increasingly paramount, exploring the relational dynamics within the arXiv authorship network is not just informative but essential for fostering collaborative opportunities and steering the future course of scientific inquiry. Furthermore, we will analyze the arXiv dataset's publication patterns over the years, which will show us which researchers are consistently active and the ebb and flow of research across various fields. This insight is crucial—it helps pinpoint where the scientific community is focusing its efforts and where potential breakthroughs are brewing. Moreover, the study of this dataset will reveal valuable information about recent developments in research and the influential techniques that are transforming various academic fields. This includes advancements in large language models (LLMs), such as ChatGPT [11], and their ongoing influence on interdisciplinary research over time.

Recently, Large Language Models (LLMs) like GPT-4 [12], Gemini [6], LLaMA-2 [13] have taken the world by storm. The exploration of LLMs is vital due to their broad implications beyond computer science. These models, powered by extensive data and computational advances, offer groundbreaking tools for language understanding that are reshaping numerous fields. For example, in linguistics, LLMs can analyze complex language patterns, while in the healthcare sector, they improve patient interaction through sophisticated language processing. The study of LLMs is also critical for identifying and mitigating built-in biases and ensuring ethical use. Therefore, multidisciplinary research into LLMs is essential not only for technological innovation but also for ensuring these advancements benefit society equitably and responsibly.

[1] https://www.kaggle.com/datasets/Cornell-University/arxiv

1.2 Related Works

Datasets from the real world usually contain a mix of different types of data such as images, text, and time series, which are intricately interconnected in a way that can be effectively represented using graphs. Recent developments in graph-based models have enabled us to better leverage the detailed characteristics and complex connections present in real-world data [1, 7, 8].

Since its inception in 1991, the arXiv[2] has emerged as the quintessential pre-print repository for a variety of disciplines including Computer Science, Mathematics, Physics, and numerous interdisciplinary fields. It has become a trusted platform for researchers to disseminate their findings before formal peer-review, often serving as the main literature source for many in these fields. With a collection exceeding 2.4 million papers, it presents a vast multigraph dataset, encompassing the full text of articles, metadata, and internal citation networks.

Notably, the arXiv has frequently been utilized as a research dataset in various studies. Examples include Liben-Nowell et al.'s [9] exploration of the arXiv co-authorship graph for link prediction, and Dempsey et al.'s [2] application of the authorship graph in testing network models. Furthermore, predictive modeling of future research trends was explored by Eger et al. [4] and Liu et al. [10] in machine learning and physics. The arXiv data also underpinned the 2003 KDD Cup [5], a competition focused on citation prediction, download estimation, and data cleaning. However, the use of different data subsets in these studies poses challenges for future comparative research. However, existing research has not delved into the complex relationships between authors conducting interdisciplinary research, nor has it considered the timing of research publications by various authors and their impact in their respective areas. Moreover, there has been a lack of focus on the implications of recent advancements in Large Language Models (LLMs). In light of this, we are optimistic that our study will provide valuable and important insights in these areas.

1.3 Problem Definition

In this project, we aim to perform an exploration and comprehensive analysis of the arxiv dataset, which aggregates scholarly articles from multiple academic disciplines over a 30-year timeline. Our approach involves a meticulous preprocessing phase where we apply sophisticated Natural Language Processing (NLP) techniques to manage and interpret the textual data effectively. A crucial aspect of this phase is the implementation of the Stanford Named Entity

[2] https://arxiv.org

Recognizer (NER) for the precise identification of individual authors' names, thereby avoiding misclassification with non-personal entities.

Following the preprocessing, the project will shift focus to a specialized form of network analysis. Utilizing tools like Python and Networkx, we plan to conduct an in-depth co-authorship network analysis. The goals of this analysis include identifying key figures in various academic fields, exploring the structure and interconnectedness of scholarly communities, and tracing the dissemination of ideas and methodologies among researchers, which is vital in understanding the evolution of academic thought and collaboration. Furthermore, we will explore the impact of recent advancements in LLMs and analyze their influence on interdisciplinary research.

2 METHODOLOGY

2.1 Dataset

The dataset we will use for our project is the arxiv dataset which is available publicly in Kaggle[3]. It covers a wide range of academic disciplines such as Mathematics, Computer Science, Quantitative Biology, and Quantitative Finance, to name a few, and compiles data collected over a period of 30 years. This dataset is substantial, encompassing approximately 2.4 million entries across 8 primary disciplines with 61 sub-specializations. Domains include Computer Science, Economics, Electrical Engineering and Systems Science, Mathematics, Physics, Quantitative Biology, Quantitative Finance, and Statistics. However, computational limitations prevented full utilization of the dataset for co-author network analysis. We addressed this by applying filtering criteria, specifically targeting papers with titles or abstracts referencing 'language model', 'pretrained language model', 'large language model', 'foundation model', 'BERT', 'XLNet', 'GPT-2', 'GPT-3', 'GPT-4', 'GPT-Neo', 'GPT-J', 'ChatGPT', 'PaLM', and 'LLaMA'.

2.2 Data Preprocessing

As the dataset consists of textual data, we employ various Natural Language Processing techniques in order to process our data. We discuss the data preprocessing pipeline in this section.

2.2.1 ***Data Acquisition and Initial Filtering.*** The initial stage involves acquiring the data and performing some basic filtering operations. A custom function is employed to convert the data stored in JSON format into a *Pandas* DataFrame, a versatile data structure for manipulation and analysis. We focus on the field of *Artificial Intelligence (AI)* and *Computational Linguistics (CL)*, thus the data is filtered to retain entries from the corresponding categories (`cs.CL` and `cs.AI`) within the dataset, as they contain the most occurrences of the keywords of interest. Further, we chose papers published within a designated timeframe, specifically from the year 2020 and onwards. This approach concentrates the analysis on the contemporary developments in ML and AI research, aligning the study with the current state of the field and ensuring its relevance to today's academic and industry standards.

2.2.2 ***Keyword-Based Refinement.*** Following the initial filtering, we perform further refinement of the data based on keywords related to a specific area of interest within the broader fields of AI and CL. In this case, we focus on research pertaining to large language models (LLMs) and ChatGPT. Specifically, we search by the following keywords, such as 'language model', 'pretrained language model', 'large language model', 'foundation model', 'BERT', 'XLNet', 'GPT-2', 'GPT-3', 'GPT-4', 'GPT-Neo', 'GPT-J', 'ChatGPT', 'PaLM', 'LLaMA', etc. We perform a case-insensitive search on both the title and abstract fields of each publication to identify relevant papers mentioning these keywords. This keyword-based filtering helps narrow down the dataset to research topics directly to LLMs.

2.2.3 ***Final data Preprocessing for Co-authorship Network Construction.*** After refining the data based on publication categories and keywords, we perform some further preprocessing steps to prepare the data for network analysis.

Parsing Metadata: The task of extracting and parsing metadata from the Arxiv dataset was crucial. It allowed us to obtain comprehensive details on each paper, including titles, abstracts, authors, and publication dates. This step is the backbone of all further analysis, enabling a structured approach to data handling.

Author Name Standardization: Given the varied formats of author names, we implemented a process to standardize these names across the dataset. This involved addressing complex cases such as authors with multiple names or incomplete name entries, ensuring a uniform format that facilitates accurate author identification and comparison.

Extraction of Co-Author Pairs: A pivotal part of our analysis involved identifying all possible pairs of co-authors for each paper. This step required the application of combinatorial logic, forming the basis for the subsequent construction of the co-author network graph. It is a critical process that enables the analysis of collaboration patterns within the dataset.

We retain essential columns containing information like title, abstract, author names, and publication for further processing. Further, we clean up the author data to extract individual author names and eliminate any unnecessary whitespace. This crucial step involves generating all possible pairs of co-authors from the author lists associated with each publication. Next, this co-authorship information serves as the foundation for constructing the network, where nodes represent researchers (authors) and edges represent collaborations between them. We take into account the frequency of co-authorship between authors, potentially indicating stronger collaborative relationships between frequently co-authoring researchers.

2.3 Network Analysis

In this section, we discuss our initial network analysis of Co-author network analysis on the filtered dataset.

2.3.1 ***Co-author Network Construction.*** Constructing a co-author network involves mapping the collaborative relationships between authors based on their shared academic outputs. In this process, *authors* are represented as nodes, and their *co-authorships* are represented as edges connecting these nodes. The innovative aspect of this construction lies in the weighting of the edges, which are determined by the number of papers co-authored by any given pair of authors. This weighted approach is pivotal as it provides a quantifiable measure of collaboration intensity, distinguishing between

[3]https://www.kaggle.com

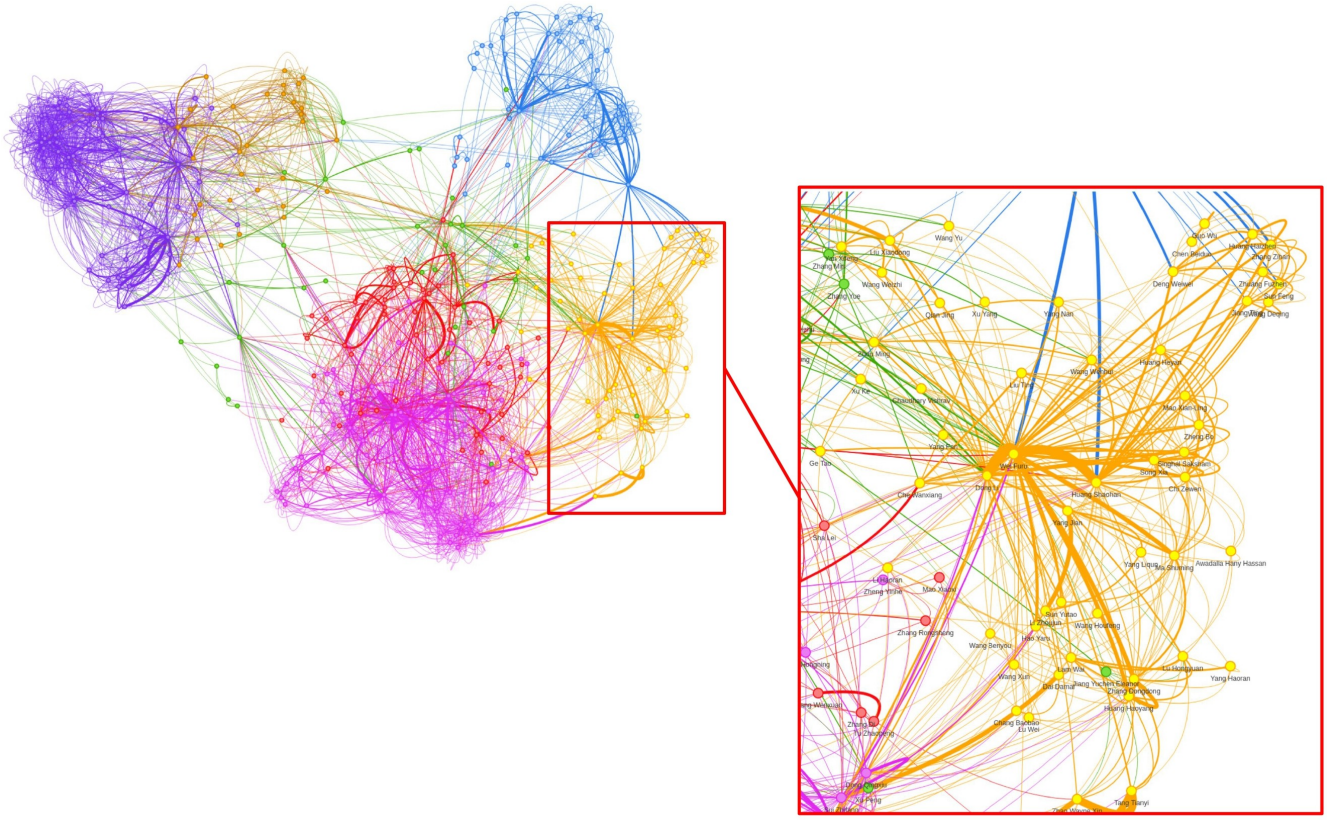

Figure 1: An example Co-Author network.

occasional collaborators and those with a more substantial joint research output. This methodological step is crucial for several reasons. Firstly, it enables the identification of key researchers within a field based on their collaborative behaviors, offering insights into how scholarly communities are structured and how knowledge flows within these communities. Secondly, the weighted edges contribute to a nuanced understanding of the network, allowing for the differentiation between central, highly collaborative authors and those with fewer, but potentially more strategic, collaborations. We employ Python's NetworkX library for the network construction process.

2.3.2 ***Refining the Network.*** After the initial construction, the network undergoes a refinement process to focus on the most relevant components of the scholarly community. This refinement involves the creation of a subgraph, which includes only those authors who meet specific criteria, such as a minimum number of publications, i.e., at least two publications from 2020 onwards. Such filtering ensures that the analysis remains focused on active and significant contributors to the field, thereby enhancing the network's relevance and manageability. Moreover, this phase involves the removal of isolated nodes, which are authors without any connections within the selected subset. Isolated nodes can occur for various reasons, such as authors having fewer than the minimum number of required co-authored papers with the selected group. By eliminating these nodes, our analysis concentrates on the main component of the network, where the interactions are most dense and, presumably, the most scientifically fruitful collaborations occur.

2.3.3 ***Visualization.*** After creating the co-authorship network, we generate the visual representation of the graph that the network represents. Using the *PyVis* library, we render the network as an interactive graph, where nodes represent authors, and edges represent co-authorships, with the thickness of the edges reflecting the strength of collaboration. This visual representation allows for an immediate understanding of the network's structure, highlighting the most densely connected clusters and the central nodes within them, as users can interact with the visualization, exploring connections between authors and identifying key contributors in specific areas of research. Figure 1 depicts such a network diagram.

3 EVALUATION

3.1 Visual Analytics

To deepen our understanding of this analysis, we will implement comprehensive visual analytics. These will clarify how authors are connected, using a variety of visual tools such as charts and graphs. These visualizations aim to effectively showcase the latest trends in research and the development of academic work over time. We will place a special emphasis on the substantial impact that large

language models have had on recent research. This study is crucial as it uses visual methods to simplify complex data, making it easier to understand. We will employ several visualization libraries i.e., matplotlib [4], seaborn [14], plotly[5], etc.

3.2 Challenges

As previously stated, our dataset comprises a substantial 2.4 million entries. In addition, the dataset contains eight major categories with 61 different sub-categories. While working with the dataset, we found that it requires a lot of RAM to load the dataset into the experiment environment. Further, due to the large volume of authors, significantly building and expanding the co-author network to include every research field and with all the authors is severely computationally expensive. Thus we had to narrow down the scope of our experiments and apply several different filtering techniques. Moving ahead, we want to further analyze Co-authorship networks from several different angles, such as research topics, date and time of publication, etc.

REFERENCES

[1] Peter W. Battaglia, Jessica B. Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, Caglar Gulcehre, Francis Song, Andrew Ballard, Justin Gilmer, George Dahl, Ashish Vaswani, Kelsey Allen, Charles Nash, Victoria Langston, Chris Dyer, Nicolas Heess, Daan Wierstra, Pushmeet Kohli, Matt Botvinick, Oriol Vinyals, Yujia Li, and Razvan Pascanu. 2018. Relational inductive biases, deep learning, and graph networks. arXiv:1806.01261 [cs.LG]

[2] Walter Dempsey, Brandon Oselio, and Alfred Hero. 2019. Hierarchical network models for structured exchangeable interaction processes. arXiv:1901.09982 [stat.ME]

[3] David Easley and Jon Kleinberg. 2010. *Networks, Crowds, and Markets*. Cambridge University Press, UK.

[4] Steffen Eger, Chao Li, Florian Netzer, and Iryna Gurevych. 2019. Predicting Research Trends From Arxiv. arXiv:1903.02831 [cs.CL]

[5] Johannes Gehrke, Paul Ginsparg, and Jon Kleinberg. 2003. Overview of the 2003 KDD Cup. *SIGKDD Explor. Newsl.* 5, 2 (dec 2003), 149–151. https://doi.org/10.1145/980972.980992

[6] Google. 2023. https://blog.google/technology/ai/google-gemini-ai.

[7] Palash Goyal and Emilio Ferrara. 2018. Graph embedding techniques, applications, and performance: A survey. *Knowledge-Based Systems* 151 (July 2018), 78–94. https://doi.org/10.1016/j.knosys.2018.03.022

[8] William L. Hamilton, Rex Ying, and Jure Leskovec. 2018. Representation Learning on Graphs: Methods and Applications. arXiv:1709.05584 [cs.SI]

[9] David Liben-Nowell and Jon Kleinberg. 2003. The link prediction problem for social networks. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management* (New Orleans, LA, USA) *(CIKM '03)*. Association for Computing Machinery, New York, NY, USA, 556–559. https://doi.org/10.1145/956863.956972

[10] Wenyuan Liu, Stanisław Saganowski, Przemysław Kazienko, and Siew Ann Cheong. 2019. Predicting the Evolution of Physics Research from a Complex Network Perspective. *Entropy* 21, 12 (Nov. 2019), 1152. https://doi.org/10.3390/e21121152

[11] OpenAI. 2022. https://openai.com/chatgpt.

[12] OpenAI. 2022. https://openai.com/research/gpt-4.

[13] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288 [cs.CL]

[14] Michael L. Waskom. 2021. seaborn: statistical data visualization. *Journal of Open Source Software* 6, 60 (2021), 3021. https://doi.org/10.21105/joss.03021

[4] https://matplotlib.org

[5] https://plotly.com